


IFIP AICT 340

Zhongzhi Shi
Sunil Vadera
Agnar Aamodt
David Leake
(Eds.)



Intelligent Information Processing V

6th IFIP TC 12 International Conference, IIP 2010
Manchester, UK, October 2010
Proceedings

 Springer

Editor-in-Chief

A. Joe Turner, Seneca, SC, USA

Editorial Board

Foundations of Computer Science

Mike Hinchey, Lero, Limerick, Ireland

Software: Theory and Practice

Bertrand Meyer, ETH Zurich, Switzerland

Education

Bernard Cornu, CNED-EIFAD, Poitiers, France

Information Technology Applications

Ronald Waxman, EDA Standards Consulting, Beachwood, OH, USA

Communication Systems

Guy Leduc, Université de Liège, Belgium

System Modeling and Optimization

Jacques Henry, Université de Bordeaux, France

Information Systems

Barbara Pernici, Politecnico di Milano, Italy

Relationship between Computers and Society

Chrisanthi Avgerou, London School of Economics, UK

Computer Systems Technology

Paolo Prinetto, Politecnico di Torino, Italy

Security and Privacy Protection in Information Processing Systems

Kai Rannenberg, Goethe University Frankfurt, Germany

Artificial Intelligence

Max A. Bramer, University of Portsmouth, UK

Human-Computer Interaction

Annelise Mark Pejtersen, Center of Cognitive Systems Engineering, Denmark

Entertainment Computing

Ryohei Nakatsu, National University of Singapore

IFIP – The International Federation for Information Processing

IFIP was founded in 1960 under the auspices of UNESCO, following the First World Computer Congress held in Paris the previous year. An umbrella organization for societies working in information processing, IFIP's aim is two-fold: to support information processing within its member countries and to encourage technology transfer to developing nations. As its mission statement clearly states,

IFIP's mission is to be the leading, truly international, apolitical organization which encourages and assists in the development, exploitation and application of information technology for the benefit of all people.

IFIP is a non-profitmaking organization, run almost solely by 2500 volunteers. It operates through a number of technical committees, which organize events and publications. IFIP's events range from an international congress to local seminars, but the most important are:

- The IFIP World Computer Congress, held every second year;
- Open conferences;
- Working conferences.

The flagship event is the IFIP World Computer Congress, at which both invited and contributed papers are presented. Contributed papers are rigorously refereed and the rejection rate is high.

As with the Congress, participation in the open conferences is open to all and papers may be invited or submitted. Again, submitted papers are stringently refereed.

The working conferences are structured differently. They are usually run by a working group and attendance is small and by invitation only. Their purpose is to create an atmosphere conducive to innovation and development. Refereeing is less rigorous and papers are subjected to extensive group discussion.

Publications arising from IFIP events vary. The papers presented at the IFIP World Computer Congress and at open conferences are published as conference proceedings, while the results of the working conferences are often published as collections of selected and edited papers.

Any national society whose primary activity is in information may apply to become a full member of IFIP, although full membership is restricted to one society per country. Full members are entitled to vote at the annual General Assembly, National societies preferring a less committed involvement may apply for associate or corresponding membership. Associate members enjoy the same benefits as full members, but without voting rights. Corresponding members are not represented in IFIP bodies. Affiliated membership is open to non-national societies, and individual and honorary membership schemes are also offered.

Zhongzhi Shi Sunil Vadera Agnar Aamodt
David Leake (Eds.)

Intelligent Information Processing V

6th IFIP TC 12 International Conference, IIP 2010
Manchester, UK, October 13-16, 2010
Proceedings

Volume Editors

Zhongzhi Shi
Chinese Academy of Sciences
Institute of Computing Technology
Beijing 100190, China
E-mail: shizz@ics.ict.ac.cn

Sunil Vadera
University of Salford
School of Computing, Science and Engineering
Salford M5 4WT, UK
E-mail: s.vadera@salford.ac.uk

Agnar Aamodt
Norwegian University of Science and Technology
Department of Computer and Information Science
7491 Trondheim, Norway
E-mail: agnar.aamodt@idi.ntnu.no

David Leake
Indiana University
Computer Science Department
Bloomington, IN 47405, USA
E-mail: leake@cs.indiana.edu

Library of Congress Control Number: 2010935792

CR Subject Classification (1998): I.2.4, I.2, H.3, I.4, I.5, I.2.7

ISSN 1868-4238
ISBN-10 3-642-16326-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-16326-5 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.com

© IFIP International Federation for Information Processing 2010
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 219/3180

Preface

This volume comprises the 6th IFIP International Conference on Intelligent Information Processing. As the world proceeds quickly into the Information Age, it encounters both successes and challenges, and it is well recognized nowadays that intelligent information processing provides the key to the Information Age and to mastering many of these challenges. Intelligent information processing supports the most advanced productive tools that are said to be able to change human life and the world itself. However, the path is never a straight one and every new technology brings with it a spate of new research problems to be tackled by researchers; as a result we are not running out of topics; rather the demand is ever increasing. This conference provides a forum for engineers and scientists in academia and industry to present their latest research findings in all aspects of intelligent information processing.

This is the 6th IFIP International Conference on Intelligent Information Processing. We received more than 50 papers, of which 35 papers are included in this program as regular papers and 4 as short papers. We are grateful for the dedicated work of both the authors and the referees, and we hope these proceedings will continue to bear fruit over the years to come. All papers submitted were reviewed by two referees.

A conference such as this cannot succeed without help from many individuals who contributed their valuable time and expertise. We want to express our sincere gratitude to the Program Committee members and referees, who invested many hours to reviews and deliberations. They provided detailed and constructive review reports that significantly improved the papers included in the program.

We are very grateful to have the sponsorship of the following organizations: IFIP TC12, The University of Salford, and at the Institute of Computing Technology, Chinese Academy of Sciences.

We hope you enjoy this diverse and interesting volume.

August 2010

Zhongzhi Shi
Sunil Vadera
Agnar Aamodt
David Leake

Organization

General Chairs

S. Vadera (UK)

M. Musen (USA)

R. Mizoguchi (Japan)

Program Chairs

Z. Shi (China)

A. Aamodt (Norway)

D. Leake (USA)

Program Committee

A. Aamodt (Norway)

J. Alvarez (France)

A. Bernardi (Germany)

N. Bredeche (France)

C. Bryant (UK)

L. Cao (Australia)

E. Chang (Australia)

C. Chen (USA)

E. Chen (China)

H. Chen (UK)

K. Chen (UK)

F. Coenen (UK)

I. Cohen (USA)

Z. Cui (China)

H. Dai (Australia)

S. Ding (China)

Z. Duan (China)

S. Dustdar (Austria)

J. Ermine (France)

P. Estrailier (France)

W. Fan (UK)

D. Feng (Australia/HK)

L. Hansen (Denmark)

T. Hong (Taiwan)

T. Honkela (Finland)

Z. Huang (Netherlands)

P. Ibarquengoyatia (Mexico)

G. Kayakutlu (Turkey)

J. Liang (China)

H. Leung (HK)

E. Mercier-Meziani (France)

F. Meziane (UK)

S. Nefti-Meziani (UK)

T. Nishida (Japan)

G. Osipov (Russia)

M. Owoc (Poland)

R. Pfeifer (Switzerland)

A. Rafea (Egypt)

K. Rajkumar (India)

T. Ritchings (UK)

D. Ruan (Belgium)

M. Saraee (UK)

F. Segond (France)

E. Sucar (Mexico)

Q. Shen (UK)

Z.P. Shi (China)

K. Shimohara (Japan)

A. Skowron (Poland)

M. Stumptner (Australia)

K. Su (China)

H. Tianfield (UK)

I.J. Timm (Germany)

S. Tsumoto (Japan)

Z. Vetulani (Poland)

X. Wang (China)

H. Xiong (USA)

J. Yang (Korea)

X. Yao (UK)

Y. Yao (Canada)

J. Yu (China)

J. Zhou (China)

Z.-H. Zhou (China)

J. Zucker (France)

Table of Contents

Keynote Presentations

Case-Based Reasoning Tomorrow: Provenance, the Web, and Cases in the Future of Intelligent Information Processing	1
<i>David Leake</i>	
Knowledge Mining Biological Network Models	2
<i>Stephen Muggleton</i>	
Multivariate Bandits and Their Applications	3
<i>John Shawe-Taylor</i>	
Image Semantic Analysis and Understanding	4
<i>Zhongzhi Shi</i>	

Semantic Web Services

Collaboration in Agent Grid Based on Dynamic Description Logics	6
<i>Limin Chen and Zhongzhi Shi</i>	
Requirement Driven Service Composition: An Ontology-Based Approach	16
<i>Guangjun Cai</i>	
Multi-agent and Workflow-Based Web Service Management Model	26
<i>Wenjia Niu, Quansheng Dou, Xu Han, Xinghua Yang, and Zhongzhi Shi</i>	
Semantic Approach for Service Oriented Requirements Modeling	35
<i>Bin Zhao, Guang-Jun Cai, and Zhi Jin</i>	

Automatic Reasoning

Extend Atomic Action Definitions of DDL to Support Occlusions and Conditional Post-Conditions	45
<i>Liang Chang, Zhongzhi Shi, and Tianlong Gu</i>	
Preservative Translations between Logical Systems	55
<i>Yuming Shen, Yue Ma, Cungen Cao, Yuefei Sui, and Ju Wang</i>	
The Description Logic for Relational Databases	64
<i>Yue Ma, Yuming Shen, Yuefei Sui, and Cungen Cao</i>	

Non-Functional Requirements Elicitation and Incorporation into Class Diagrams 72
Xiaoyu Song, Zhenhua Duan, and Cong Tian

Case-Based Reasoning

Architectures Integrating Case-Based Reasoning and Bayesian Networks for Clinical Decision Support 82
Tore Bruland, Agnar Aamodt, and Helge Langseth

Event Extraction for Legal Case Building and Reasoning 92
Nikolaos Lagos, Frederique Segond, Stefania Castellani, and Jacki O’Neill

Applications of CBR in Oil Well Drilling: A General Overview 102
Samad Valipour Shokouhi, Agnar Aamodt, and Pål Skalle

Data Mining

Associated Clustering and Classification Method for Electric Power Load Forecasting 112
Quansheng Dou, Kailei Fu, Haiyan Zhu, Ping Jiang, and Zhongzhi Shi

Two Improvement Strategies for PSO 122
Quansheng Dou, Shasha Liu, Ping Jiang, Xiuhua Zhou, and Zhongzhi Shi

Mining Temporal Patterns of Technical Term Usages in Bibliographical Data 130
Hidenao Abe and Shusaku Tsumoto

Automated Empirical Selection of Rule Induction Methods Based on Recursive Iteration of Resampling Methods 139
Shusaku Tsumoto, Shoji Hirano, and Hidenao Abe

Web Mining

Adaptive Web-Based Instruction for Enhancing Learning Ability 145
Wawta Techataweewan

Extracting Comparative Commonsense from the Web 154
Yanan Cao, Cungen Cao, Liangjun Zang, Shi Wang, and Dongsheng Wang

Detecting Temporal Pattern and Cluster Changes in Social Networks: A Study Focusing UK Cattle Movement Database	163
<i>Puteri N.E. Nohuddin, Frans Coenen, Rob Christley, and Christian Setzkorn</i>	

Unstructured P2P-Enabled Service Discovery in the Cloud Environment	173
<i>Jing Zhou and Zhongzhi Shi</i>	

Web Search

Using Global Statistics to Rank Retrieval Systems without Relevance Judgments	183
<i>Zhiwei Shi, Bin Wang, Peng Li, and Zhongzhi Shi</i>	

Rule Learning with Negation: Issues Regarding Effectiveness	193
<i>Stephanie Chua, Frans Coenen, and Grant Malcolm</i>	

Integrating Web Videos for Faceted Search Based on Duplicates, Contexts and Rules	203
<i>Zhuhua Liao, Jing Yang, Chuan Fu, and Guoqing Zhang</i>	

An Efficient Data Indexing Approach on Hadoop Using Java Persistence API	213
<i>Lai Yang and Zhongzhi Shi</i>	

Knowledge Representation

Knowledge Engineering for Non-engineers	225
<i>Tatiana Gavrilova</i>	

Attribute Exploration Algorithms on Ontology Construction	234
<i>Ping Qin, Zhongxiang Zhang, Hualing Gao, and Ju Wang</i>	

Intelligent Business Transaction Agents for Cross-Organizational Workflow Definition and Execution	245
<i>Mohammad Saleem, Paul W.H. Chung, Shaheen Fatima, and Wei Dai</i>	

Knowledge Granularity and Representation of Knowledge: Towards Knowledge Grid	251
<i>Maria A. Mach and Mieczyslaw L. Owoc</i>	

Natural Language Processing

Combining the Missing Link: An Incremental Topic Model of Document Content and Hyperlink	259
<i>Huifang Ma, Zhixin Li, and Zhongzhi Shi</i>	

A General Approach to Extracting Full Names and Abbreviations for Chinese Entities from the Web. 271
Jiang Guang, Cungen Cao, Yuefei Sui, Lu Han, and Shi Wang

An English-Arabic Bi-directional Machine Translation Tool in the Agriculture Domain 281
Khaled Shaalan, Ashraf Hendam, and Ahmed Rafea

A Laplacian Eigenmaps Based Semantic Similarity Measure between Words 291
Yuming Wu, Cungen Cao, Shi Wang, and Dongsheng Wang

Image Processing

A Filter-Based Evolutionary Approach for Selecting Features in High-Dimensional Micro-array Data 297
Laura Maria Cannas, Nicoletta Dessì, and Barbara Pes

A Novel Distribution of Local Invariant Features for Classification of Scene and Object Categories 308
LiJun Guo, JieYu Zhao, and Rong Zhang

Adult Image Detection Combining BoVW Based on Region of Interest and Color Moments 316
Yizhi Liu, Shouxun Lin, Sheng Tang, and Yongdong Zhang

Pattern Recognition

Multimedia Speech Therapy Tools and Other Disability Solutions as Part of a Digital Ecosystem Framework 326
David Calder

Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment 336
Urmila Shrawankar and Vilas Thakare

Proximity User Identification Using Correlogram 343
Shervin Shahidi, Parisa Mazrooei, Navid Nasr Esfahani, and Mohammad Sarae

Erratum

An Efficient Data Indexing Approach on Hadoop Using Java Persistence API E1
Lai Yang and ZhongZhi Shi

Author Index 353

Case-Based Reasoning Tomorrow: Provenance, the Web, and Cases in the Future of Intelligent Information Processing

David Leake

School of Informatics and Computing, Bloomington
Indiana University

Abstract. The World Wide Web and grid computing provide new opportunities and challenges for artificial intelligence. This talk examines how case-based reasoning can respond to these challenges by leveraging large-scale information sources. It highlights opportunities for exploiting naturally arising cases and augmenting them with additional open sources, to enable robust support for human reasoning. It illustrates with examples from current research, focusing especially on how CBR can leverage frameworks being developed in burgeoning research activity in provenance capture and storage.

Knowledge Mining Biological Network Models

Stephen Muggleton

Department of Computing
Imperial College London

Abstract. In this talk we survey work being conducted at the Centre for Integrative Systems Biology at Imperial College on the use of machine learning to build models of biochemical pathways. Within the area of Systems Biology these models provide graph-based descriptions of bio-molecular interactions which describe cellular activities such as gene regulation, metabolism and transcription. One of the key advantages of the approach taken, Inductive Logic Programming, is the availability of background knowledge on existing known biochemical networks from publicly available resources such as KEGG and BioCyc. The topic has clear societal impact owing to its application in Biology and Medicine. Moreover, object descriptions in this domain have an inherently relational structure in the form of spatial and temporal interactions of the molecules involved. The relationships include biochemical reactions in which one set of metabolites is transformed to another mediated by the involvement of an enzyme. Existing genomic information is very incomplete concerning the functions and even the existence of genes and metabolites, leading to the necessity of techniques such as logical abduction to introduce novel functions and invent new objects. Moreover, the development of active learning algorithms has allowed automatic suggestion of new experiments to test novel hypotheses. The approach thus provides support for the overall scientific cycle of hypothesis generation and experimental testing.

Multivariate Bandits and Their Applications

John Shawe-Taylor

Centre for Computational Statistics and Machine Learning
University College, London, UK

Abstract. We will review the multi-armed bandit problem and its application to optimizing click-through for Web site banners. We will present multi-variate extensions to the basic bandit technology including the use of Gaussian Processes to model relations between different arms. This leads to the consideration of infinitely many arms as well as applications to grammar learning and optimization.

Image Semantic Analysis and Understanding

Zhongzhi Shi

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

Abstract. Image understanding is widely used in many areas like satellite imaging, robotic technologies, sensory networks, medical and biomedical imaging, intelligent transportation systems, etc. But it is difficult by traditional image processing. Recently semantic analysis has become an active research topic aimed at resolving the gap between low level image features and high level semantics which is a promoting approach in image understanding.

This talk highlights the sophisticated methodologies of image semantic analysis, including discriminative, generative, cognitive methodology. Discriminative methodology is a data driven and uses classical machine learning, such as Kernel function, ensemble method, multi-instance, etc. Generative methodology is model driven and utilizes graphical models with text semantic analysis and each note with conceptual definition. Cognitive methodology can achieve four levels of generic computer vision functionalities: detection, localization, recognition, and understanding which are very useful for semantic knowledge exploration and image understanding. The feature binding computational model proposed by the Intelligence Science Laboratory will be presented in this talk.

People understand the nature of the image scene is through the internal syntactic structure of image. Originally syntactic structure of the sentence is generated through a series of production rules that the words are divided into a number of interrelated terms portfolio, reflecting the constraints between words within syntactic relations. Image parsing studies the image semantics directly. An image with certain hierarchical entities can be represented by and-or graph, that is, the parse tree. Syntactic reasoning with and-or graphs usually adopts a top-down and bottom-up strategy. Semantic Web technology promotes the development of image semantic analysis. In particular, the Web Ontology Language OWL provides a rich syntactic structure of semantics for image syntactic description in which different ontologies have explicit knowledge of dependencies, different text by OWL ontology mapping connected with high reusability. In terms of OWL and-or graphs are usually converted to RDF format. The semantic representation of image syntactic structure will achieve image-text standardized output possible. In this talk event exploring will be used to illustrate the procedure and principle ideas of visual syntax analysis which is easy to catch the scenic context.

This talk also concerns granular computing which is a new viewpoint and will impact on image understanding. We have proposed a tolerant granular

space model and applied it to image processing. Finally, the directions for further research on image semantic analysis and understanding will be pointed out and discussed.

Acknowledgement. This work is supported by National Basic Research Priorities Programme (No. 2007CB311004), National Natural Science Foundation of China (No. 60775035, 60933004, 60903141, 60970088), National Science and Technology Support Plan (No. 2006BAC08B06). We will review the multi-armed bandit problem and its application to optimizing.

Collaboration in Agent Grid Based on Dynamic Description Logics

Limin Chen¹ and Zhongzhi Shi²

¹ The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China
Graduate University of Chinese Academy of Sciences, Beijing 100049, China
chenlm@ics.ict.ac.cn

² The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China
shizz@ics.ict.ac.cn

Abstract. The global expansion of the Web brings the global computing; and the increasing number of problems with increasing complexity & sophistication also makes collaboration desirable. In this paper, we presented a semantics-based framework for collaborative problem solving in agent grid by coupling joint intention and dynamic description logics (DDL), our previous work to extend description logics (DL) with a dynamic dimension to model the dynamic world. Capabilities and attitudes of agents were captured by actions, and formulas in DDL respectively. Thus representation components in our framework were conferred with well-defined semantics by relating them to some domain ontologies. We could employ reasoning on actions in DDL to help agents to find proper colleagues when collaboration is necessary, and the philosophy underlying Joint Intention to bind their actions to achieve their overall goal. The main strengths of our framework include: i) finding probably helpful agents in a semantically accurate way due to the employment of semantic information; ii) going much closer to industrial implementations while retaining the main express power of classical joint intention model.

Keywords: Agent grid, dynamic description logics, joint intention, collaborative computing.

1 Introduction

To solve the problems with increasing complexity and sophistication, substantial time, effort and finances have been devoted to developing complex and sophisticated software systems, which places greater demand on the knowledge content and executing power of software agents. Collaboration might be a promising way to ease this tension. Furthermore, the need for collaboration among intelligent systems has also been fuelled by the global expansion of the Web and the advent of the paradigm of service oriented computing (SOC).

The emerging computational grid as well as agent technologies provides a good basis for building super collaboration frames for complex problem solving. The grid community has historically focused on infrastructure, tools, and applications for reliable and secure resource sharing within dynamic and geographically distributed virtual organizations, while the agent community has focused on autonomous problem solvers that can act flexibly in uncertain and dynamic environments [1]. As the scale and ambition of both grid and agent deployments increase, a convergence of interests in the agent and grid communities emerges: agent systems require robust infrastructure and grid systems require autonomous, flexible behaviours.

The past decades have witnessed researchers' attempts to apply agent technologies to realize the grid vision that enables resource sharing, provides the basic mechanisms for forming and operating dynamic distributed collaborations, or virtual organizations [2], and facilitates the unification of geographically dispersed computing systems to form a more powerful one. The most interesting work might be DARPA ISO's Control of Agent-Based Systems (*CoABS*) program, which firstly proposes the concept of 'Agent Grid' [3]. Shi et al. propose a model for agent-based grid computing from a point view of implementation, and develop *AGEGC*, an agent grid using a multi-agent environment platform based on the model [4]. Due to the merits inherited from agent and grid, agent grid has some advantages over the traditional approaches dealing with more demanding problems.

This paper addressed collaboration in agent grid by coupling dynamic description logics and joint intention. Sec.2 was devoted to a brief overview of dynamic description logics (*DDL*), a dynamic extension of *DL* [5]-[6]. In Sec.3, we proposed a collaboration model based on joint intention and dynamic description logics. Sec.4 concluded the paper with a discussion on the future work.

2 An Overview of Dynamic Description Logics

DDL is PDL-like dynamic extensions of *DL* [7]. Actually it is a family of languages, depending on the underlying *DL*. When necessary to be specific, we write *D-ALC*, *D-SHOIQ*, and the like. For simplicity, we choose *ALCO* as the underlying logic, and refer to the resulted *DDL* as *D-ALCO*.

(Syntax). Primary alphabets of *D-ALCO* include: i) N_R for role names; ii) N_C for concept names; iii) N_I for individual names; and iv) N_A for atomic action names.

The concepts and roles in *D-ALCO* are the same as that in *ALCO* with " $C, D \rightarrow C_i \mid \{o\} \mid \neg C \mid (C \sqcap D) \mid \exists R. C$ " & " $R \rightarrow P$ ", where $C_i \in N_C$, $o \in N_I$, $P \in N_R$. We use \perp , \top , $(C \sqcup D)$, and $\forall R. C$ to shorthand $(C \sqcap \neg C)$, $\neg \perp$, $\neg(\neg C \sqcap \neg D)$, and $\neg \exists R. \neg C$, resp..

Formulas in *D-ALCO* are built with: $\varphi, \psi \rightarrow C(u) \mid R(u, v) \mid \neg \varphi \mid \varphi \vee \psi \mid \langle \pi \rangle \varphi$, where $u, v \in N_I$, $R \in N_R$ and π is an action defined later. We define the logical connectives " \rightarrow " and " $\langle \pi \rangle$ " in terms of " \neg ", " \vee ", as usual, and define " $[\pi]\varphi$ " as " $\neg \langle \pi \rangle \neg \varphi$ ". **ABox** and **TBox** in *D-ALCO* are the same as that in *ALCO*. Here we assume that readers have some familiarity with *DL* and omit some details of them.

An atomic action in D -ALCO is defined as $\langle \alpha, C, Pre, Eff \rangle$, where i) $\alpha \in N_A$ is the name of the atomic action; ii) C is a concept denoting the category the atomic action belongs; iii) Pre is a finite set of formulas; and iv) Eff is a finite set of formulas of the form $C(a)$ or $R(a,b)$, or their negations. D -ALCO actions are built with: $\pi, \pi' \rightarrow a \mid \varphi? \mid \pi \cup \pi' \mid \pi$; $\pi \mid \pi^*$, where a is an atomic action, and φ is a formula i. An action box is a finite set of atomic actions in D -ALCO.

A domain specification is defined as: $DS = \langle T, A, ActBox, SSet \rangle$, where T is a TBox, consisting of domain constraints; A is an ABox for the initial world; $SSet$ is the set of Aboxes for the world in a time slice; $ActBox$ is an action box, the dynamic aspects of the world evolvments.

(Semantics). A D -ALCO interpretation is a pair (I, \mathcal{W}) , where $I = (\Delta^I, \cdot^I)$ is an ALCO-interpretation and \mathcal{W} is a set of ALCO-interpretations. I consists of nonempty domain Δ^I and mapping \cdot^I that assigns atomic concepts to a subset of Δ^I , each $a \in N_I$ to an Δ^I -element, and each role to a subset of $\Delta^I \times \Delta^I$. The semantics of D -ALCO is formally defined in Table 1, where A is an atomic concept, R a role, C, D concepts, α is an atomic action, and π_1, π_2 denote actions:

Table 1. Semantics of D -ALCO

1) $A^{I, \mathcal{W}} = A^I$;	2) $R^{I, \mathcal{W}} = R^I$;	3) $\{u\}^{I, \mathcal{W}} = \{u^I\}$;
4) $(\neg C)^{I, \mathcal{W}} = \Delta^I \setminus C^{I, \mathcal{W}}$;	5) $(C \sqcap D)^{I, \mathcal{W}} = C^{I, \mathcal{W}} \cap D^{I, \mathcal{W}}$;	
6) $(\forall R. C)^{I, \mathcal{W}} = \{x \in \Delta^I \mid \forall y. (x, y) \in R^{I, \mathcal{W}} \text{ implies } y \in C^{I, \mathcal{W}}\}$;		
7) $(\exists R. C)^{I, \mathcal{W}} = \{x \in \Delta^I \mid \exists y. (x, y) \in R^{I, \mathcal{W}} \text{ and } y \in C^{I, \mathcal{W}}\}$;		
8) $(\alpha)^{I, \mathcal{W}} = (Pre, Eff)^{I, \mathcal{W}} = \{(I, I) \mid I \text{ satisfies each } \varphi_i \in Pre, \text{ and } C^{I, \mathcal{W}} = (C^{I, \mathcal{W}} \cup \{u^I \mid C(u) \in Eff\}) \setminus \{u^I \mid \neg C(u) \in Eff\}; R^{I, \mathcal{W}} = (R^{I, \mathcal{W}} \cup \{(u^I, v^I) \mid R(u, v) \in Eff\}) \setminus \{(u^I, v^I) \mid \neg R(u, v) \in Eff\}\}$;		
9) $(\varphi?)^{I, \mathcal{W}} = \{(I, I) \mid I \text{ satisfies } \varphi\}$;		
10) $(\pi_1 \cup \pi_2)^{I, \mathcal{W}} = (\pi_1)^{I, \mathcal{W}} \cup (\pi_2)^{I, \mathcal{W}}$;		
11) $(\pi_1; \pi_2)^{I, \mathcal{W}} = \{(I, I) \mid \text{there exists some } I_t \in \mathcal{W} \text{ such that } (I, I_t) \in (\pi_1)^{I, \mathcal{W}} \text{ and } (I_t, I) \in (\pi_2)^{I, \mathcal{W}}\}$		
12) $(\pi_1^*)^{I, \mathcal{W}} = \text{the reflective transitive close of } (\pi_1)^{I, \mathcal{W}}$.		

We still need to make precise the meaning of ‘‘satisfies’’ in 8) and 9). The satisfaction of a formula F in (I, \mathcal{W}) , written as $(I, \mathcal{W}) \models F$, is defined in Table 2:

Table 2. The Satisfaction of F in (I, \mathcal{W})

1) $(I, \mathcal{W}) \models C(a)$ iff $a^{I, \mathcal{W}} \in C^I$;	2) $(I, \mathcal{W}) \models R(a, b)$ iff $(a^{I, \mathcal{W}}, b^{I, \mathcal{W}}) \in R^{I, \mathcal{W}}$;
3) $(I, \mathcal{W}) \models \neg \varphi$ iff $(I, \mathcal{W}) \not\models \varphi$;	4) $(I, \mathcal{W}) \models \varphi \vee \psi$ iff $(I, \mathcal{W}) \models \varphi$ or $(I, \mathcal{W}) \models \psi$;
5) $(I, \mathcal{W}) \models \langle \pi \rangle \varphi$ iff there exists an $I' \in \mathcal{W}$ such that $(I, I') \in \pi^I$ and $(I', \mathcal{W}) \models \varphi$.	

A formula φ is satisfiable w.r.t a TBox T , if a *D-ALCO* interpretation models T and φ . The main reasoning tasks in *DDL* can be reduced to satisfiability checking of formulas, which is *decidable* in *D-ALCO* [5]. Due to the space limitation, we do not elaborate on reasoning tasks in *D-ALCO* (See [5] for further details).

3 Collaboration by Joint Intention and Dynamic Description Logics

In this section, we proposed a collaboration model which employed reasoning on actions in *DDLs* to find probably helpful agents and the philosophy underlying Joint Intention to bind their teamwork.

Intention is the notion to characterize both our action and our mental states in our commonsense psychology [8]. While in a collaboration environment, intentions are not sufficient in that teamwork involves more than just the union of the simultaneous individual actions even if they are coordinated [9]. As a joint commitment to act in a shared belief state, joint intention binds team members, and enables the team to overcome misunderstandings and surmount obstacles [10].

Cohen and Levesque devise a formalism to express joint intentions with strong semantics and a high level of maturity [9, 11-12]. The main drawback of C&L's formalism is that it stays too far from the industrial application. In our model, rather than as modal operators to model the attitudes of community members, beliefs and goals of agents are expressed as *DDL*-formulas and simply grouped into the corresponding classes. Intentions of an agent are defined as its aim to perform some actions to fulfil some goals. The joint attitudes or intentions of a team are defined in terms of those of its members. Capabilities of agents are represented by *DDL*-actions, thus the problem of capability matching, i.e., to find proper peers to do proper jobs, can be solved by *DDL*-reasoning. Then the philosophy underlying joint intention is employed to bind their actions to achieve their overall goal.

3.1 Conceptualizing Agent Grid

This subsection gives some primary definitions about agents and the specifications of agents and agent grid. In the sequel, we use Greek symbols such as α β γ , possibly with subscripts, to name agents.

Defn 3.1 (Belief). Let w be an environment, for an agent α , a formula b is a belief of α if α believes to be true in w , where s is the ABox describing w . We denote by $B_{\alpha,s}$ the set of α 's beliefs in w .

We assume agents are rational but not logically omniscient. So belief of an agent in any environment is consistent, but may be not close under logical references.

Beliefs of an agent are environment-dependent, i.e., changing along with the evolvement of environments. We assume that agents have records of their past beliefs. The sequence of beliefs of an agent α in an evolvement path is a belief record of α .

Defn 3.2 (Mutual Belief). For an environment w described by state s , formula b is a mutual belief of agent α and β in w if and only if $b \in B_{\alpha,s} \cap B_{\beta,s}$.

Of course, agents have capabilities to change the environments; and we assume that these capabilities are the only factors that cause environment changes. This assumption rules out the possibility of unpredictable changes on environments.

Defn 3.3 (Capability). A capability of agent α is an action π in *DDL* describing some ways the agent can change the environments, written as $\langle \alpha, \pi \rangle$.

We can also safely state that an agent changes the environments with the aim to enter into environments with certain properties. Typically, agents' goals can be classified into two types: maintenance goals and achievement goals. The former are something initially true and are not allowed to be changed over the evolvments of the system, while the latter are currently false and agents attempt to make true later. For example, consider an environment where a frail vase stands on a table, and if only one side is lifted, then the vase slides down and crumbles. If a team of agents have an achievement goal to lift the table, they may simultaneously try to keep the table flat in the lift process, which is a maintenance goal. In this paper we focus on achievement goals, and assume that an agent retains a goal until the goal has been fulfilled or the motivation becomes irrelevant (persistent goals).

Defn 3.4 (Goal). A goal of an agent α relative to m is an ordered pair of the form $\langle m, g \rangle$, where m is a formula in *DDLs* stating the motivation, and g a formula characterizing the desirable environments of α .

Agents' goals direct their attempts and determine the actions agents take to fulfil these goals. Intentions of an agent are persistent goals to do some actions.

Defn 3.5 (Intention). An intention of an agent α relative to m is an ordered pair of the form $\langle m, actExp \rangle$, where i) m is a *DDL*-formula stating the motivation (usually, a goal of α); and ii) $actExp$ is a *DDL*-action.

Agents' intentions are goal-dependent. Given a goal and a set of actions available to perform, the *planning problem* is to find an action list whose execution reaches an environment the goal holds. Such plans may pre-exist as a part of agents' knowledge, or can be computed dynamically.

Joint intentions are a special kind of intentions whose component actions are relative to other agents.

Defn 3.6 (Joint Intention). A joint intention of agent α relative to m with agents $\{\beta_1, \beta_2, \dots, \beta_k\}$ is a ternary $\langle m, \pi, AgS \rangle$, where i) m is a formula; ii) π is an action; and iii) AgS is the set concerned agents.

An agent cannot drop a joint intention arbitrarily, and once the intended action has been performed or becomes irrelevant, the agent must make its new mental state about the intention known to the relative agents before its discard of the intention.

Defn 3.7 (Agent Specification). An agent is described by a tuple: $\langle \alpha, G, B, C, I, JI, K_{on}PC, K_{on}E \rangle$, where i) α : the agent's name; ii) G : the set of agent's goals; iii) B : agent's current beliefs; iv) C : $\{\pi | \langle \alpha, \pi \rangle\}$, the set of actions describing α 's capabilities; v) I : the set of agent's intentions; vi) JI : the set of agent's joint intentions; vii) $K_{on}PC$: a set of capabilities of other agents; viii) $K_{on}E$: a subset of the domain specification, the knowledge about the environment.

An agent grid is specified by the following definition:

Defn 3.8 (Agent Grid). An agent grid is a tuple: $\langle W, T, L, AgS \rangle$, where i) W : the set of possible environments; ii) T : a subset of $W \times W$, the binary relation on W , for $e_i, e_j \in W$, $\langle e_i, e_j \rangle \in T$ iff some agent in the agent grid can perform some actions transforming e_i to e_j ; iii) L : $W \rightarrow 2^{SSet}$, a labeling function, i.e., mapping any $w \in W$ to a subset of $SSet$, the set of ABoxes in domain specification (cf. Sec. 2); and iv) AgS : the set of community members of the grid.

T is intended to model transitions on environments and is transitive, whose transitive close is **accessibility**, denoted by A . Evolvement paths, or paths for short, are defined as chains in the partial order set $\langle W, A \rangle$.

3.2 Planning for Goals

Once an agent α has a goal g , the next thing for α is to get a plan for g . As mentioned above, plans may pre-exist as a part of agents' knowledge, or can be computed dynamically. Here, we propose a planning algorithm where the *planning problem* can be reduced to the satisfiability checking of *DDL*-formulas [5].

For agent α with a goal g , α computes a plan for g based on its knowledge on the current environment, on its capability and on its peers' capabilities. Before presenting the algorithm, we first give some results concerned, and the interested reader can refer to [5] for deeper investigations on topics of reasoning on actions in *DDLs*.

Suppose $PlanCandidate = \langle a_1, a_2, \dots, a_k \rangle$ is an action list, and that s is the ABox for the initial environment. In the sequel, we denote by $Conj(s)$ the conjunction of member formulas in a formula set s . Let us explore the following two formulas.

$1) [(a_1 \cup \dots \cup a_k)^*] \Pi \wedge Conj(s) \rightarrow \langle PlanCandidate \rangle true$, where Π is $\bigwedge_{i=1}^k (Conj(P_i) \rightarrow \langle a_i \rangle true)$ and P_i the precondition of a_i for each $i: 1 \leq i \leq k$.

The formula $[(a_1 \cup \dots \cup a_k)^*] \Pi \wedge Conj(s)$ can be viewed as an evolvement axiom: started from the environment described by s , during the process of environment evolvements, any action can be performed if its preconditions are satisfied. The whole formula says that *PlanCandidate* can be executed in the world respecting the evolvement axiom, and its validity requires its negation unsatisfiable!

$$[(a_1 \cup \dots \cup a_k)^*] \Pi \wedge Conj(s) \wedge \neg \langle PlanCandidate \rangle true \quad (1)$$

2) $Conj(s) \rightarrow [PlanCandidate]g$.

This formula states that after the execution of *PlanCandidate* in the environment described by *s*, the goal *g* holds, whose validity requires the following unsatisfiable.

$$Conj(s) \wedge \neg[PlanCandidate]g \quad (2)$$

The above facts can be employed by agent α to compute a plan for a goal from a set of actions. Table 3 shows the algorithm to compute a plan to fulfill a goal *g* from the set *CapSet* of the available actions, which travels all the possible action lists to find a plan, if any. All inputs of the algorithm are from the *G*, *C*, K_{onPC} and K_{onE} of α , including the ABox *s* describing the current environment, TBox *T* capturing the domain axioms, *CapSet*, the set of actions. In the algorithm, $\langle PlanCandidate, \alpha_i \rangle$ denotes the action list resulted by appending α_i to the rear of *PlanCandidate*.

Table 3. Plan computation

Algorithm 1 CapabilityComposition(*s*, *T*, *CapSet*, *g*)

[Input: state *s*, TBox *T*, Capabilities of agents *CapSet*, and goal formula *g*.]
 [Output: a unempy action sequence as a plan; or false (as a failure).]
 [Begin]

Initialize queue *QueOfPlanCandidates*;
 Enqueue *QueOfPlanCandidates* with empty action list $\langle \rangle$;
while (*QueOfPlanCandidates* is unempty) **do**
 Dequeue(*QueOfPlanCandidates*, *PlanCandidate*);
 if (Formula (1) is unsatisfiable w.r.t *T*) **then**

if (Formula (2) is unsatisfiable w.r.t *T*) **then**
 Return *PlanCandidate* as a successful plan;
else
 for each α_i : enqueue *QueOfPlanCandidates* with new candidate
 $\langle PlanCandidate, \alpha_i \rangle$;
end if;
end if
end while
 Return false;
 [End]

3.3 Working Collaboratively

When a task arrives, the agent concerned does a means-end analysis based on its capabilities, current intentions, and environment information to see whether the task can be solved locally. Suppose that α has belief $B_{\alpha,s}$ in the environment described by state *s*. For each $\langle m, g \rangle$, if $K_{onE}, B_{\alpha,s} \models m$, then α chooses *g* as a current goal. Then α computes plans for these goals and takes these plans as its current intentions. After a process of capability matching between the intended actions and its capabilities, α realizes the need for collaboration, and a further negotiation with other agents is invoked. If the task can be solved locally, then a set of local objectives are identified. Such objectives may contradict the agent's current intentions, thus a phase to check and resolve the inconsistency is need. *Goals* and *Beliefs* of the agent are also involved in the phase. Failures in this phase can also lead to the considerations of the task's feasibility. Then, a set of consistent intentions are generated and added to the agent's current intentions. Once a new (joint) intention is formed, another kind of *Inconsistency*

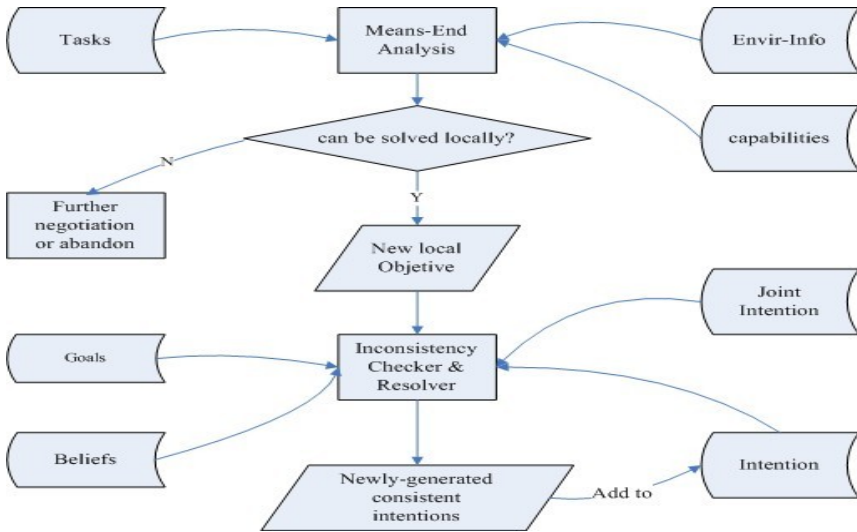


Fig. 1. An adoption process of a new task

Checker & Resolver will be invoked to ensure the consistency between individual intentions and joint ones. Figure 1 sketched such a process.

We further investigate what is involved when agents prefer to work collaboratively. The main consideration includes the following two questions:

What conditions should be satisfied before collaboration can proceed? Once an agent decides to achieve a goal through teamwork, the first thing for this originator agent is to figure out the capable and potential interested agents. The key point is to compute a semantically accurate match between the actions to be performed and those peers' capabilities.

The originator employs the classical reasoning tasks in description logics to find the capable agents by a matching between the intended actions and its *knowledge on peers' capabilities*. For the moment, we just simply employ the concepts to categorize the intended actions and agents' capabilities, and the match between agents' capabilities and the intended actions can be solved by a reduction to the subsumption between concepts: the agent with the capability C has the ability to perform the action T if C subsumes T . Further, classifying agents' capabilities by relating them to concepts in some ontology also structures the originator's knowledge about agents' capabilities in a hierarchical way, and such a structure can be taken into the originator's account in the process of finding and ranking the potential participants. Communication becomes necessary in finding the interested agents. The originator sends collaboration proposals to its target agents, promises to help in future may also be offered. The recipients undertake some rounds of the process depicted in Figure. 1, and decide the proposals should be accepted or refused. Finally, agreement on the details about the tasks will be reached through negotiations. Once such an agreement has been reached, the intentions concerned become joint intentions of the agents involved with its co-workers.

After the above phase, all the participants should acknowledge their acceptances to their own parts in the final plan. Only after all the participants have promised to contribute to the whole plan and all the details have been fixed, collaboration can proceed.

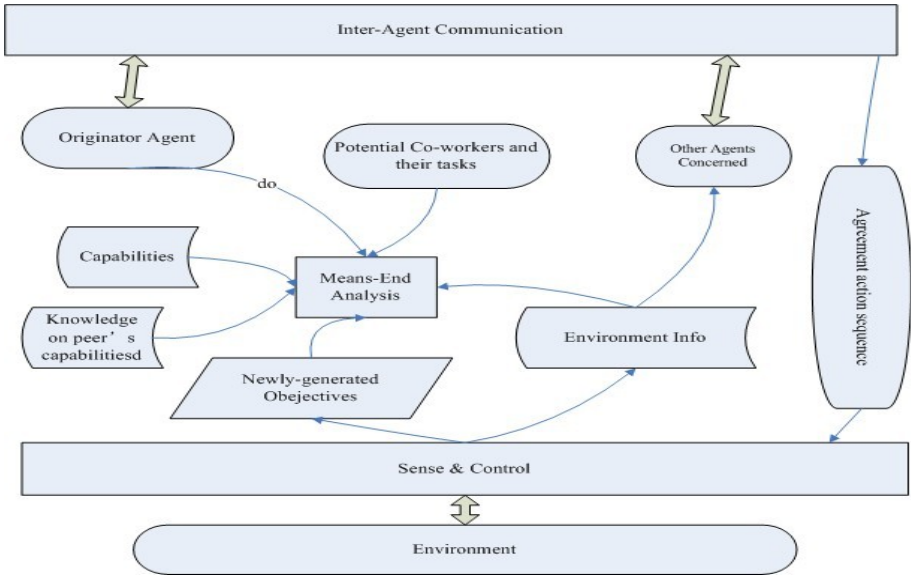


Fig. 2. Collaboration in agent grid

When and how agents should interact with their co-workers? Once cooperation begins, the participants are jointly committed to their individual goals and the overall goal until the collaboration is finished. When a participant comes to believe the goal is finished, or impossible to finish, or the motivations are not relevant any longer but that is not a mutually known, then the participant takes a new persistent goal to make its new discovery a mutual known. Then communications between the observant participant and its fellows are initiated, and the persistent goal terminates if all participants acknowledge their new known about the new status of the goal. The community deals with the newly-emerged situation accordingly, be it a reconsideration of the validity of the goal, a totally discard, a further negation or a new attempt to something less ambitious. Figure 2 sketches the architecture of collaboration in agent grid.

4 Conclusions and Future Work

We have presented a framework toward collaboration in agent grid. The framework is based on our ongoing work on extending description logics with a dynamic dimension and the philosophy behind Cohen & Levesque’s joint intention model. Rather than as modal operators to model the attitudes of community members, we just simply group

beliefs, goals, and intentions into corresponding classes. The joint attitudes of a community are defined in terms of the joint ones of its members. The problem of capability matching, i.e., finding the peers to do proper jobs, can be solved by *DDL*-reasoning; and the philosophy underlying joint intentions is employed to bind the team. While the collaboration model employs a *DDL*-based planning and capability matching to choose the potential participants and joint intention to bind and coordinate their actions, its higher-level nature in abstraction allows the potential tailors for proper fitness for the problems at hand.

One future work is to investigate the relation among *DDL*-actions further to compute a sophisticated specification of agents' capabilities, rather than the preliminary classifications by relating actions to *DL*-concepts in this paper. It is helpful to capture the partial matching with quantification degrees and to find the potential co-workers.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 60775035, 60933004, 60903141, 60970088), National Basic Research Priorities Programme (No. 2007CB311004), and National Science and Technology Support Plan (No. 2006BAC08B06).

References

1. Foster, I., Jennings, N., Kesselman, C.: Brain Meets Brawn. In: AAMAS 2004, pp. 8–15 (2004)
2. Foster, I., Kesselman, C.: The Grid, 2nd edn. Morgan Kaufmann, San Francisco (2003)
3. Manola, F., Thompson, C.: Characterizing the Agent Grid (1999), <http://www.objs.com/agility/tech-reports/990623-characterizing-the-agent-grid.html>
4. Shi, Z., Huang, H., Luo, J., et al.: Agent-based Grid Computing. Applied Mathematical Modeling 30(7), 629–640 (2006)
5. Chang, L., Lin, F., Shi, Z.: A Dynamic Description Logic for Representation and Reasoning about Actions. In: Zhang, Z., Siekmann, J.H. (eds.) KSEM 2007. LNCS (LNAI), vol. 4798, pp. 115–127. Springer, Heidelberg (2007)
6. Shi, Z., Dong, M., Jiang, Y., et al.: A Logic Foundation for the Semantic Web. Science in China, Series F 48(2), 161–178 (2005)
7. Baader, F., Calvanese, D., McGuinness, D., et al.: The description logic handbook. Cambridge University Press, Cambridge (2003)
8. Bratman, M.: Two Faces of Intention. The Phil. Review 93, 375–405 (1984)
9. Levesque, H., Cohen, P.: Teamwork. Nous 25, 487–512 (1991)
10. Cohen, P., Levesque, H.: Confirmations and Joint Action. In: IJCAI 1991, pp. 951–957 (1991)
11. Cohen, P., Levesque, H.: Intention is choice with commitment. AI 42, 213–261 (1990)
12. Levesque, H., Cohen, P., Nunes, J.: On Acting Together. In: AAI 1990, pp. 94–99 (1990)

Requirement Driven Service Composition: An Ontology-Based Approach

Guangjun Cai^{1,2}

¹ The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

² Graduate University of Chinese Academy of Sciences, Beijing 100049, China
caiguangj@mails.gucas.ac.cn

Abstract. Service-oriented computing is a new computing paradigm that utilizes services as fundamental elements for developing applications. Service composition plays a very important role in it. This paper focuses on service composition triggered by service requirement. Here, the processes modeling the requirement should be treated in parallel with describing service and a same ontology should be adopted for allowing the understanding between the requirement and services. An effect-based approach has been proposed based on our previous work on service description. This approach could be promising for tackling the challenge of services composition.

Keywords: Service-oriented computing, environment ontology, service composition, service discovery.

1 Introduction

Service-oriented computing (SOC) is a new computing paradigm that utilizes services as fundamental elements for developing applications [1]. Web service composition, which aims at solving complex problems by combining available basic services and ordering them to best suit the requirement and can be used to accelerate rapid application development, service reuse, and complex service consummation [2], is key to the success of SOC.

Many works, including standards, languages and methods, have been done to promote web service composition. But facing the challenge of high complexity composition problem with massive dynamic changeable services and on-demand request, most of them fail to provide an effective solution. Most approaches, such as [3] and [4], have only considered the requirement described by IO or IOPE. Second, few composite approaches use requirement as part of the composite service building process. Though the approach introduced by [4], [5]) consider the request in the composition process, the former fail to address the behavior and the latter leave all the task of requirement description to user. Consequently, they cannot cope with the challenge which the rapid change of user demands.

Different from them, we think that the composition process indeed relates to the requirement and user, the requirement should play a greater role in a service-oriented computing paradigm. Thus, it is necessary to consider the problem what the requirement

is in detail. This paper proposes a requirement-driven services composition method. We use environment ontology as the unified description of web service and requirement. From this starting point, a process modeling the requirement is processed and the environment ontology is adopted for allowing the understanding between them. Based on these effect-based structured descriptions, we design the algorithm for decomposing the requirement in order to identify suitable component services and take it as the basic step for composing service. It not only helps reduce the complexity of the problem and to improve concurrency, but also provides flexibility among various part of the requirement. And then we present the discovery algorithm for the precise matching by utilizing the unified explicit formal semantic description. Finally, we end with a method generating the composition service.

For the space limitation, this paper concentrates on the behavior aspect of service or requirement in a single domain. The rest of this paper is structured as follows. In section 2, we introduce environment ontology and a usage scenario. Section 3 proposes the base work of composing a web service, the descriptions of a web service or a requirement. Section 4 presents the algorithms decomposing the requirement, discovering suitable service and generating composition service. After discussing some related work in section 5, we conclude in section 6.

2 Environment Ontology

Environment ontology, which is proposed according to the thinking of requirement engineering based on the environment modeling, is extended further in [6]. In this approach, the environment of a Web service is viewed to be composed of those environment entities that the Web service can interact with. The concepts and associations of environment ontology are shown in figure 1. These associations form a general conceptualization of the particular entity. Entity type a web service imposed on is suggestive of some property of the service.

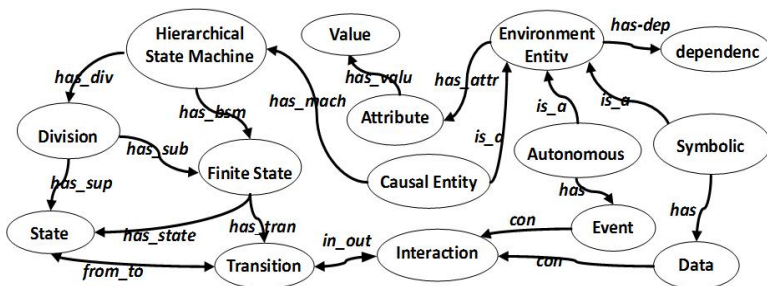


Fig. 1. Associations of environment ontology

In this paper, we embed a usage scenario for a travel system as an example. Part of the environment ontology about this domain is shown in figure 2, where the left figure show the environment entities and the relation among them, the right one is the part description about a causal entity ticket.

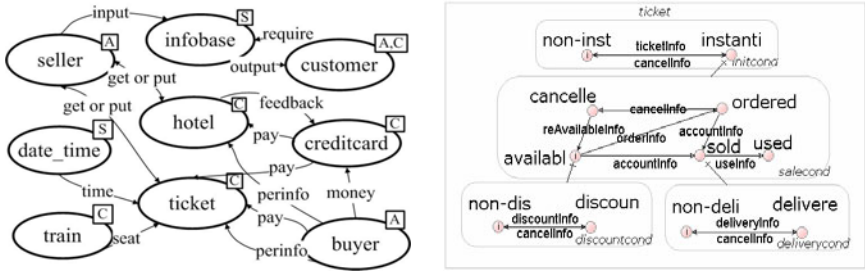


Fig. 2. Environment ontology of travel domain

3 Services Description and Requirement Description

Service description and requirement description is the prerequisite of service composition. Moreover, full automatic service composition needs a complete, formal specification. According to [3], it is difficult to provide a behavior-based requirement description for user.

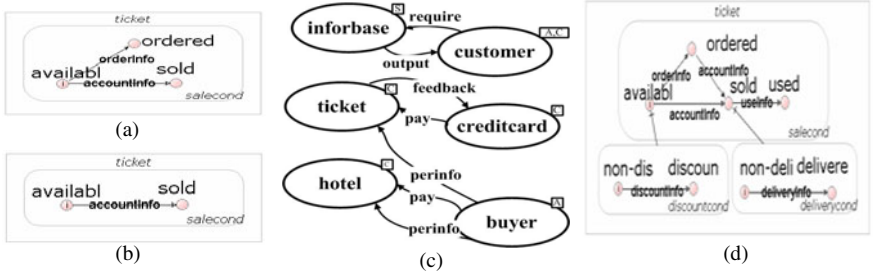


Fig. 3. Descriptions of service and requirement

Instead of focusing on the services of their own, environment ontology-based service description give attention on the effects imposed on their environment entities. The capabilities of services are expressed by the traces of the changes of the environment entities caused by the interactions of the services with these entities. We have presented the method for describing services[6]. Here some description results about ticket are presented in figure 3(a, b). They are part of environment ontology of the ticket and illustrate the changing process a service supports.

The process (figure 4) to describe the requirement is similar with that depicts web services, besides needing to facilitate the user to input and modify user information by using ontology. At the same time, the description result can be seen as the effect description of the composing web service if succeeded. The description result is also illustrated in figure 3, where part (c) shows the environment entities and the relations among them, and part (d) presents the requirement about the environment entity ticket.

```

ReqSpecification(userInfo, Onto, req) //requirement description based on environment ontology
Input: userInfo, Onto
Output: req
1   UserPro:=Generate(userInfo,Onto); //generate user profile in interaction manner,
    //which including environment entities selection, output and target states selection,
    // input and initial state selection
2   req:=generate(Onto, UserPro);// similar with the algorithm generating service description
3   if(ismulti(req)){
4     reqTem:=trim(req,userInfo);//determine the req according to user information
5     req:=ReqSpecification(userInfo, reqTem, req);
6   }
7   if(req=∅){
8     userPro:=Modify(UserPro, Onto);
9     req:=ReqSpecification(userInfo, Onto, req);
10  }
11  return req;

```

Fig. 4. Pseudocode of the requirement decomposition among environment entities

4 Services Composition

The task of web service composition is to search appropriate service and to arrange them in suitable order. Based on the environment ontology, we present a method through decomposing the requirement to search the available service and using the precise matching rules to judge. Thus, we divide the task into three parts: decomposing the requirement, discovering the suitable component services, generating the composition service.

4.1 Decomposing the Requirement

In our method, service requirement is described in a formal detailed way. It means that the same functional services have the same unique description. So we can acquire a composite service through a way which firstly decomposes the requirement into various parts to discover and compose the respecting services. The decomposing algorithms are introduced below.

Decomposition based relations among environment entities. The essence of the task in this section is to classify the set of environment entities, which need not consider the internal changing process in the environment entity under the effects of services. The behavior of decomposition among environment entities is summarized by the algorithm listed in figure 5, where req and reqSet denote the input and output of each algorithm, respectively. The requirement is firstly divided by DivByEntDep() according to the dependability among entities. The reason is that there is no dependence among different sub requirements. Secondly, we decompose a requirement by DivByEntIntTyp() according to the type num of dependences or entity type. The reason is that there is little constrain than other situation. Finally, the decomposition is

```

DivByEntDep(req, reqSet) //req contains unrelated environment entities
1. while(EnvEntNum(req) ≠ 0) { //req contains two or more environment entities
2.   e:=getEnvEnt(req); //gets environment entity from req
3.   if(Relate(e)≠EnvEnt(req)) { // Relate(e) is a set of environment entities which directly or
   //indirectly depend on the environment entity e
4.     eReq:=create(Relate(e), req); //create() generate the requirement responding to relate(e)
5.     reqSet:= reqSet ∪ {eReq};
6.     req:= Remove(Relate(e), req); //removes the requirement corresponding to relate(e)
7.   } //end if
8. } //end while

DivByEntIntTyp(req, reqSet) //req contains many related entities
11. if(∃e((e ∈ Req) ∧ (inDep(e)=0 ∨ outDep(e)=0 ∨ type(e)=A ∨ type(e)=S))) {
   //inDep(e), outDep(e)repents the number of dependents of environment entity e
   //type(e) represents the type of e
9.   newreq:= create(e, Req);
10.   ReqSet:= {newreq} ∪ DivByEntDep (req-newreq, reqSet);
11. }

DivByEnt(req, reqSet) //each req in reqSet contains only one environment entity
12. for(each e in Req) { // e represents environment entity
13.   ReqSet := ReqSet ∪ create(e, Req); //create() generate the requirement responding to e
14. } //end for

```

Fig. 5. Pseudocode of the requirement decomposition among environment entities

done by DivByEnt() based on environment entity. That is because there is only message dependence among different entities.

Decomposition in the environment entity. The requirement decompositions in the environment entity vary by type of entity. The decompositions in autonomous entity and symbolic entity, similar with transition-based decomposition in causal entity, can be directly done based on their basic behavior event or data by the algorithm DivByBeha() in figure 6. But for causal entity, we need to consider its hierarchical structure in detail.

```

DivInHsm(req, reqSet) //req contains exactly one hsm
1. rootfsm:=GetRootfsm(Req); //rootfsm, a finite machine having no super state
2. for(each subhsm of rootfsm) { //subhsm.rootfsm.superstate ∈ rootfsm
3.   ReqSet:=ReqSet ∪ create(subhsm, Req); // create(subhsm, Req) generate
   // the requirement responding to subhsm
4. }
5. ReqSet:=ReqSet ∪ create(rootfsm, Req); // create(rootfsm, Req) generate
   // the requirement responding to rootfsm

```

Fig. 6. Pseudocode of the requirement decomposition in environment entity

```

DivInFsm(req, kstate, reqSet) //req contains exact one fsm
7.  if((req.State-(From(kstate)  $\cup$  To(kstate)))= $\emptyset$ ){ //req.State represents all the state of req
8.    From(kstate):=From(kstate)-{kstate}; //From(kstate) represents all the states kstate can reach
9.    To(kstate):= To(kstate)-{kstate}; //To(kstate) represents all the states which can reach kstate
10. }//end if
11. req1:=create(req.State-(From(kstate)  $\cup$  To(kstate)), req); //the req in different branch
12. req2:=create(From(kstate)-To(kstate), req); //the req before the kstate or the loop containing it
13. req3:=create(To(kstate)-From(kstate), req); //the req after the kstate or the loop containing it
14. req4:=create(From(kstate) $\cap$ To(kstate),req); //the req in the same loop with kstate
15. reqSet:= {req1}  $\cup$  {req2}  $\cup$  {req3}  $\cup$  {req4};

DivByBeha(req, reqSet) //each req in reqSet only contains one basic behaviour
16.  for(each b in req){ //b can represent a transition, data or event
17.    reqSet := reqSet  $\cup$  {create({b}, req)};
18.  } //end for

```

Fig. 6. (continued)

For a requirement in one causal entity, we firstly divide it using the algorithm `DivInHsm()`, based on the reason the requirement of the finite state machines at the same sub-tree has greater relevance than at different sub-trees. The decomposition process could be repeated until finding available service or reaching that each requirement only contains one finite state machine. Then, we decompose the requirement according to the key states, which can be divided into two classes. The first class includes that the initial, middle or target states in it, the second contains the states which is the super state of some finite state machine. The reason for this is that for user should have capabilities to choose what to do next step on the key state, the functionality on both side of the state could be separable. If there were still no available services for the result requirement, the algorithm `DivByBeha()` is used.

Table 1. Decomposition result of each step by the decomposition algorithm

Algorithm	Decomposition result
DivByEntDep	{infobase, customer}, {hotel, creditcard, ticket, buyer}
DivByEntEntTyp	{infobase}, {customer}, {buyer}, {hotel}, {creditcard, ticket}
DivByEnt	{infobase}, {customer}, {buyer}, {hotel}, {creditcard}, {ticket}
DivInHsm	{salecond}, {discountcond}, {deliverycond}
DivInFsm	{available, ordered, sold}, {sold, used}
DivByBeha	{available, ordered}, {ordered, sold}, {available, sold}, {sold, used}

Taking the requirement in figure 3(d) as an example to present how to use these algorithms. For illustrating how to use them, we assume that each match fails below. The decomposition results are listed in table 1, where the set of entities, fsms and states denote corresponding requirement in different level requirement, respectively.

4.2 Selecting the Suitable Component Services

Service discovery, which enables a service requester to locate counterpart, plays a critical role in web service composition. With the increasing number of web services with similar functionality, measuring the match degree will become more and more important. However, there is a gap in most current approaches between service advertisement and service requirement. Environment ontology facilitates this, not only providing a unified semantic description, with additional knowledge about context, but also prompting to mark descriptions with the weight on modularity process ontology. Based on this description, we propose a precise match degree, as a sound criterion, to measure and select service. The functionalities can be measured by various aspects, such as the number of transitions, events, or data. The formulas calculating the matching degrees are shown as follows:

$$\text{ComDeg} = (\text{UseSerFun} / \text{ReqFun}) \times 100\%. \quad (1)$$

$$\text{NecDeg} = |\text{UseSerFun}| / (|\text{UseSerFun}| + \text{NumUseFun}) \times 100\% \quad (2)$$

$$\text{ValDeg} = (\text{UseSerFun} / \text{SerFun}) \times 100\% \quad (3)$$

$$\text{GloDeg} = W_c \times \text{ComDeg} + W_n \times \text{NecDeg} + W_v \times \text{ValDeg} \quad (4)$$

In the formulas above, SerFun, UseSerFun and ReqFun denotes all the functionalities, the useful functionalities for the requirement a service can provide and all the functionalities the requirement needs. Moreover, NumUseFun represents the number of the functionalities in UseSerFun which other services can provide; W_c , W_n and W_v represent the weights of ComDeg, NecDeg and ValDeg respectively and the sum of them equals 1.

Based on the definition of the matching degree, we describe a service discovery method in figure 7. The service is firstly selected according to their globe matching degree in line 4. Then based on the matching type of a service, we process them in corresponding method. Taking the available services a, b in figure 3, the requirement

```

Discovery(req, serSet, t, service, subSerSet)
Input: req, serSet, t; //t represent the threshold between the req and the service
Output: service, subSerSet; //subSerSet is be used when service is nil
1  if(serSet=∅){return nil;}
2  for(service in serSet){
3    compares services with req and calculates their matching degree;
4    if(gloDeg>t){matchSerSet:= matchSerSet ∪ {service}; }
5  }
6  while(matchSerSet≠∅){
7    chooses matSer with necDeg=100% or with maximal gloDeg;
8    if(matSer.comDeg=100%∧ matSer.valDeg=100%)//exact matching
9      return service;
10   if(matSer.comDeg=100%∧service.valDeg<100%){//subsume matching

```

Fig. 7. Pseudocode of the component services discovery

```

11     if(isdivide(matSer, req)){ return divide(matSer, req);}
12     else {continue;}
13     }//end subsume matching
14     if(matSer.comDeg<100%){ //interaction or plug-in matching
15         add matSer in subSerSet prepare for future choosing;
16     }//endif
17 } //end while

```

Fig. 7. (continued)

d in figure 3 as an example, ComDeg, NecDeg and ValDeg of a is $2/6*100\%$, $2/(1+2)*100\%$ and $2/2*100\%$, respectively. After getting the above values, we can easily calculate GloDeg of it to choose service.

4.3 Generating the Composition Service

The task of this step is to determine the relation among selected component service according to the relations between functionalities in a requirement and the relations between each component service with the responding sub requirements. The Pseudocode is listed in figure 8, where Req, Onto, serSet and t denote input, ComSerModel denotes the output.

```

GenComModel(Req, Onto, serSet, t, ComSerModel)
Input: Req, Onto, serSet, t
Output: ComSerModel
1  ReqSet.add (Req); //inserts the copy of req into ReqSet
2  while(ReqSet≠∅){
3    for(each req in ReqSet){
4      using discovery(req, serSet, t, subSerSet) discovery service and generate subSerSet;
5      ReqSet:=ReqSet-{req};
6      if(service≠nil){// discovery success
7        LabeledReq:=LabeledReq ∪ label(req, service.name); //labels Req using service.name
8        break; //end for
9      }
10   if(subSerSet=∅)//req cannot be satisfied by service in serSet
11     return nil; // composition failure
12   newreqSet:=divide(req, choosealg(ruleset)); //divides req according the selected algorithm
13   ReqSet:=ReqSet ∪ newreqSet -{req};
14   }//end for
15 }//end while
16 ComSerModel:=Generate(LabeledReq, req);

```

Fig. 8. Pseudocode of the composition service generation

We have simulated our method in a Java platform, where the worst-case time complexity is shown in table 2. In it, $lentl$, $lbehl$, $ltranl$, $leventl$, $ldata$ and $lserl$ denote the number of the behaviours, environment entities, transitions, events, data and services.

Moreover, m denotes the repetitions, and n shows the times for discovering all the requirements, and $Lev(hsm)$ denotes the average level number of all the hsm. Hence, the overall complexity of our method is at the polynomial level.

Table 2. The complexity degree of each algorithm

Algorithm	Des	Div1	Div2	Div3	Div4	Div5	Div6	Div7	Dis	Com
Worst-case complexity	$m \times lbeh^3$	$lentl^2$	$lentl^2$	$lentl^2$	$lentl$	$Lev(hsm)$	$ltran^3$	$Max(ldata, lser \times ltran, levent)$	$lser \times lbeh$	$n \times T(dis)$

5 Related Work

Web service composition is a research topic attracting attention daily. The differences between our method with others are illustrated in table 3, where I, O, P, E, B and “-” denote input, output, precondition, effect, behaviour and unspecified explicitly, respectively.

Table 3. Comparison of various approaches to service composition

Approach	Service	Request	Content	Composition method
McIlraith [4]	IOPE	IOPE	-	agent
Hamadi [9]	IO	-	-	-
Bultan [10]	B	B	conversation	behavior equivalence
Fensel [11]	IOPEB	IOPEB	service, goal	mediator
Maamar [12]	IOB	IOB	context	agent technology
Berardi [5]	B	B	-	behavior equivalence
Sirin [7]	IOPE	IOPE	service	complex service-based decomposition
Brogi [3]	IOB	IO	-	Graph-constructing and coloring
This paper	IOPEB	IOPEB	environment	Requirement-driven

6 Conclusions

This paper proposes that the essence of the composition of web services is the combination of effects of these services on their environment, and illustrates the requirement can play more important role in it. Compared with the existing efforts in this field, this work advances the state of art in the following aspects:

-The sharable environment ontology serves as a common knowledge background of both the services and the requirement. That enables the capability matching at semantic and behavior level.

-The ontology-based requirement description method reduces difficulty of requirements description as well as provides a more understandable and more expressive specification.

-The structured effect-based requirement specification prompts hierarchical effective decomposition and composition-oriented service discovery.

This paper describes an on-going work for tackling the issue of automatic service composition. In the next step, we will extend the ontology for supporting the service composition with different granularity and also in various domains. And then we will enhance the service composition procedure for considering the non-functional concerns. Moreover we will also focus on the verification of the capability profiles for the correctness of the composite services.

Acknowledgments. This work is partially supported by the National Natural Science Fund for Distinguished Young Scholars of China under Grant No.60625204, the Key Project of National Natural Science Foundation of China under Grant No. 60736015 and 90818026, the National 973 Fundamental Research and Development Program of China under Grant No. 2009CB320701.

References

1. Papazoglou, M.P., Georgakopoulos, D.: Service-Oriented Computing. *Communications of the ACM* 46(10), 25–29 (2003)
2. Nikola, M., Miroslaw, M.: Current Solutions for Web Service Composition. *IEEE Internet Computing* 8(6), 51–59 (2004)
3. Brogi, A., Corfini, S., Popescu, R.: Semantics-Based Composition-Oriented Discovery of Web Services. *ACM Transactions on Internet Technology* 8(4), 19, 1–39 (2008)
4. McIlraith, S., Son, T.C.: Adapting Golog for Composition of Semantic Web Services. In: 8th International Conference on Knowledge Representation and Reasoning, Toulouse, France, pp. 482–496 (2002)
5. Berardi, D., Calvanese, D., Giuseppe, D.G., et al.: Automatic Composition of Transition-based Semantic Web Services with Messaging. In: 31st International Conference on Very Large Databases, VLDB Endowment, Norway, pp. 613–624 (2005)
6. Puwei, W., Zhi, J., Lin, L., Guangjun, C.: Building towards Capability Specifications of Web Services Based on an Environment Ontology. *IEEE Transactions on Knowledge and Data Engineering* 20(4), 547–561 (2008)
7. Sirin, E., Parsia, B., Wu, D., Hendler, J., Nau, D.: HTN Planning for Web Service Composition using SHOP2. *Journal of Web Semantics* 1(4), 377–396 (2004)
8. Martin, D., Burstein, M., Hobbs, J., et al.: OWL-S: Semantic Markup for Web Services. The OWL Services Coalition (2004), <http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/>
9. Hamadi, R., Benatallah, B.: A Petri Net-based Model for Web Service Composition. In: 14th Australasian Database Conference, pp. 191–200. Australian Computer Society, Australia (2003)
10. Bultan, T., Fu, X., Hull, R., Jianwen, S.: Conversation Specification: A New Approach to Design and Analysis of E-Service Composition. In: 12th International World Wide Web Conference, Hungary, pp. 403–410 (2003)
11. Fensel, D., et al.: Ontology-based Choreography of WSMO Services. WSMO Final Draft (2007), <http://www.wsmo.org/TR/d14/v0.4/>
12. Mamar, Z., Mostefaoui, S.K., Yahyaoui, H.: Toward an Agent-Based and Context-Oriented Approach for Web Services Composition. *IEEE Transactions on Knowledge and Data Engineering* 17(5), 686–697 (2005)

Multi-agent and Workflow-Based Web Service Management Model

Wenjia Niu^{1,2}, Quansheng Dou³, Xu Han^{1,2}, Xinghua Yang², and Zhongzhi Shi¹

¹ Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100080, Beijing, China

² Graduate School of the Chinese Academy of Sciences, 100039, Beijing, China

³ Shandong Institute Of Business and Technology, 264005, Yantai, China

{niuwenjia, douqsl, hanxu, yangxh, shizz}@ics.ict.ac.cn

Abstract. The coordination between agent service and Web service is the key factor for intelligent Web service management in the multi-agent based Web service framework. In view of the drawbacks of existing coordination approaches for agent service and Web service, this paper proposed a multi-agent and workflow-based Web service management model. Through analyzing the interaction relations between agent service and Web service in the logical action-based task environment, a uniform task view for intelligent Web service is built. And based on such task view, a workflow towards special task is designed to realize intelligent Web service discovery and cooperation and composition. This model provides a more flexible Web service management.

Keywords: Multi-agent, Workflow, Intelligent Web Service, Task View.

1 Introduction

Web service technology has already been an interesting research problem in distributed computing field. Based on the Internet standard protocol, it aims to blend the heterogeneous applications, and realize interaction between different platforms. As the developing trend of the future Internet, its basic architecture is based on the interaction among the three roles: service provider, service broker and service requester. The service provider uses WSDL[1] to describe the service and publishes it to the service broker. The service requester submits request to the service broker, and uses SOAP[2] to invoke Web service. The service broker registers the services the provider released, and helps the requester search and assembly the services.

From the angle of role management, the traditional Web service has some shortages, e.g. lack of self-organization capabilities, possible resource conflicts during the execution and rigidified processing mode. Agent has intelligent characters such as autonomy, interaction and initiative etc., so it is very suitable to construct flexible and intelligent Web service management system with complex structures. Web service technology is complementary with agent technology, and the combination of such two technologies has now become a major research direction in the intelligent Web service field. Although some work has been done on the combination of agent and Web

service at home and abroad[3,4,5,6], but from the global aims of the intelligent Web service, those key issues about the semantics expression, service management and intelligent Web mining, have not been resolved very well. Especially in the intelligent Web service management field, how to realize flexible interactions between the Web service roles is currently a key problem which extraordinarily needs to be resolved.

In intelligent Web service management, many researchers have contributed helpful attempts and researches, and workflow as an important technology has also been tried in the coordination and composition of intelligent Web service, and typical work was described in [7,8,9,10,11]. Through the analysis of related work, there are two major methods for workflow to coordinate tasks. One is to directly coordinate Web service tasks with agent technology, while the other is to indirectly invoke Web service by coordinating agent tasks. These two methods proposed the idea of Web service management based on workflow from different angles. In fact, there exists tight inherent relevance between Web service task and agent task. Therefore, the two tasks are separated in processing, which will result in the lack of unified task views and not beneficial to the global workflow task coordination and flexible Web service management.

For the problem mentioned above, this paper built unified intelligent Web task view, and based on such view we designed the global task view oriented workflow to realize the intelligent management of Web service, which can be used to improve the global coordination and management capabilities of the Web service.

The remainder of this paper is organized as follows. Section 2 presents the intelligent Web task view. Section 3 presents the service workflow management. In Section 4, we give the intelligent Web service management model. Section 5 draws conclusions.

2 Intelligent Web Task View

2.1 Agent Task

There are three types of Agents in intelligent Web: Provider Agent (PA), Requester Agent (RA), Broker Agent (BA). PA supplies Web service, and RA requests Web service from PA, while BA helps RA to locate PA, find and combine corresponding services. The interactions between the three agents are shown in Fig. 1. From the

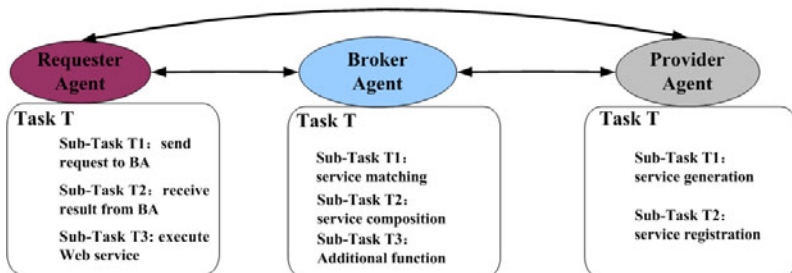


Fig. 1. Multi-agent Interaction in the Intelligent Web

angle of ultimate goal, these agents supply Web service for users by mutual negotiation. From the angle of role that the agent acts as, one agent can be either a service provider agent or a requester agent. From the angle of functional component, each agent encapsulates specific function. In the interaction process, each agent implements its own task for the global aim, and each task can be also divided into several sub-tasks, these tasks are called agent tasks in short.

Agent task environment can be formalized as a two-tuples $\langle P ; R \rangle$. $P = \{A, T, \text{Sub-T}\}$, where A represents agent, T represents abstract tasks of agent, and Sub-T represents sub-tasks after the division and instantiation of abstract task; $R = \{E, F\}$, where $E = \langle A, T \rangle$ represents the abstract task T that agent A needs to realize; $F = \langle T, \text{Sub-T} \rangle$ represents that abstract task T can be implemented by an instantiated single Sub-T or composite Sub-Ts. The above formalized description shows that it is easy to translate the invoke and operation of agent task into the function realization in the program. For example, RA.T.Sub-T2 represents that RA invokes its Sub-T T2 to obtain the results returned by BA.

2.2 Web Service Task

Web Service can be defined as a group of service access points through WSDL description, and requester can access the service through these points. The WSDL service file, first abstractly describes the access operation, the request and response messages, which will be bound into concrete transfer protocol and message format. WSDL includes service interface and service realization. Analysis from the functional structure angle, the execution of each concrete service means the implementation of one task. For example, the task of ticket booking service is to realize the users' ticket booking request, and this task can be divided into several sub-tasks, such as ticket query task and cash settlement task etc.. In addition, for the software development, Axis2[12] has already realized the bidirectional auto generation both from Java class to WSDL and from WSDL to Java class, which means that Web service is a functional entity which can execute the task just like a Java program.

Except the WSDL, the Semantic Web Service technology uses the semantics description language (e.g. OWL-S[13]) which can be understood by the computer, and enrich the service semantics with domain ontology. It aims to intelligently operate the service through inference function of the ontology. OWL-S describes a Web service by describing Service Profile, Process Model and Service Grounding, in which Service Profile describes the input, output, precondition and effect (IOPE). As description of the service function attributes, in fact IOPE also describes the task which the Web service functional entity will execute.

The above analysis shows that Web service task just means the function execution of the Web service. Take the suggested description language OWL-S of Web service as standard, with IOPE to describe the service function, the task environment of Web service can also be formalized into a two-tuples $\langle P ; R \rangle$. $P = \{S, T\}$, where S represents Web service, T represents Web service task; $R = \langle S, T \rangle$ represent the task T that Web service S will implement; and $T = \{I, O, P, E\}$, which represents that if the input and precondition are satisfied, executing task T can obtain the output and effect of the service. That is the task execution will change the current status $\{I, P\}$ into the effect status $\{O, E\}$.

2.3 Task View

From the point of view of theory and engineering implementation, the agent and Web service have already overcome the description and communication barrier in the integration aspect in JADE, which provides a basic infrastructure for the relation mapping between the agent task and Web service task. Nguyen proposed WS2JADE framework[14], which realized the encapsulation from Web service to agent service through GateWay technology. Varga discussed the compatible problem between the FIPA ACL on JADE platform and the Web service protocol, and utilized Wrapper mechanism to encapsulate the agent service of the JADE platform into the standard published Web service. Furthermore, it makes the Web service invoke agent service and return corresponding results.

Agent service description language SDLSIN[16] is the improved edition based on CDL, SDL and LARKS. It clearly defines that agent service is composed of several tasks, i.e. $AS = \bigcup_i (\text{Action:Concept})_i (i \geq 1)$, where Action is the action expression in dynamic description logic (DDL)[17]; Concept is the concept expression. Agent service executes the corresponding agent task through invoking the logic action. According to WSDL, an atomic Web service is composed of one or multi operations (tasks), i.e. $WS = \bigcup_i WS_Ti (i \geq 1)$, Shi[17] has proved that Semantic Web Service is equivalent to logic action in semantics. The execution of corresponding Web service task is equivalent to the execution of action in DDL. Through above analysis, the interaction of agent service and Web service, and the action logic of the task together set up a bridge for the unified logic view of Web service task and agent task(See Fig.2).

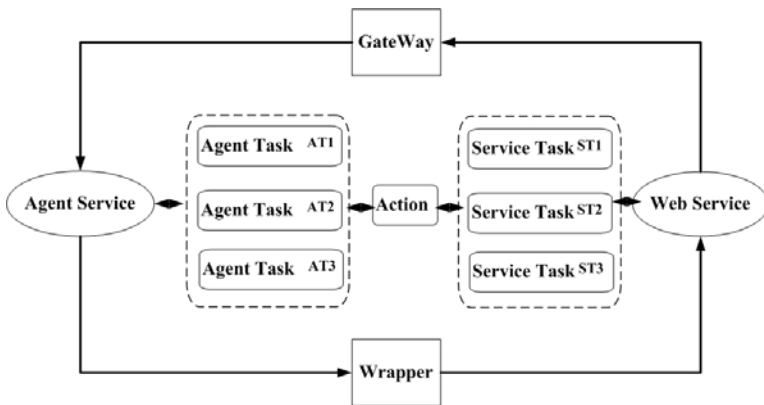


Fig. 2. Logic Relations between Agent Task And Web Service Task

In fact, agent service and Web service formed the unified task view based on action. This task environment can be formalized as a two-tuples $\langle P: R \rangle$. $P = \{AT, ST, Action\}$, where AT represents agent task, ST represents Web service task, action represents logic action of the task; $R = \{E, F, G\}$, where $E = \langle AT, Action \rangle$ represents logic action the AT will implement; $F = \langle ST, Action \rangle$ represents logic action the Web service will implement; G represents the relation between logic action π of agent task and logic action π' of Web service (See Table 1).

Table1. Relations among Logical Action of Task

	Logic Description	Explanation
Sequence(π, π')	$\pi ; \pi'$	π and π' execute sequentially, i.e. the execution effect of π is the execution precondition of π'
Choice(π, π')	$\pi \cup \pi'$	Either π or π' is chosen to execute, i.e. π is similar with π' , choose one to execute
Any-Order (π, π')	$(\pi ; \pi') \cup (\pi' ; \pi)$	π and π' execute sequentially , π' and π execute sequentially , one of the compound action $(\pi ; \pi')$ and $(\pi' ; \pi)$ is chosen to execute.

3 Service Flow Management

Workflow is a managing process which organizes a group of tasks to implements some special function. Each task is implemented by one or more programs. The trigger sequence and condition are defined in the workflow, which implements the task trigger and synchronization as well as the workflow transmission. Intelligent Web service workflow management is to realize the workflow definition and management for the task of agent service and Web service, and propel the execution of the task according to the logic defined in advance. The whole workflow management model needs to consider the following issues. First, control role design to control and supervise the execution of workflow. Second, workflow design is to realize the Web service discovery and composition, and automatic or semi-automatic invoke and design and organize the subtasks. Third, conflict resolution is to resolve the conflicts of synchronous and asynchronous cooperation among the tasks.

3.1 Design of Control Role

The control role needs to control all the resources. In intelligent Web service, BA has features to easily interact with other agent to get the global view of the resources. BA with workflow control structure aims at implementing the message transferring between the workflows as well as the task execution in the workflow, so as to intelligently control the workflow. The major design has the following five modules: communication interface module, workflow supervision module, policy repository, chief control module and status repository. Communication interface module is responsible for message transmission and communication between the workflow tasks. Currently for the message transmission mechanism on the MAGE platform, in order to avoid marshalling and unmarshalling procedure, message is coded into Java object rather than string to transmit. When passing the platform border, message is automatically transformed to FIPA compatible grammars, codes and transfer protocols. Policy repository mainly stores the policy of executing the corresponding task in the workflow according to the current status, e.g. conflict resolution policy etc.. The workflow

status supervision module mainly supervises the execution status of the task and conflict generation. The chief control module works like the nerve center to coordinate the execution of other modules. The status repository is constructed as a form, and mainly stores the execution status of the tasks.

3.2 Workflow Design

The workflow of Web service mainly includes service discovery workflow and service composition workflow. Service discovery is the precondition and foundation of service composition, so in the whole workflow, service composition generally includes discovery workflow. As shown in the right part of Fig.3 : ①agent task AT0 receives request from user, and decomposes request into several sub-goal, then distributes them to other agent task; ②Agent task invoke Web service task to realize the sub-goal based on the sequence relation between tasks; ③if agent task doesn't find suitable Web service or execution exception, then it will start other agent task and reduce the service discovery standard, then continue discovering task until find similar Web service task based on the Choice relation between tasks; ④Agent task combines the execution of Web service task; ⑤return execution results to users; ⑥if the composed Web service task executes with exceptions or errors, then ask AT0 to decompose the goal, and repeat the flow. The whole workflow shows that what the ②③ done is in fact to realize the service discovery.

The logic decomposition of tasks in step① can be done automatically by plan inference based on DDL actions. The plan inference can be simply defined as the following: if and only if action sequence satisfies goal ϕ for the ABox, RBox and ActionBox, then the action sequence is a plan to achieve the goal, the concrete inference algorithm is described in document[17]. As shown in the left part of Fig. 3, DDL inference machine generates the action plan through atomic inference. By atomic mapping algorithm and manual variable assigning mapping, finally realize atomic task workflow deployment.

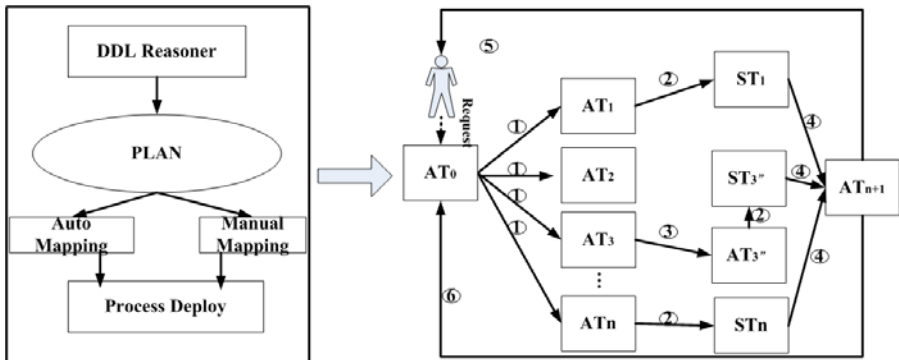


Fig. 3. Workflow Design for Web Service Composition

3.3 Conflict Resolution

For the asynchronous cooperation of the task, the conflict generally occurs in the circumstance that the resources are not released in time after task execution. We mainly discuss the conflict resolution methods for synchronous cooperation of the tasks, and coordinate these synchronous tasks so that they can get the required resources, and insure that the tasks can execute fluently. Based on above design of control roles and workflow, the simple and effective coordinate algorithm is described in the following:

(1)Workflow supervision module supervises the task execution through the communication interface. When detecting resource operating conflicts during the execution process of two tasks, supervision module will inform the chief control module to instantiate a coordinating task (Co-Task) and an inference task (R-Task).

(2)The Co-Task will collect the status and resource information of the conflicting tasks, and submit to the global status repository as a form.

(3)Combined with the information of policy repository and global status repository, the inference task will infer and compute, and return the result to the chief control module.

(4)According to the returned result, the chief module will restart one conflicting task and set waiting time for the other conflicting task through the communication interface.

4 Intelligent Web Service Management

The task-oriented intelligent Web service management model is to map the agent and Web service logic into the unified task environment, through the operation of workflow on the task, realize the service discovery and composition etc.. As shown in Fig. 4, the bottommost layer is the logic foundation of intelligent Web service, which mainly includes dynamic description logic and action theory etc., and the main goal of importing action is to represent and infer the static and dynamic knowledge in the intelligent Web according to the basic variation features of the dynamic world, and then provides the logic foundation for the unified view construction of task environment in the upper layer. The intelligent Web service description combines agent service and Web service description. Map operations of the two services into operations of tasks, so as to be invoked and managed directly by workflow. Workflow management module is composed of four parts: discovery module, negotiation module, cooperation module and composition module. Discovery module mainly judges if the providing service can satisfy the requesting service. Negotiation module mainly implements the interaction between the agents to achieve the consistent protocol by the operations on the tasks. Composition and cooperation modules are based on the service discovery, what they need to resolve is to satisfy the requester's request by the "reasonable organization" of Agent task and Web service task.

The unified-task-view-oriented workflow management module is the central part of the module. On the hand, the task abstract of agent and Web service forms the unified task view which provides a new visual angle and flexible process mechanism

for the Web service management. The workflow designers don't need to consider the complex questions such as the role variation of the agent, the isomerism of Web service as well as the interaction of agent service and Web service etc., but emphasize on how to coordinate the execution of the tasks to achieve the goal in the environment with several tasks. On the other hand, the workflow and task are two entities closely related to each other. Unified task view enriches the task type and quantity that the workflow can coordinate to a certain extent with the automatical workflow design based on inference. The workflow design is now more flexible, which makes the Web service management module press close to industrial application to a higher degree.

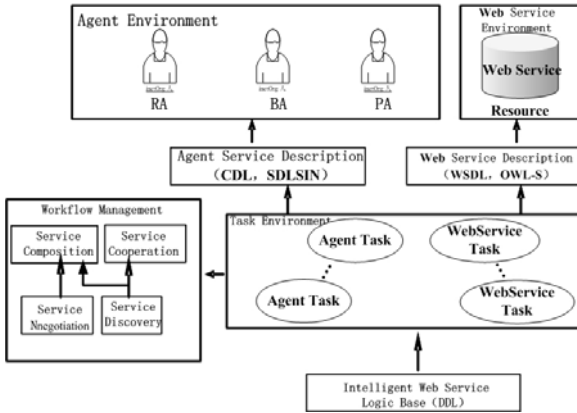


Fig. 4. Intelligent Web Service Management Model

5 Conclusion

This paper proposed a new Web service management model, by abstracting the unified task view of agent service and Web service, specific-task-oriented workflow management method was studied. But this model has many aspects which need to be improved further, e.g. further study the coordination method and complex conflict resolution method of the task, and further refine the operating steps of the workflow management etc.. In addition, refine the mapping rules of the task, and how to tightly combine with the industrial standard BPEL is the issue which needs to be studied further in the future.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No. 60775035, 60970088, 60903141), the National High-Tech Research and Development Plan of China (No. 2007AA01Z132), National Basic Research Priorities Programme(No. 2007CB311004) and National Science and Technology Support Plan (No.2006BAC08B06), Dean Foundation of Graduate University of Chinese Academy of Sciences(O85101JM03).

References

- [1] Christensen, E., Curbera, F., Meredith, G., et al.: Web services description language (WSDL) 1.1, W3C note. Tech. Rep., W3C (2001)
- [2] Box, D., Ehnebuske, D., Kajuvata, G., Layman, et al.: Simple object access protocol (SOAP) 1.1. Tech. Rep., W3C (2000)
- [3] Gibbins, N., Harris, S., Shadbolt, N.: Web Semantics: Science, Services and Agents on the World Wide Web. In: WWW 2003, pp. 141–154 (2003)
- [4] Kumar, S., Mishra, R.B.: Multi-Agent Based Semantic Web Service Composition Models. *INFOCOMP Journal of Computer Science* 7(3), 42–51 (2008)
- [5] Payne, T.R.: Web Services from an Agent Perspective. *IEEE Intelligent Systems* 23(2), 12–14 (2008)
- [6] Jia, Y., Sun, Y., Shi, Z.: Research on Service Model of Multi-agent System. *Computer Engineering* 32(22), 89–90 (2006)
- [7] Blanchet, W., Stroulia, E., Elio, R.: Supporting Adaptive Web-Service Orchestration with an Agent Conversation Framework. In: ICWS 2005, pp. 11–15 (2005)
- [8] Arpacı, A., Bener, A.: Agent Based Dynamic Execution of BPEL Documents. In: Yolum, p., Güngör, T., Gürgeç, F., Özturan, C. (eds.) *ISCIS 2005*. LNCS, vol. 3733, pp. 332–341. Springer, Heidelberg (2005)
- [9] Ricci, A., Omicini, A., Denti, E.: Virtual enterprises and workflow management as agent coordination issues. *International Journal of Cooperative Information Systems* 11(3-4), 355–379 (2002)
- [10] Buhler, P.A., et al.: Towards adaptive workflow enactment using multiagent systems. *Information Technology and Management Journal* 6(1), 61–87 (2005)
- [11] Dong, W.: Multi-agent test environment for BPEL-based web service composition. In: *ICSOC 2008*, pp. 855–860 (2008)
- [12] <http://ws.apache.org/axis2/index.html>
- [13] The OWL Services Coalition. OWL-S: Semantic Markup for Web Services (2004), <http://www.w3.org/Submission/OWL-S/>
- [14] Nguyen, T., et al.: WS2JADE: Integrating Web Service with Jade Agents, SUTICT-TR (2005)
- [15] Varga, L.Z., Hajnal, A.: Engineering Web Service Invocations from Agent Systems. In: Mařík, V., Müller, J.P., Pěchouček, M. (eds.) *CEEMAS 2003*. LNCS (LNAI), vol. 2691, pp. 626–635. Springer, Heidelberg (2003)
- [16] Jiang, Y., Shi, Z., Zhang, H., et al.: Dynamic Service Matchmaking in Intelligent Web. *Journal of Web Engineering* 2(3), 131–147 (2004)
- [17] Shi, Z., Chang, L.: Reasoning About Semantic Web Services with an Approach Based on Dynamic Description Logics. *Chinese Journal of Computers* 31(9), 1599–1611 (2008)

Semantic Approach for Service Oriented Requirements Modeling

Bin Zhao^{1,2}, Guang-Jun Cai^{1,2}, and Zhi Jin³

¹ The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

² Graduate University of the Chinese Academy of Sciences, Beijing, China
zhaobin.ict@gmail.com, caiguangj@mails.gucas.ac.cn

³ Software Engineering Institute, Peking University, Beijing, China
zhijin@sei.pku.edu.cn

Abstract. Services computing is an interdisciplinary subject that devotes to bridging the gap between business services and IT services. It is recognized that Requirements Engineering is fundamental in implementing the service oriented architecture. It takes traditional RE techniques great efforts to model business requirements and search for the appropriate services. In this paper, we propose an ontological approach to facilitate the service-oriented modeling framework. The general idea is to establishing a common semantic language to describe both the business requirements and services capabilities based on their effects on the environment. After that, we used a case study to illustrate this method and showed that substantial efforts can be spared to construct a service model from business requirements.

Keywords: Knowledge representation, ontology engineering, requirements engineering, SOA, services computing.

1 Introduction

Services computing is an interdisciplinary subject that devotes to bridging the gap between business services and IT services. It covers the body of knowledge from business process modeling to IT infrastructure implementation. Services Computing is primarily carried out through web services enabled Service Oriented Architecture (SOA), in which, service requesters invoke local or remote black-box services that are provided by various service providers. SOA can bring flexibility to business solutions by reusing services, and it is the de facto infrastructure for cloud computing. Service orientation as a business and computational paradigm is shifting the software development methodology.

To correctly develop a business solution by utilizing SOA, it is vital to first correctly model the business requirements[1]. As many systems fail due to poorly understood, or ill-conceived, or misinterpreted business requirements[2]. While some researchers tried to adapt traditional RE techniques that are commonly used in OO and component based development to model requirements for SOA, they are labor intensive and error-prone. The first issue is that they missed the point that the SOA

paradigm differs greatly from traditional software development paradigms with its emphasis on reuse and business agility[3]; and the SOA solution should be designed for change[4]. The second issue is how to establish mutual understandings between service providers and service requesters; this is more fundamental than the first issue. There are already works attempt to add semantic sugar to the description of services(e.g. OWL-S, WSDL-S, SAWSDL, etc.); these works carry the SOA forward towards an automatic and intelligent service modeling, discovery, binding, and invocation vision. Ontology in computer science functions as a knowledge representation frame[5], it is commonly used for service description, as the reusable SOA assets are domain specific. In SOA, there are atomic services and composite services. This brings up the granularity problem to service description, and it cannot be effectively handled by existing semantic approaches.

In this paper, we propose a semantic framework for describing the functional business requirements and service functionalities. This is a major improvement to environment ontology since EC4WS was proposed: we describe service functionalities from two aspects, namely, information transformation and state transition. Besides, we also make the first attempt to enable the inference mechanism for environment ontology.

This paper is organized as follows. In section 2, related semantic approaches for services description are introduced. The environment-based service ontology is given in section 3. In section 4, we used a case study to illustrate the service modeling framework. Conclusions and future works are presented in section 5.

2 Related Works

One of the characteristic of SOA is to use heterogeneous black-box services that are published by various vendors. However, web services with the same WSDL description may have dramatically different functionalities. In the academia world, semantic web services is believed to be the remedy for the service ambiguity; techniques such as ontology is recognized as the silver bullet for precise service description.

2.1 OWL-S

OWL-S[6] is the state-of-the-art service ontology, it is widely accepted in both the academia world and industry world. OWL-S utilizes OWL as its core ontology, it describes services from three aspects: service profile, service model and service grounding. The IOPE (Input, Output, Precondition, and Results) description of service profile describes services from two aspects: IO describes information transformation and PE describes state change. OWL-S allows users to use their preferred logical language to express preconditions and effects. The detailed perspective on how to interact with a service is given in *servicemodel* by viewing it as a process. The *grounding* of a service specifies the technical details on how to access a service, e.g. protocol, message format.

There are two features that make OWL-S stand out in a crowd of semantic web service techniques. The first is that it lays its semantic foundation on OWL; it is convenient to use existing OWL inference engines, such as Pellet and Jena. The second is

that it allows user to use their preferred logical expression to describe the preconditions and effects of a service process. This enables the flexibility and can be easily adopted as needed. However, this feature can also lead to confusion among users using different logical languages.

2.2 Web Services Capability Description Based on Environment Ontology

The main idea underlying this environment-based approach to web service capability description is that services can have effects on their environments, and by portraying environment changes, services capabilities descriptions could be obtained[7][8]. This idea was originally borrowed from Jackson's Problem Frames[9].

The state changes of environment entities are formally represented using hierarchical state machines. Though this makes it straight to model the granularities of services, it cannot effectively handle web services composition problems due to the inherent incomputability of state machines composition.

The ontology of this EC4WS mainly functions as some enumerations of concepts and relationships, it only enables simple inference mechanism, such as subsumption and equivalence. Though EC4WS is still in its infancy, it has the advantage of modeling business requirements in a manner easily understandable by business stakeholders; And we believe that it will be prosperous in the future.

2.3 Service-Oriented Modeling and Architecture

SOMA[10] has been used to conduct service modeling in multiple industries around the world since it was proposed in 2004. As a technical guideline that integrates SOA life-cycle management and service-oriented principles, SOMA provides a software engineering method for building end-to-end SOA solutions. The SOMA service development lifecycle includes seven phases; it is a highly structured and regular guideline for carrying out service modeling. In general, SOMA is heavyweight, top-down, model-driven iterative software development method.

SOMA had attempted to use capability patterns and solution templates to help speed up the solution specification process; however, it still heavily relies on the experience and expertise of software engineers. Hopefully, semantic techniques powered domain knowledge can be captured and reused to automate the business modeling and service identification tasks.

3 Environment-Based Ontology for Service-Oriented Architecture

In the paradigm of services computing, black-box IT Services are usually published and requested by different participants. There exist not only a semantic gap between service providers and service requesters but also a semantic gap between business requirements and IT service capabilities[11]. The notion of environment-base software engineering, proposed by Jackson and Zave[9], is considered to be a "silver bullet" for bridging the above two gaps.

3.1 Principle of Environment-Based RE Methodology

The environment of software includes everything but the software itself. According to Michael J., requirements are located in the environment and the software system is to be used within a specific environment. One principle of environment-based RE is that the interactions between a system and its environment are the interfaces between the system and the environment. By observing the shared phenomenon between software and its environment[12], one can indirectly infer the functionalities (capabilities) of the software. In general, appropriate capabilities are required to realize a business goal, whilst the software exposes some capabilities. Thus, we can use the capability of both ends to bridge the gap between requirements and software functionalities.

Another principle is that while the software may change frequently, its environment stays relatively unchanged. This makes it suitable for requirements engineering for the service-oriented paradigm, because the SOA paradigm is characteristic of loose coupling, reuse, flexibility, and constantly needs to react to business changes. In contrast to that, IT services are frequently subject to extensions and modifications.

3.2 Environment Ontology for Services Computing

IT services have two aspects of functionalities: information transformation or state changes to the environment (Specifically, state changes of some environment entities). For example, to finish a sales order, credit card number is required as input, the status of the transaction is the output information, as a result, the credit card is charged and the ownership of the product transfers from the seller to the buyer[6].

We use ontology to formally model the semantics of a service's functionality, and to cope with its semantic heterogeneities and interface ambiguities. The main concept of environment ontology is *environment entity*. By an environment entity we mean some independent unit of being that is identifiable from the environment[13]. It can be abstract or concrete, such as a message, an event, a record in database, a book (conceptually or physically), etc.

Conceptual Model of Environment-based Ontology

This section specifies the ontology model for representing domain specific environment entities. The model is built on-top of OWL2 using the OWL2 datatypes and vocabularies. The OWL2 datatype map is a 6-tuple:

$$D ::= \langle N_{DT}, N_{LS}, N_{FS}, \cdot^{DT}, \cdot^{LS}, \cdot^{FS} \rangle, \text{ where}$$

- N_{DT} is a set of names of datatypes.
- N_{LS} is a function that assigns each datatype $DT \in N_{DT}$ the lexical form of strings.
- N_{FS} is the constraining facet of values of the form (F, v) , where F is the constraining facet and v is a value.
- The interpretation function \cdot^{DT} assigns each $DT \in N_{DT}$ a value space.
- The interpretation function \cdot^{LS} assigns each $DT \in N_{DT}$ a lexical form.
- The interpretation function \cdot^{FS} assigns each $DT \in N_{DT}$ a constraint $\langle F, v \rangle$.

Using the above notion, a datatype *NaturalNumber* has the name “Natural Number, with the lexical form “Integer” and constraint $\langle \text{minValue}, 0 \rangle$.

The vocabulary of the environment ontology is a 7-tuple over a datamap D :

$V ::= \langle V_C, V_I, V_{OP}, V_{DP}, V_{DT}, V_{AN}, V_{LF} \rangle$, where

- V_C is a set of classes that contains at least *owl:Thing* and *owl:Nothing*
- V_I is a set of individuals used to represent a specific environment entity, individual can be named or anonymous.
- V_{OP} is a set of object properties; this can be used to define the mereology of environment entity.
- V_{DP} is a set of datatype properties
- V_{DT} is a set of datatypes containing D .
- V_{AN} is a set of annotations that are used to comment purpose.
- V_{LF} is the vocabulary of a logic language that is used to construct the expression of constraints and axioms.

Given the datatype map D and vocabulary V , the environment ontology is defined as a 3-tuple as follows:

$EnvOnt ::= \langle Entities, Expression, Rel, Axiom \rangle$

- $Entities \subseteq V_C \cup V_I$ is the set of identified environment entities. For a specific domain, the set of entities may vary.
- $Expression \subseteq Entities \times (V_{OP} \cup V_{DP} \cup V_{DT})$, the expression describes the properties of an environment entity.
- Rel is the set of relations between entities, $Rel \subseteq Entities \times Entities$. The details of these relations are given in section 3.2.2.
- $Axiom \subseteq V \times V \cup Entities \times Entities \cup Entities \times Expression$. Axioms are used for inference purpose, e.g., discover new relationships, automatically analyzing the control of the data, and discovering possible inconsistencies.

The environment entities identified are domain specific and have granularities. For example, in a census application, the entity “family” is composed of concrete family members. In contrast, in the domain of healthcare, human beings are view as composed of different part. This also has the characteristic of granularities.

The granularity of an environment entity is captured using the tree structure of ontology concepts. By a *composite entity*, we mean some entity that can be further decomposed into different parts. The constituent entities are called its *sub-entity*. *Atomic entity* is not decomposable. In the hierarchy of entity mereology, atomic entity lies at the bottom layer. The property of an atomic entity is totally determined by its properties, whilst the property of a composite entity is determined by both its attributes and its mereology (the way how it is composed).

Axioms and Relation of Environment Ontology

The semantics of the environment ontology is defined by its interpretation. In this section we give the fundamental interpretations of properties, relationships, and axioms. They are subject to extension depending on the purpose of the knowledge engineer and the characteristics of the modeling domain.

The basic *EnvOnt* building blocks are its classes and individuals. The axioms about the relationships between classes are shown in table 1.

Table 1. Class Relationship Expression

Class Axiom	Interpratation	Description
<i>SubClassof</i> (C_1, C_2)	$C_1 \subseteq C_2$	C_1 is the subclass of C_2
<i>EquivalentClass</i> (C_1, C_2)	$C_1 \subseteq C_2 \cap C_1 \supseteq C_2$	C_1 and C_2 are semantically equivalent
<i>ClassIntersection</i> (C_1, C_2)	$\{x x \in C_1 \wedge x \in C_2\}$	It is the class that are the intersection of C_1 and C_2
<i>ClassUnion</i> (C_1, C_2)	$C_1 \cup C_2$	It is used for concept extension
<i>ComplementClass</i> (C_1, C_2)	$\{x x \in C_1 \wedge x \notin C_2\}$	It is the set of concepts that belong to C_1 but not C_2

For an environment entity, its associations with other entities are specified by the object property axioms and data property axioms. A sample of the typical property axioms are shown in table 2. These are not the complete property axioms due to lack of space.

The *state* of an environment entity is the snapshot of its properties. It is a summary of the entity description. The state changes if some of the properties of an individual changes. Following the principle of environment-based RE methodology in section 3.1, the software's functionality can be observed by the state changes of some environment entities.

The environment ontology *EnvOnt* can be used to describe service capabilities in SOA. More detail on how to describe the services will be given in section 3.2.3. It can also be utilized to represent knowledge of an application domain, or to describe resources etc.

Table 2. Entity property axioms

Property Axiom	Description
$HasA(I_1, I_2)$	Used to describe the whole-part relation or specify attributes, se.g., $HasA(Person, Age)$ means $Person$ has Age attribute.
$IsA(I, C)$	Individual I is an instantiation of class C .
$DataProperty(I, DPE)$	DPE is the data property expression. Data from I must satisfy DPE .
$Cardinality(I, DPE)$	For a property specified by DPE , constraint its cardinality.
$ObjectProperty(I, OPE)$	Individual I satisfies the property OPE .
$HasAnnotation(I, An)$	Annotation is used to provide further information that is not part of the ontology.

Using environment ontology to model service capability

The SOA services are black-box units of functionality that are published by different service providers. The traditional syntactical interface description cannot handle the inherent heterogeneity of service semantics. *EnvOnt* is purposed to handle this heterogeneity by providing the semantic terms that can be used by both ends.

A service accepts some messages, and then performs some functionality accordingly or sends out some messages. This can be categorized to information transformation and state changes of its environment. Formally, we define the capability of services as a 3-tuple:

$SerCap ::= \langle (Input, Output)?, Conds, StateTrans^* \rangle$, where

- *Input* is the message that service requires.
- *Output* is the message that service generates as response to *Input*. A service can have zero or exact one IO pairs, this is denoted by the question mark“?”. When errors happen, an exception is generated as *Output*.
- *Conds* is the logical expressions that specify conditions under which a service can perform its functionality.
- *StateTrans* describes the semantics of services by means of its effect on the environment entities. $StateTrans ::= \{(\Delta Entity.property)^+\}$, this means that a service can effect on more than one entities, and causes their properties to change.

The message interchanged is a 4-tuple, as defined below:

$Message ::= \langle Sender, Receiver, Content, Type \rangle$, where

- $Sender \in Entities$ and $Receiver \in Entities$ are the participants in a message exchange. The sender and receiver can be services, users, or other applications.
- *Content* includes the parameters to invoke a service, or the results generated by services.
- $Type = one-way \mid broadcast \mid request-response$, this specifies the patterns of message interchange.

Based on the above definition of *message* and *service capability*, a services is described as a 4-tuple,

$Serv ::= \{Des, Process, SerCap, Groundings\}$, where

- *Des* uses natural language and annotations from V_{AN} to describe service profiles, this is used for service discovery.
- *Process* specifies the collaborations between service partners. For a composite service, it specifies how its component services choreograph.
- *SerCap* as defined above is for the service functionality description.
- *Groundings* provides the detail on integrating with existing IT standards.

In reality, there are various web services, e.g., VoIP, VOD (information transformation), online docs (state changes), e-business, etc. Quite often there are services that have the same functionality, e.g., their effects on environment are indistinguishable. We call these services functionally **isomorphism**.

4 Case Study: Online Shopping

We use online shopping to demonstrate the environment-based service oriented requirements modeling technique. The modeling process starts by identifying the environment entities. However, one may, with different purposes in mind, look at a system quite differently. So, this is the place where domain knowledge can function as vocabularies that help different users unify their terminology.

In this case, the entities are Customer, PurchaseOrder, Invoice, Manifest, Schedule, Shipping, InvoiceMessage, and POMessage. PurchaseOrder has one Invoice, one or more Manifests, and one or more Schedule as its sub-entity. PrurchaseOrder has the property ‘id’ of type string, ‘totalPrice’ of type Integer, and property ‘priority’ of type Integer. As illustrated in Figure 1.

The capability requirements for Purchasing Service is that it can receive the POMessage, calculate the price of Invoice, and schedule the shipping according to

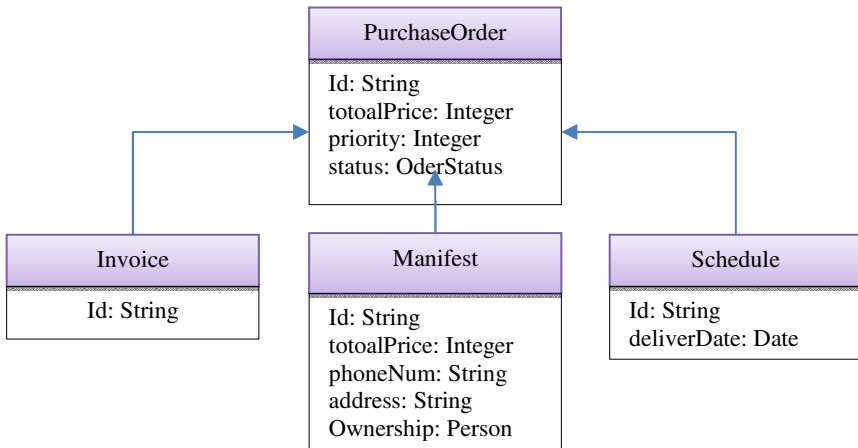


Fig. 1. PurchaseOrder Entity

PurchaseOrder.Schedule. This changes the PurchaseOrder's *totalPrice* property to some amount with constrains greater than 0. When the order is processed, the ownership of the manifests is transferred to the customer, and the *status* property of this PurchaseOrder instance is changed to "Processed and Done".

The example illustrates how to capture the business requirements, and the remaining work is to look up the service repository to discovery available services that expose the required capabilities. This work can be facilitated by the inference mechanism of the environment ontology for service description.

5 Conclusions and Future Work

Requirements engineering for services computing have different features compared to traditional OO or component-based RE. It is an interdisciplinary subject that still in its infancy. In this paper we proposed an environment-based semantic model for service capability modeling.

This work builds a semantic model based on environment ontology to establishing a common semantic description for both the business service capability requirements and services capabilities. We use an online shopping case study to illustrate how to capture business requirements using the environment-based method. It can be seen that, now we can model the service functionalities using the semantics provided by environment ontology. As a result, substantial efforts can be spared to construct a service model from business requirements.

In the future, we will extend this modeling framework to model the non-functional requirements. The quality of a business solution depends on both its functional requirements and quality requirements.

Acknowledgement

This work was supported by the National Natural Science Foundation for Distinguished Young Scholars of China (Grant No. 60625204), the National Basic Research Program of China (Grant No. 2009CB320701), the Key Projects of National Natural Science Foundation of China (Grant Nos. 90818026, 60736015).

Appreciation also goes to our research team; this work would not have been possible without their generous contribution of our research team.

References

1. Michael, B.: Service-oriented modeling: Service Analysis, Design and Architecture. John Wiley & Sons, Inc., Chichester (2008)
2. Gary, C., Eric, N.: The Value of Modeling, A technical discussion of software modeling, <http://www.ibm.com/developerworks/rational/library/6007.html>
3. Erl, T.: SOA Principles of Service Design. Prentice Hall, Englewood Cliffs (2007)
4. Ian, G.: Requirements Modeling and Specifications for Service Oriented Architecture. Wiley, Chichester (2008)

5. Nicola, G.: Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human-Computer Studies* 43, 625–640 (2005)
6. OWL-S: Semantic Markup for Web Services, <http://www.w3.org/Submission/OWL-S/>
7. Tsai, W.T., Zhi, J., Xiaoying, B.: Internetware Computing: Issues and Perspective. In: *Internetware 2009*, October 17-18, Beijing, China (2009)
8. Puwei, W., Zhi, J., Hongyan, L.: Capability Description and Discovery of Internetware Entity. *Science China* 53(4), 685–703 (2010)
9. Pamela, Z., Michael, J.: Four dark corners of requirements engineering. *ACM Transactions on Software Engineering and Methodology* 6(1), 1–30 (1997)
10. Ali, A., Shuvanker, G., Abdul, A., Tina, A., Sella, G., Kerrie, H.: SOMA: A Method for Developing Service-oriented Solutions. *IBM System Journal* 47(3), 377–396 (2008)
11. Liang-Jie, Z., Jia, Z., Hong, C.: *Services Computing*. Springer-Verlag GmbH & Tsinghua University Press (2007)
12. Dines, B.: *Software Engineering: Abstraction and Modeling*. Springer, Heidelberg (2006)
13. Saltzer, J.H., Frans Kaashoek, M.: *Principles of Computer System Design—An Introduction*. Elsevier, Amsterdam (2009)

Extend Atomic Action Definitions of DDL to Support Occlusions and Conditional Post-conditions

Liang Chang¹, Zhongzhi Shi², and Tianlong Gu¹

¹ School of Computer and Control, Guilin University of Electronic Technology, Guilin, 541004, China

² The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100080, China
{changl,cctlgu}@guet.edu.cn, shizz@ics.ict.ac.cn

Abstract. The dynamic description logic *DDL* provides a kind of action theories based on description logics (DLs). Compared with another important DL-based action formalism constructed by Baader et.al., a shortcoming of *DDL* is the absence of occlusions and conditional post-conditions in the description of atomic actions. In this paper, we extend atomic action definitions of *DDL* to overcome this limitation. Firstly, we introduce extended atomic action definitions in which the occlusions and conditional post-conditions are incorporated. Secondly, for each atomic action specified by an extended atomic action definition, a function named *Expand* is introduced to transform it into a choice action which is composed of atomic actions defined by ordinary atomic action definitions. Finally, based on the *Expand* function, the satisfiability-checking algorithm of *DDL* is extended to support occlusions and conditional post-conditions.

1 Introduction

Description logics (DLs) are well-known for representing and reasoning about knowledge of static application domains. The main strength of description logics is that they offer considerable expressive power going far beyond propositional logic, while reasoning is still decidable.

The study of integrating description logics with action formalisms is driven by two factors. One is the demand to represent and reason about semantic web services [7]. Another factor is the fact that there is an expressive gap between existing action formalisms: they are either based on first- or higher-order logics and do not admit decidable reasoning, like the Situation Calculus [9] and the Fluent Calculus [11], or are decidable but only propositional, like those based on propositional dynamic logics [5] or propositional temporal logics [2].

One approach to integrate description logics with action formalisms was proposed by Baader et.al. [1]. That approach is characterized by constructing action formalisms over description logics of the *ALCQIO* family. In that formalism, acyclic TBoxes and ABox assertions of description logics are used to specify the

domain constraints and the states of the world respectively. Each atomic action is described as a triple $(pre, occ, post)$, where the set pre is composed of ABox assertions for specifying the pre-conditions under which the action is applicable; the set $post$ is composed of conditional post-conditions of the form φ/ψ with φ an ABox assertion and ψ a primitive literal, each conditional postcondition φ/ψ says that if φ is true before executing the action, then ψ should be true after the executions; the set occ is composed of oclussions for indicating these primitive literals that can change arbitrarily as while as the action is executed, where each oclussion is of the form $A_i(p)$ or $r(p, q)$, with A_i a primitive concept name and r a role name. The semantics of atomic actions is defined according to the minimal-change semantics; each atomic action is defined as a transition relation on DL-interpretations. Taking each finite sequence of atomic actions as a composite action, Baader et.al. investigated the executability problem and the projection problem of actions, and demonstrated that both of them can be reduced to standard inference problems of description logics and therefore were remained decidable.

A limitation of Baader et.al.'s formalism is that atomic actions can only be organized as finite sequences. Many complex control structures required by Web services [8], such as the "Choice", "Any-Order", "Iterate", "If-Then-Else", "Repeat-While" and "Repeat-Until" structures specified in the OWL-based Web service ontology OWL-S [6], are not supported by it.

Another typical approach to integrate description logics with action formalisms was proposed by Shi et.al. [10]. That approach is characterized by constructing a kind of dynamic description logics named *DDL*, which is in fact a combination of description logics, propositional dynamic logics and action formalisms. In that approach, domain knowledge of each action theory is captured by acyclic TBoxes of description logics; based on these knowledge, both the states of the world and the pre- and post-conditions of each atomic action are described by ABox assertions. Starting from atomic actions and ABox assertions, complex actions are constructed with the help of regular program constructors of propositional dynamic logics, so that not only the sequence structure, but also the "Choice", "Any-Order", "Iterate", "If-Then-Else", "Repeat-While" and "Repeat-Until" structures required by Web services are all supported by the formalism. Finally, both atomic actions and complex actions are used as modal operators to construct formulas, so that properties on actions can be stated explicitly by formulas. Chang et.al. [4] provided a terminable, sound and complete algorithm for checking the satisfiability of *DDL*-formulas; based on that algorithm, reasoning tasks on the realizability of actions, the executability of actions and the consequence of executing actions can all be effectively carried out [3].

Compared with Baader et.al.'s formalism, a merits of *DDL* is the capability of representing complex actions. However, a shortcoming of *DDL* is that oclussions and conditional post-conditions are not supported in the description of atomic actions. In this paper, we extend atomic action definitions of *DDL* to include oclussions and conditional post-conditions.

The rest of this paper is organized as follows. A brief introduction of *DDL* is presented in Section 2. The atomic action definitions of *DDL* is extended to include occlusions and conditional post-conditions in Section 3. Section 4 provides a satisfiability-checking algorithm for *DDL*-formulas in the case that occlusions and conditional post-conditions are embraced in the description of atomic actions. Section 5 concludes the paper.

2 The Dynamic Description Logic *DDL*

As a kind of dynamic description logics, *DDL* is constructed by embracing an action theory into description logics. Be corresponding to the family of description logics, *DDL* is embodied as different logic systems. In this section, we take the description logic *ALCQIO* as an example and introduce the dynamic description logic constructed over it.

Primitive symbols of the logic $DDL(ALCQIO)$ are a set N_I of individual names, a set N_R of role names, a set N_C of concept names, and a set N_A of action names. Basic citizens of $DDL(ALCQIO)$ are roles, concepts, actions and formulas; all of them are defined inductively by constructors starting from primitive symbols.

Roles of $DDL(ALCQIO)$ are formed according to the following syntax rule:

$$R ::= R_i \mid R^-$$

where $R_i \in N_R$.

Concepts of $DDL(ALCQIO)$ are constructed according to the following syntax rule:

$$\begin{aligned} C, C' ::= & A_i \mid \neg C \mid C \sqcup C' \mid C \sqcap C' \\ & \mid \forall R.C \mid \exists R.C \mid \leq nR.C \mid \geq nR.C \mid \{p\} \end{aligned}$$

where $A_i \in N_C$, $p \in N_I$, and R is role.

A *concept definition* is of the form $A \equiv C$, where A is a concept name and C is a concept. A *TBox* of $DDL(ALCQIO)$ is a finite set of concept definitions with unique left-hand sides. A TBox is said to be *acyclic* if there are no cyclic dependencies between the definitions.

With respect to an acyclic TBox \mathcal{T} , a concept name $A_i \in N_C$ is called *defined* if and only if it occurs on the left-hand side of some concept definition contained in \mathcal{T} , and is called *primitive* otherwise.

Formulas of $DDL(ALCQIO)$ are formed according to the following syntax rule:

$$\varphi, \varphi' ::= C(p) \mid R(p, q) \mid \langle \pi \rangle \varphi \mid [\pi] \varphi \mid \neg \varphi \mid \varphi \vee \varphi' \mid \varphi \wedge \varphi'$$

where $p, q \in N_I$, R is a role, C is a concept and π is an action.

An *ABox assertion* is of the form $C(p)$, $R(p, q)$ or $\neg R(p, q)$, where $p, q \in N_I$, C is a concept, and R is a role. A finite set of ABox assertions is called an *ABox* of $DDL(ALCQIO)$.

With respect to an acyclic TBox \mathcal{T} , an ABox assertion ψ is called a *primitive literal* if and only if it is of the form $A_i(p)$, $(\neg A_i)(p)$, $R(p, q)$ or $\neg R(p, q)$, with A_i a primitive concept name, R a role and $p, q \in N_I$.

Actions of $DDL(ALCQIO)$ are formed according to the following syntax rule:

$$\pi, \pi' ::= \alpha \mid \varphi? \mid \pi \cup \pi' \mid \pi; \pi' \mid \pi^*$$

where $\alpha \in N_A$, and φ is an ABox assertion.

With respect to an acyclic TBox \mathcal{T} , an *atomic action definition* of $DDL(ALCQIO)$ is of the form $\alpha \equiv (P, E)$, where

- $\alpha \in N_A$,
- P is a finite set of ABox assertions for describing the pre-conditions, and
- E is a finite set of primitive literals for describing the post-conditions.

An *ActBox* of $DDL(ALCQIO)$ is a finite set of atomic action definitions with unique left-hand sides.

An atomic action α is said to *be defined in an ActBox* \mathcal{A}_C if and only if α occurs on the left-hand side of some atomic action definition contained in \mathcal{A}_C . A formula φ is said to *be defined w.r.t. an ActBox* \mathcal{A}_C if and only if all the atomic actions occurring in φ are defined in \mathcal{A}_C .

A *knowledge base* of $DDL(ALCQIO)$ is of the form $K = (\mathcal{T}, \mathcal{A}_C, \mathcal{A})$, where \mathcal{T} , \mathcal{A}_C and \mathcal{A} are respectively a TBox, an ActBox and an ABox.

The semantic model of $DDL(ALCQIO)$ is of the form $M = (W, T, \Delta, I)$, where,

- W is a non-empty set of states;
- $T : N_A \rightarrow 2^{W \times W}$ is a function which maps action names into binary relations on W ;
- Δ is a non-empty set of individuals; and
- I is a function which associates with each state $w \in W$ a DL-interpretation $I(w) = \langle \Delta, \cdot^{I(w)} \rangle$, where the function $\cdot^{I(w)}$
 - maps each concept name $A_i \in N_C$ to a set $A_i^{I(w)} \subseteq \Delta$,
 - maps each role name $R_i \in N_R$ to a binary relation $R_i^{I(w)} \subseteq \Delta \times \Delta$, and
 - maps each individual name $p \in N_I$ to an element $p^{I(w)} \in \Delta$, with the constraints that $p^{I(w)} = p^{I(w')}$ for any state $w' \in W$, and $p^{I(w)} \neq q^{I(w)}$ for any individual name q which is different from p . Since interpretations of p are the same in every state, the interpretation $p^{I(w)}$ is also represented as p^I .

Given a model $M = (W, T, \Delta, I)$, the semantics of concepts, formulas and actions of $DDL(ALCQIO)$ are defined inductively as follows.

Firstly, with respect to any state $w \in W$, each role R will be interpreted as a binary relation $R^{I(w)} \subseteq \Delta \times \Delta$, and each concept C will be interpreted as a set $C^{I(w)} \subseteq \Delta$; the definition is as follows:

1. $(R^-)^{I(w)} = \{(y, x) \mid (x, y) \in R^{I(w)}\}$;
2. $(\neg C)^{I(w)} = \Delta \setminus C^{I(w)}$;

3. $(C \sqcup D)^{I(w)} = C^{I(w)} \cup D^{I(w)}$;
4. $(C \sqcap D)^{I(w)} = C^{I(w)} \cap D^{I(w)}$;
5. $(\forall R.C)^{I(w)} = \{x \in \Delta \mid \text{for all } y \in \Delta: \text{if } (x, y) \in R^{I(w)}, \text{ then } y \in C^{I(w)}\}$;
6. $(\exists R.C)^{I(w)} = \{x \in \Delta \mid \text{there is a } y \in \Delta \text{ with } (x, y) \in R^{I(w)} \text{ and } y \in C^{I(w)}\}$;
7. $(\leq nS.C)^{I(w)} = \{x \in \Delta \mid \#\{y \in \Delta \mid (x, y) \in S^{I(w)} \text{ and } y \in C^{I(w)}\} \leq n\}$;
8. $(\geq nS.C)^{I(w)} = \{x \in \Delta \mid \#\{y \in \Delta \mid (x, y) \in S^{I(w)} \text{ and } y \in C^{I(w)}\} \geq n\}$;

Secondly, for any formula φ and any state $w \in W$, the truth-relation $(M, w) \models \varphi$ is defined inductively as follows:

9. $(M, w) \models C(p)$ iff $p^I \in C^{I(w)}$;
10. $(M, w) \models R(p, q)$ iff $(p^I, q^I) \in R^{I(w)}$;
11. $(M, w) \models \langle \pi \rangle \varphi$ iff there is a state $w' \in W$ with $(w, w') \in T(\pi)$ and $(M, w') \models \varphi$;
12. $(M, w) \models [\pi]\varphi$ iff for every $w' \in W$: if $(w, w') \in T(\pi)$ then $(M, w') \models \varphi$;
13. $(M, w) \models \neg\varphi$ iff it is not the case that $(M, w) \models \varphi$;
14. $(M, w) \models \varphi \vee \psi$ iff $(M, w) \models \varphi$ or $(M, w) \models \psi$;
15. $(M, w) \models \varphi \wedge \psi$ iff $(M, w) \models \varphi$ and $(M, w) \models \psi$;

Finally, each action π is interpreted as a binary relation $T(\pi) \subseteq W \times W$ according to the following definitions:

16. $T(\varphi?) = \{(w, w) \mid w \in W \text{ and } (M, w) \models \varphi\}$;
17. $T(\pi \cup \pi') = T(\pi) \cup T(\pi')$;
18. $T(\pi; \pi') = \{(w, w') \mid \text{there is a state } u \in W \text{ with } (w, u) \in T(\pi) \text{ and } (u, w') \in T(\pi')\}$;
19. $T(\pi^*) = \text{reflexive and transitive closure of } T(\pi)$.

Let $K = (\mathcal{T}, \mathcal{A}_C, \mathcal{A})$ be a knowledge base and let $M = (W, T, \Delta, I)$ be a semantic model, then:

- a state w of the model M satisfies the ABox \mathcal{A} , denoted by $(M, w) \models \mathcal{A}$, if and only if $(M, w) \models \varphi_i$ for every ABox assertion $\varphi_i \in \mathcal{A}$;
- M is a model of the TBox \mathcal{T} , denoted by $M \models \mathcal{T}$, if and only if $A^{I(w)} = C^{I(w)}$ for every state $w \in W$ and every concept definition $A \equiv C \in \mathcal{T}$; and
- M is a model of the ActBox \mathcal{A}_C , denoted by $M \models \mathcal{A}_C$, if and only if the following equation holds for every atomic action definition $\alpha \equiv (P, E) \in \mathcal{A}_C$:

$$\begin{aligned}
T(\alpha) = \{ & (w, w') \mid w \in W, w' \in W, (M, w) \models P, \\
& C^{I(w')} = (C^{I(w)} \cup \{p^I \mid C(p) \in E\}) \setminus \{p^I \mid (\neg A)(p) \in E\} \text{ for} \\
& \text{each concept name } A \text{ which is primitive w.r.t. } \mathcal{T}, \text{ and} \\
& R^{I(w')} = (R^{I(w)} \cup \{(p^I, q^I) \mid R(p, q) \in E\}) \setminus \{(p^I, q^I) \mid \neg R(p, q) \in E\} \\
& \text{for each role name } R.\}.
\end{aligned}$$

A primary inference problem for $DDL(ALCQIO)$ is to decide the satisfiability of formulas. A formula φ is called *satisfiable* w.r.t. a TBox \mathcal{T} and an ActBox \mathcal{A}_C if and only if there exists a model $M = (W, T, \Delta, I)$ and a state $w \in W$ such that $M \models \mathcal{T}$, $M \models \mathcal{A}_C$ and $(M, w) \models \varphi$.

In the literature, a terminable, sound and complete algorithm for deciding the satisfiability of $DDL(ALCQIO)$ -formulas is presented [\[4\]](#).

3 Extended Atomic Action Definitions

In this section, we extend atomic action definitions of *DDL* to include occlusions and conditional post-conditions. To be distinguished from original atomic action definitions discussed in *DDL*, we refer to these extended definitions as extended atomic action definitions.

With respect to an acyclic TBox \mathcal{T} , an *extended atomic action definition* is of the form $\alpha \equiv (P, O, E)$, where

- $\alpha \in N_A$;
- P is a finite set of ABox assertions for describing the pre-conditions;
- O is a finite set of occlusions of the form $A(p)$ or $r(p, q)$, with A primitive concept name, r role name, and $p, q \in N_I$; and
- E is a finite set of conditional post-conditions of the form φ/ψ , where φ is an ABox assertion and ψ is a primitive literal.

Intuition of the above definition is as follows. The pre-conditions specify under which conditions the action is applicable. Each conditional postcondition φ/ψ says that, if φ is true before executing the action, then ψ should be true after the execution. The occlusions indicate those primitive literals that can change arbitrarily as while as the action is executed.

The semantics of extended atomic action definitions is defined as follows: a semantic model $M = (W, T, \Delta, I)$ satisfies an extended atomic action definition $\alpha \equiv (P, O, E)$, in symbols $M \models \alpha \equiv (P, O, E)$, if and only if

$$T(\alpha) = \{ (w, w') \in W \times W \mid (M, w) \models P, \text{ both } A_w^+ \cap A_w^- = \emptyset \text{ and } A^{I(w')} \cap I_A^w = ((A^{I(w)} \cup A_w^+) \setminus A_w^-) \cap I_A^w \text{ for each concept name } A \text{ which is primitive w.r.t. } \mathcal{T}, \text{ and both } R_w^+ \cap R_w^- = \emptyset \text{ and } R^{I(w')} \cap I_R^w = ((R^{I(w)} \cup R_w^+) \setminus R_w^-) \cap I_R^w \text{ for each role name } R. \},$$

where A_w^+ , A_w^- , I_A^w , R_w^+ , R_w^- and I_R^w denote some sets constructed as follows:

- $A_w^+ := \{ p^I \mid \varphi/A(p) \in E \text{ and } (M, w) \models \varphi \}$,
- $A_w^- := \{ p^I \mid \varphi/(\neg A)(p) \in E \text{ and } (M, w) \models \varphi \}$,
- $I_A^w := (\Delta \setminus \{ p^I \mid A(p) \in O \}) \cup A_w^+ \cup A_w^-$,
- $R_w^+ := \{ (p^I, q^I) \mid \text{there is some role } S \text{ with } S \sqsubseteq_{\mathcal{R}}^* R, \varphi/S(p, q) \in E \text{ and } (M, w) \models \varphi \}$,
- $R_w^- := \{ (p^I, q^I) \mid \text{there is some role } S \text{ with } R \sqsubseteq_{\mathcal{R}}^* S, \varphi/\neg S(p, q) \in E \text{ and } (M, w) \models \varphi \}$,
- $I_R^w := (\Delta \times \Delta) \setminus \{ (p^I, q^I) \mid R(p, q) \in O \} \cup R_w^+ \cup R_w^-$.

Be similar with the semantics of atomic action definitions, this definition is also based on the minimal-change semantics [12]. For any pair $(w, w') \in T(\alpha)$, any primitive concept name A and any role name R , it must be $A_w^+ \subseteq A^{I(w')}$, $A_w^- \cap A^{I(w')} = \emptyset$, and nothing else changes from $A^{I(w)}$ to $A^{I(w')}$ with the possible exception of the occluded literals. Similarly, the interpretations $R^{I(w)}$ and $R^{I(w')}$

should satisfy that $R_w^+ \subseteq R^{I(w')}$, $R_w^- \cap R^{I(w')} = \emptyset$, and nothing else changes from $R^{I(w)}$ to $R^{I(w')}$ with the possible exception of the occluded literals. Those parts that might change arbitrarily by the presence of occluded literals are captured by the set I_A^w and the set I_R^w .

For example, consider a Web service system in which customers are able to buy books online, a Web service for the customer *Tom* to buy the book *KingLear* can be described according to the following extended atomic action definition:

$$\begin{aligned}
 & \text{BuyBook}_{Tom, KingLear} \\
 \equiv & (\{ \text{customer}(Tom), \text{book}(KingLear) \}, \{ \}, \\
 & \{ \text{instore}(KingLear)/\text{bought}(Tom, KingLear), \\
 & \text{instore}(KingLear)/\neg\text{instore}(KingLear), \\
 & \text{instore}(KingLear)/\text{notify}(Tom, \text{NotifySucceed}), \\
 & \neg\text{instore}(KingLear)/\text{notify}(Tom, \text{NotifyBookOutOfStock}) \})
 \end{aligned}$$

According to this definition, if the book *KingLear* is in store, then the formula $\text{bought}(Tom, KingLear)$, $\neg\text{instore}(KingLear)$ and $\text{notify}(Tom, \text{NotifySucceed})$ will become true after the execution of the service, otherwise *Tom* will be notified by the notification *NotifyBookOutOfStock*.

4 Reasoning Mechanisms for Extended Atomic Action Definitions

In order to provide reasoning services for extended atomic action definitions, for any atomic action α defined by some extended atomic action definition $\alpha \equiv (P, O, E)$, we introduce a procedure $Expand(\alpha)$ to transform it into some action of the form $\alpha_1 \cup \dots \cup \alpha_n$, where each α_i ($1 \leq i \leq n$) is an atomic action defined by an atomic action definition.

More precisely, for the definition $\alpha \equiv (P, O, E)$, let $O = \{ \phi_1, \dots, \phi_m \}$ and let $E = \{ \varphi_1/\psi_1, \dots, \varphi_k/\psi_k \}$, then the procedure $Expand(\alpha)$ operates according to the following steps:

1. Construct an empty set \mathcal{A}_C on atomic action definitions.
2. According to the set P and these k conditional post-conditions contained in E , construct 2^k atomic action definitions as follows:

$$\begin{aligned}
 \alpha_0 & \equiv (P \cup \{ \varphi_k^-, \dots, \varphi_3^-, \varphi_2^-, \varphi_1^- \}, \{ \}) \\
 \alpha_1 & \equiv (P \cup \{ \varphi_k^-, \dots, \varphi_3^-, \varphi_2^-, \varphi_1^- \}, \{ \psi_1 \}) \\
 \alpha_2 & \equiv (P \cup \{ \varphi_k^-, \dots, \varphi_3^-, \varphi_2, \varphi_1^- \}, \{ \psi_2 \}) \\
 \alpha_3 & \equiv (P \cup \{ \varphi_k^-, \dots, \varphi_3^-, \varphi_2, \varphi_1 \}, \{ \psi_2, \psi_1 \}) \\
 \alpha_4 & \equiv (P \cup \{ \varphi_k^-, \dots, \varphi_3, \varphi_2^-, \varphi_1^- \}, \{ \psi_3 \}) \\
 & \dots \\
 \alpha_{2^k-1} & \equiv (P \cup \{ \varphi_k, \dots, \varphi_3, \varphi_2, \varphi_1 \}, \{ \psi_k, \dots, \psi_3, \psi_2, \psi_1 \})
 \end{aligned}$$

I.e., start from the set P of pre-conditions and an empty set of post-conditions, be corresponding to each conditional post-condition φ_i/ψ_i ($1 \leq i \leq k$), add either φ_i or φ_i^- into the pre-condition set, and add ψ_i into the postcondition set as while as φ_i is added into the pre-condition set.

3. For each atomic action definition $\alpha_i \equiv (P_i, E_i)$ ($0 \leq i \leq 2^k - 1$) constructed above, if it is consistent w.r.t. \mathcal{R} and \mathcal{T} , then do the following operations sequentially:
 - (a) Construct an empty set O_i on primitive literals.
 - (b) For each occlusion ϕ_j ($1 \leq j \leq m$), if both $\phi_j \notin E_i^*_{\mathcal{R}}$ and $\phi_j^- \notin E_i^*_{\mathcal{R}}$, then add ϕ_j into the set O_i .
 - (c) Let $\phi_{i,1}, \dots, \phi_{i,m_i}$ be all the primitive literals contained in the set O_i , construct 2^{m_i} atomic action definitions as follows:

$$\begin{aligned}
 \alpha_{i,0} &\equiv (P_i, E_i \cup \{\phi_{i,m_i}^-, \dots, \phi_{i,3}^-, \phi_{i,2}^-, \phi_{i,1}^-\}) \\
 \alpha_{i,1} &\equiv (P_i, E_i \cup \{\phi_{i,m_i}^-, \dots, \phi_{i,3}^-, \phi_{i,2}^-, \phi_{i,1}\}) \\
 \alpha_{i,2} &\equiv (P_i, E_i \cup \{\phi_{i,m_i}^-, \dots, \phi_{i,3}^-, \phi_{i,2}, \phi_{i,1}^-\}) \\
 \alpha_{i,3} &\equiv (P_i, E_i \cup \{\phi_{i,m_i}^-, \dots, \phi_{i,3}^-, \phi_{i,2}, \phi_{i,1}\}) \\
 \alpha_{i,4} &\equiv (P_i, E_i \cup \{\phi_{i,m_i}^-, \dots, \phi_{i,3}, \phi_{i,2}^-, \phi_{i,1}^-\}) \\
 &\dots \\
 \alpha_{i,2^{m_i}-1} &\equiv (P_i, E_i \cup \{\phi_{i,m_i}, \dots, \phi_{i,3}, \phi_{i,2}, \phi_{i,1}\})
 \end{aligned}$$

I.e., start from the set P_i of pre-conditions and the set E_i of post-conditions, be corresponding to each primitive literal $\phi_{i,j}$ ($1 \leq j \leq m_i$), add either $\phi_{i,j}$ or $\phi_{i,j}^-$ into the post-condition set.

- (d) For each atomic action definition constructed above, if it is consistent w.r.t. \mathcal{R} and \mathcal{T} , then add it into the set $\mathcal{A}_{\mathcal{C}}$.
4. If the set $\mathcal{A}_{\mathcal{C}}$ is empty, then construct an atomic action definition $\beta_0 \equiv (\{false\}, \emptyset)$ and return the action β_0 , else let $\beta_1 \equiv (P_{\beta_1}, E_{\beta_1}), \dots, \beta_n \equiv (P_{\beta_n}, E_{\beta_n})$ be all the atomic action definitions contained in $\mathcal{A}_{\mathcal{C}}$, construct a choice action $\beta_1 \cup \dots \cup \beta_n$ and return it.

As an example, taken the atomic action $BuyBook_{Tom, King}$ defined in previous example as an input, the procedure $Expand(BuyBook_{Tom, King})$ will return a choice action as follows:

$$BuyBookNotified_1 \cup BuyBookNotified_2$$

where $BuyBookNotified_1$ and $BuyBookNotified_2$ are defined by the following atomic action definitions respectively:

$$\begin{aligned}
 &BuyBookNotified_1 \\
 &\equiv (\{ customer(Tom), book(KingLear), instore(KingLear) \}, \\
 &\quad \{ bought(Tom, KingLear), \neg instore(KingLear), \\
 &\quad \quad notify(Tom, NotifyOrderSucceed) \}) \\
 &BuyBookNotified_2 \\
 &\equiv (\{ customer(Tom), book(KingLear), \neg instore(KingLear) \}, \\
 &\quad \{ notify(Tom, NotifyBookOutOfStock) \})
 \end{aligned}$$

The procedure *Expand()* is technically designed to guarantee the following property.

Theorem 1. *Let α be an atomic action defined by some extended atomic action definition $\alpha \equiv (P, O, E)$ w.r.t. an acyclic TBox \mathcal{T} , let $\alpha_1 \cup \dots \cup \alpha_n$ be the action returned by the procedure *Expand*(α), and let \mathcal{A}_C be an ActBox composed of atomic action definitions of every α_i ($1 \leq i \leq n$). Then, for any model $M = (W, T, \Delta, I)$ with $M \models \mathcal{T}$, $M \models \alpha \equiv (P, O, E)$ and $M \models \mathcal{A}_C$, it must be $T(\alpha_1 \cup \dots \cup \alpha_n) = T(\alpha)$.*

Now we are ready to present a satisfiability-checking algorithm for formulas that might contain atomic actions defined by extended atomic action definitions.

Algorithm 1. *Let \mathcal{A}_C be an ActBox which might contains some extended atomic action definitions, let φ be a formula defined w.r.t. \mathcal{A}_C . Then, the satisfiability of φ w.r.t. a TBox \mathcal{T} and ActBox \mathcal{A}_C is decided according to the following steps.*

1. *Construct a formula φ' and an ActBox \mathcal{A}_C' according to the following steps:*
 - (a) *set $\mathcal{A}_C' := \emptyset$ and $\varphi' := \varphi$;*
 - (b) *for each atomic action occurring in φ' , if it is defined in \mathcal{A}_C by some atomic action definition $\alpha \equiv (P, E)$, then add $\alpha \equiv (P, E)$ into the set \mathcal{A}_C' ;*
 - (c) *for each atomic action occurring in φ' , if it is defined in \mathcal{A}_C by some extended atomic action definition $\alpha \equiv (P, O, E)$, then do the following operations sequentially:*
 - i. *call the procedure *Expand*(α) and let $\alpha_1 \cup \dots \cup \alpha_n$ be the action returned by it;*
 - ii. *add all the atomic action definitions of $\alpha_1, \dots, \alpha_n$ into the set \mathcal{A}_C' ; and*
 - iii. *for each occurrence of the action α in the formula φ' , replace it with the action $\alpha_1 \cup \dots \cup \alpha_n$.*
2. *Since every atomic action occurring in φ' is defined by original atomic action definitions discussed in DDL, we can call the procedure provided by DDL to decide whether the formula φ' is satisfiable w.r.t. \mathcal{T} and \mathcal{A}_C' ; if φ' is satisfiable w.r.t. \mathcal{T} and \mathcal{A}_C' , then return “TRUE”, otherwise return “FALSE”.*

For the formula φ' constructed in this algorithm, according to Theorem 1, it is straightforward that φ' is satisfiable w.r.t. \mathcal{T} and \mathcal{A}_C' if and only if φ is satisfiable w.r.t. \mathcal{T} and \mathcal{A}_C . Therefore, this deciding algorithm is correct; i.e.,

Theorem 2. *Algorithm 1 returns “TRUE” if and only if the formula φ is satisfiable w.r.t. \mathcal{T} and \mathcal{A}_C .*

5 Conclusion

In this paper, the dynamic description logic *DDL* is extended to support occlusions and conditional post-conditions in the description of atomic actions. As

a result, the action theory supported by *DDL* is compatible with the action formalism constructed by Baader et.al. [1].

DDL provides an approach to bring the power and character of description logics into the description and reasoning of dynamic application domains. One of our future work is to optimize the reasoning mechanisms of *DDL*. Another work is to apply *DDL* to model and reason about semantic Web services.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China under Grant Nos. 60903079, 60775035 and 60963010.

References

1. Baader, F., Lutz, C., Milicic, M., Sattler, U., Wolter, F.: Integrating Description Logics and Action Formalisms: First Results. In: Veloso, M., Kambhampati, S. (eds.) Proceedings of the 12th Nat. Conf. on Artif. Intell., pp. 572–577. AAAI Press, Menlo Park (2005)
2. Calvanese, D., De Giacomo, G., Vardi, M.: Reasoning about Actions and Planning in LTL Action Theories. In: Fensel, D., Giunchiglia, F., McGuinness, D., Williams, M. (eds.) 8th Int. Conf. on Principles and Knowledge Representation and Reasoning, pp. 593–602. Morgan Kaufmann, San Francisco (2002)
3. Chang, L., Lin, F., Shi, Z.: A Dynamic Description Logic for Representation and Reasoning about Actions. In: Zhang, Z., Siekmann, J.H. (eds.) KSEM 2007. LNCS (LNAI), vol. 4798, pp. 115–127. Springer, Heidelberg (2007)
4. Chang, L., Shi, Z., Qiu, L., Lin, F.: A Tableau Decision Algorithm for Dynamic Description Logic. Chinese Journal of Computers 31(6), 896–909 (2008)
5. De Giacomo, G., Lenzerini, M.: PDL-based Framework for Reasoning about Actions. In: Gori, M., Soda, G. (eds.) AI*IA 1995. LNCS, vol. 992, pp. 103–114. Springer, Heidelberg (1995)
6. Martin, D., Burstein, M., McDermott, D., McIlraith, S., Paolucci, M., Sycara, K., McGuinness, D., Sirin, E., Srinivasan, N.: Bringing semantics to web services with OWL-S. World Wide Web Journal 10(3), 243–277 (2007)
7. McIlraith, S., Son, T., Zeng, H.: Semantic Web Services. IEEE Intelligent Systems 16(2), 46–53 (2001)
8. Narayanan, S., McIlraith, S.: Simulation, verification and automated composition of web services. In: Proc. of the 11th Int. World Wide Web Conference (WWW 2002), pp. 77–88 (2002)
9. Reiter, R.: Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems. MIT Press, Cambridge (2001)
10. Shi, Z., Dong, M., Jiang, Y., Zhang, H.: A Logic Foundation for the Semantic Web. Science in China, Series F 48(2), 161–178 (2005)
11. Thielscher, M.: Introduction to the Fluent Calculus. Electron. Trans. Artif. Intell. 2(3-4), 179–192 (1998)
12. Winslett, M.: Reasoning about Action Using a Possible Models Approach. In: 7th Nat. Conf. on Artif. Intell., pp. 89–93. AAAI Press, Menlo Park (1988)

Preservative Translations between Logical Systems

Yuming Shen^{1,2}, Yue Ma^{1,2}, Cungen Cao¹, Yuefei Sui¹, and Ju Wang³

¹ Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences 100190, China

shenyuming@ict.ac.cn

² Graduate University of Chinese Academy of Sciences, 100049, China

³ School of Computer Science and Information Engineering, Guangxi Normal University 541004, China

Abstract. This paper investigates a subclass of translations between logical systems, called the preservative translations, which preserve the satisfiability and the unsatisfiability of formulas. The definition of preservative translation is given and connections between the preservative translation and other definitions of translations in the literature are discussed. Some properties of logical systems, such as the compactness and the decidability, which are characterized by the existence of preservative translations between logical systems are also studied.

Keywords: Translation, Faithfulness, Fullness, Preservative translation.

1 Introduction

The method of studying relations between logical systems by analyzing translations between them was originally introduced by Kolmogorov [7]. Translating or encoding one logical system (the *source logical system*) into another logical system (the *target logical system*) gives us a powerful tool for understanding logical systems from a logical, algorithmical, and computational point of view. Translating a source logical system into a target logical system have several advantages, and some of these are given below [10]:

- To grasp the expressive power, both of the source logical system and of the target logical system.
- To export any proof procedures and tools that we may have for the target logical system to the source logical system.
- To grasp the computational costs of various reasoning tasks, again both for the source logical system and for the target logical system.
- To combine two logical systems, sometimes the most natural solution is to translate both into the same target logical system, and to work inside the target logical system.

Many translations found in the literature(see, e.g., [6,8,12]) satisfy the following logical properties:

- *Soundness*: for every formula φ of the source logical system, if φ is satisfiable then φ is translated to a satisfiable formula of the target logical system;
- *Completeness*: for every formula φ of the source logical system, if the translated formula for φ is satisfiable, then φ is satisfiable.

By the definitions of soundness and completeness, we have that the satisfiability of formulas is preserved. However, translating in a satisfiability-preserving way does not immediately lead to the preservation of the unsatisfiability, if the translation between models is taken into account and the class of models of a source logical system is translated to a proper subclass of the class of models of a target logical system. For example, Fara and Williamson [3] showed that the translations from first-order modal logic into the counterpart theory given by Lewis [9], Forbes [5] and Ramachandran [14] may translate an unsatisfiable formula to a satisfiable formula.

Inspired by the observation, in [13], we give the following logical properties to describe a translation σ .

- ◇ *Faithfulness*: for every formula φ of a source logical system, φ is satisfied in a model \mathfrak{M} and a valuation v of the source logical system if and only if $\sigma(\varphi)$ is satisfied in the translated model $\sigma(\mathfrak{M})$ and valuation $\sigma(v)$, that is,

$$(\mathfrak{M}, v) \models \varphi \text{ if and only if } (\sigma(\mathfrak{M}), \sigma(v)) \models \sigma(\varphi).$$

- ◇ *Fullness*: for any formula φ of a source logical system, any model \mathfrak{M}' and any valuation v' of a target logical system, if $(\mathfrak{M}', v') \models \sigma(\varphi)$, then there exists a model \mathfrak{M} and a valuation v of the source logical system such that $(\mathfrak{M}, v) \models \varphi$ and $\sigma(\mathfrak{M}) = \mathfrak{M}'$, $\sigma(v) = v'$.

The faithfulness is the soundness and completeness of a translation. The fullness says that every model and every valuation which satisfies $\sigma(\varphi)$ has a corresponding model and valuation in the source logical system. By the definitions of faithfulness and fullness, we have that the satisfiability and unsatisfiability of formulas are both preserved.

In this paper, a translation is called a preservative translation if it is faithful and full. The aim of this paper is to investigate a subclass of translations, the preservative translations, which preserve the satisfiability and the unsatisfiability.

The paper is organized as follows. In section 2, we give the definition of preservative translation, and show that the satisfiability and unsatisfiability of formulas are preserved under the preservative translations. Two translation examples are also given in the section, the first example is the standard relational translation from propositional modal logic into first-order logic, the second is the identity translation from intuitionistic propositional logic into classical propositional logic. In section 3, we discuss connections between the definition of preservative translation and several definitions of translation given in the literature. In section 4, we study some properties of logical systems, such as the decidability and the compactness, which are characterized by the existence of preservative translations between them, and section 5 is the conclusion.

2 The Concept of Preservative Translation

In this section, we firstly present the definitions of logical system and translation between logical systems introduced in [14], and then give the definition of preservative translation. Next, we show that the satisfiability and the unsatisfiability of formulas are preserved under the preservative translations. Finally, we give the following examples: the standard relational translation from modal propositional logic (**PML**) into first-order logic (**FOL**) and the identity translation from intuitionistic propositional logic (**IPL**) into classical propositional logic (**CPL**). We show that the former is a preservative translation, whereas the latter is not.

In [14], logical systems are characterized as sets with consequence relation and translation between logical systems as consequence relation preserving maps.

Definition 1. *A logical system \mathcal{L} is a pair (\mathcal{L}, C) such that \mathcal{L} is a formal language and C is a consequence operator in the free algebra $Form(A)$ of the formulas of \mathcal{L} , that is, $C : 2^{Form(A)} \rightarrow 2^{Form(A)}$ is a function that satisfies, for $X, Y \subseteq Form(A)$, the following conditions:*

- (i) $X \subseteq C(X)$;
- (ii) If $X \subseteq Y$, then $C(X) \subseteq C(Y)$;
- (iii) $C(C(X)) \subseteq C(X)$.

It is clear that the non-monotonic logics do not match the above definition.

Definition 2. *A translation from a logical system \mathcal{L} into a logical system \mathcal{L}' is a mapping $\sigma : \mathcal{L} \rightarrow \mathcal{L}'$ such that $\sigma(C_{\mathcal{L}}(X)) \subseteq C_{\mathcal{L}'}(\sigma(X))$ for any $X \subseteq Form(\mathcal{L})$.*

By the definition [2], it clear that if the syntactical consequence relations $\vdash_{\mathcal{L}}$ and $\vdash_{\mathcal{L}'}$ are associated with $C_{\mathcal{L}}$ and $C_{\mathcal{L}'}$, respectively, then a function $\sigma : \mathcal{L} \rightarrow \mathcal{L}'$ is a translation if and only if for every $\Phi \subseteq Form(\mathcal{L})$, $\varphi \in Form(\mathcal{L})$, if $\Phi \vdash_{\mathcal{L}} \varphi$, then $\sigma(\Phi) \vdash_{\mathcal{L}'} \sigma(\varphi)$.

In the following discussion, we suppose that a translation σ is extended by translating every model and every valuation of \mathcal{L} to a model and a valuation of \mathcal{L}' , respectively. Then, we give the definition of preservative translation as follows:

Definition 3. *Let \mathcal{L} and \mathcal{L}' be logical systems. A translation σ is called a preservative translation from \mathcal{L} into \mathcal{L}' if the following conditions holds:*

- (1) *For any formula φ , any model \mathfrak{M} and any valuation v of \mathcal{L} ,*

$$(\mathfrak{M}, v) \models_{\mathcal{L}} \varphi \text{ if and only if } (\sigma(\mathfrak{M}), \sigma(v)) \models_{\mathcal{L}'} \sigma(\varphi).$$

- (2) *For any formula φ of \mathcal{L} , any model \mathfrak{M}' and any valuation v' of \mathcal{L}' , if $(\mathfrak{M}', v') \models_{\mathcal{L}'} \sigma(\varphi)$, then there exists a model \mathfrak{M} and a valuation v of \mathcal{L} such that $(\mathfrak{M}, v) \models_{\mathcal{L}} \varphi$ and $\sigma(\mathfrak{M}) = \mathfrak{M}'$, $\sigma(v) = v'$.*

Remark. In some cases, a logical system does not contain valuations, for example, propositional logic and propositional modal logic. If both \mathcal{L} and \mathcal{L}' do not contain valuations, then there is no corresponding relation between valuations of \mathcal{L} and valuations of \mathcal{L}' , for example, the translation between classical propositional logic and intuitionistic propositional logic. If one of them contains valuations, then the mapping σ may be considered as a partial function, for example, the standard relational translation from propositional modal logic into first-order logic. \square

By the definition of preservative translation, we have that the satisfiability and unsatisfiability of formulas are preserved.

Proposition 1. *If σ is a preservative translation from \mathcal{L} into \mathcal{L}' , then for any formula φ of \mathcal{L} , φ is a satisfiable formula if and only if $\sigma(\varphi)$ is a satisfiable formula.* \square

Proposition 2. *If σ is a preservative translation from \mathcal{L} into \mathcal{L}' , then for any formula φ of \mathcal{L} , φ is an unsatisfiable formula if and only if $\sigma(\varphi)$ is an unsatisfiable formula.*

Proof. For any formula φ , if φ is an unsatisfiable formula but $\sigma(\varphi)$ is a satisfiable formula, then there exists a model \mathfrak{M}' and valuation v' of \mathcal{L}' such that $(\mathfrak{M}', v') \models_{\mathcal{L}'} \sigma(\varphi)$. By the definition of preservative translation, we have that there exists a model \mathfrak{M} and valuation v such that $(\mathfrak{M}, v) \models_{\mathcal{L}} \varphi$. Contradicting the fact φ is an unsatisfiable formula. On the other hand, if $\sigma(\varphi)$ is an unsatisfiable formula but φ is a satisfiable formula, then there exists a model \mathfrak{M} and valuation v of \mathcal{L} such that $(\mathfrak{M}, v) \models_{\mathcal{L}} \varphi$. By the definition of preservative translation, we have that $(\sigma(\mathfrak{M}), \sigma(v)) \models_{\mathcal{L}'} \sigma(\varphi)$. Contradicting the fact $\sigma(\varphi)$ is unsatisfiable. \square

The following example shows that the standard relational translation from **PML** into **FOL** is a preservative translation.

Example 1. Suppose that **PML** and **FOL** both contain the connectives \neg, \rightarrow and the vocabulary of **FOL** consists of a binary predicate **R** to represent the accessibility relation and unary predicate symbols to represent proposition letters, the standard relational translation σ is given as follows:

$$\sigma(\varphi, w) = \begin{cases} p(w) & \text{if } \varphi = p \\ \neg\sigma(\psi, w) & \text{if } \varphi = \neg\psi \\ \sigma(\psi, w) \rightarrow \sigma(\theta, w) & \text{if } \varphi = \psi \rightarrow \theta \\ \forall w'(\mathbf{R}(w, w') \rightarrow \sigma(\psi, w')) & \text{if } \varphi = \Box\psi. \end{cases}$$

On the semantic side, for any **PML** model $\mathfrak{M} = (W, R, \mathcal{I})$, we construct a **FOL** model $\sigma(\mathfrak{M}) = \mathfrak{M}' = (U', \mathcal{I}')$ as follows:

- $U' = W$;
- $\mathcal{I}'(\mathbf{R}) = R$;
- $\mathcal{I}'(p) = \{w \in W : \mathcal{I}(p, w) = 1\}$;
- $v'(w) = w$.

By induction on φ , we have that for any formula φ and any model \mathfrak{M} of **PML**,

$$(\mathfrak{M}, w) \models_{\mathbf{PML}} \varphi \text{ if and only if } (\sigma(\mathfrak{M}), v') \models_{\mathbf{FOL}} \sigma(\varphi, w).$$

For any **FOL** model $\mathfrak{M}' = (U', \mathcal{I}')$, we construct a corresponding **PML** model $\mathfrak{M} = (W, R, \mathcal{J})$ as follows:

- $W = U'$;
- $R = \mathcal{I}'(\mathbf{R})$;
- $\mathcal{J}(p, w) = 1$ if and only if $w \in \mathcal{I}'(p)$.

By induction on φ , we have that for any formula φ of **PML**, and any model \mathfrak{M}' and valuation v' of **FOL**, if $(\mathfrak{M}', v') \models_{\mathbf{FOL}} \sigma(\varphi, w)$ then there exists a model \mathfrak{M} of **PML** such that $(\mathfrak{M}, w) \models_{\mathbf{PML}} \varphi, \sigma(\mathfrak{M}) = \mathfrak{M}'$. By the definition of preservative translation, we have that the standard relational translation is preservative.

The next example is the identity translation from **IPL** into **CPL**. we show that the translation is not preservative, since the law of excluded middle holds for **CPL**, not for **IPL**.

Example 2. Suppose that **IPL** and **CPL** both contain the connectives $\neg, \wedge, \vee, \rightarrow$. The identity function σ from **IPL** into **CPL** is a translation. Since the formula $p \wedge \neg p$ is a valid formula in **CPL** but it is invalid in **IPL**, the identity translation σ is not a preservative translation.

3 Connections with Several Definitions of Translations

In this section, we present several definitions of translations in the literature and analyze connections between the definition of preservative translation and them.

3.1 Translations for Epstein and Krajewski

In [2], Epstein and krajewski present a validity mapping of a propositional logic \mathcal{L} into a propositional logic \mathcal{L}' as a map σ from the language of \mathcal{L} into the language of \mathcal{L}' such that, for every formula φ ,

$$\models_{\mathcal{L}} \varphi \text{ if and only if } \models_{\mathcal{L}'} \sigma(\varphi).$$

A translation is a validity mapping σ such that, for every set Φ of formulas and every formula φ of \mathcal{L} ,

$$\Phi \models_{\mathcal{L}} \varphi \text{ if and only if } \sigma(\Phi) \models_{\mathcal{L}'} \sigma(\varphi).$$

The following proposition show that if σ is a preservative translation and satisfies the following conditions:

- (1) σ is injective at models level, that is, for any models $\mathfrak{M}_1, \mathfrak{M}_2$ of \mathcal{L} , if $\mathfrak{M}_1 \neq \mathfrak{M}_2$, then $\sigma(\mathfrak{M}_1) \neq \sigma(\mathfrak{M}_2)$.

- (2) σ is injective at valuations level, that is, for any given model \mathfrak{M} of \mathcal{L} and any valuations v_1, v_2 in \mathfrak{M} , if $v_1 \neq v_2$, then $\sigma(v_1) \neq \sigma(v_2)$.

Then Epsten's definition of translation coincides with our definition of preservative translation.

Proposition 3. *Let \mathcal{L} and \mathcal{L}' be logical systems. If σ is a preservative translation from \mathcal{L} into \mathcal{L}' and satisfies the above conditions (1) and (2), then for every formula set Φ and every formula φ of \mathcal{L} , $\Phi \models_{\mathcal{L}} \varphi$ if and only if $\sigma(\Phi) \models_{\mathcal{L}'} \sigma(\varphi)$.*

Proof. Suppose that $\Phi \models_{\mathcal{L}} \varphi$. For any model \mathfrak{M}' and any valuation v' of \mathcal{L}' , if $(\mathfrak{M}', v') \models_{\mathcal{L}'} \sigma(\Phi)$, then by σ is a preservative translation, there exists a model \mathfrak{M} and valuation v of \mathcal{L} such that $(\mathfrak{M}, v) \models_{\mathcal{L}} \Phi, \sigma(\mathfrak{M}) = \mathfrak{M}', \sigma(v) = v'$. Since $\Phi \models_{\mathcal{L}} \varphi$, we have that $(\mathfrak{M}, v) \models_{\mathcal{L}} \varphi$. By σ is a preservative translation, we have that $(\mathfrak{M}', v') \models_{\mathcal{L}'} \sigma(\varphi)$, i.e., $\sigma(\Phi) \models_{\mathcal{L}'} \sigma(\varphi)$.

On the other hand, Suppose that $\sigma(\Phi) \models_{\mathcal{L}'} \sigma(\varphi)$. For any model \mathfrak{M} and any valuation v of \mathcal{L} , if $(\mathfrak{M}, v) \models_{\mathcal{L}} \Phi$, then by σ is a preservative translation, we have that $(\sigma(\mathfrak{M}), \sigma(v)) \models_{\mathcal{L}'} \sigma(\Phi)$. Since $\sigma(\Phi) \models_{\mathcal{L}'} \sigma(\varphi)$, we have that $(\sigma(\mathfrak{M}), \sigma(v)) \models_{\mathcal{L}'} \sigma(\varphi)$. Since σ is a preservative translation and satisfies the conditions (1) and (2), we get that $(\mathfrak{M}, v) \models_{\mathcal{L}} \varphi$, i.e., $\Phi \models_{\mathcal{L}} \varphi$. \square

3.2 Translations for Prawitz and Malmnäs

In [11], Prawitz and Malmnäs define that a translation from \mathcal{L} into \mathcal{L}' is a function σ such that, for every formula of \mathcal{L} ,

$$\vdash_{\mathcal{L}} \varphi \text{ if and only if } \vdash_{\mathcal{L}'} \sigma(\varphi).$$

The following proposition shows that if \mathcal{L} and \mathcal{L}' are logical systems whose languages have the negation \neg and σ is a preservative translation which satisfies that for any formula φ of \mathcal{L} , $\sigma(\neg\varphi) = \neg\sigma(\varphi)$, then the definition of translation given by Prawitz and Malmnäs coincides with our definition of preservative translation.

Proposition 4. *Let \mathcal{L} and \mathcal{L}' be logical systems whose languages have the negation \neg and σ is a preservative translation which satisfies that for any formula φ of \mathcal{L} , $\sigma(\neg\varphi) = \neg\sigma(\varphi)$, then*

$$\vdash_{\mathcal{L}} \varphi \text{ if and only if } \vdash_{\mathcal{L}'} \sigma(\varphi).$$

Proof. For any formula φ of \mathcal{L} , if $\vdash_{\mathcal{L}} \varphi$ but $\not\vdash_{\mathcal{L}'} \sigma(\varphi)$, then there exists a model \mathfrak{M}' and a valuation v' of \mathcal{L}' such that $(\mathfrak{M}', v') \models_{\mathcal{L}'} \neg\sigma(\varphi)$. Since $\sigma(\neg\sigma) = \neg\sigma(\varphi)$ and σ is preservative, there exists a model \mathfrak{M} and a valuation v of \mathcal{L} such that $(\mathfrak{M}, v) \models_{\mathcal{L}} \neg\varphi$. Contradicting the fact $\vdash_{\mathcal{L}} \varphi$. On the other hand, if $\vdash_{\mathcal{L}'} \sigma(\varphi)$ but $\not\vdash_{\mathcal{L}} \varphi$, then there exists a model \mathfrak{M} and a valuation v such that $(\mathfrak{M}, v) \models_{\mathcal{L}'} \neg\varphi$. By $\sigma(\neg\varphi) = \neg\sigma(\varphi)$ and σ is preservative, we have that $(\sigma(\mathfrak{M}), \sigma(v)) \models_{\mathcal{L}'} \neg\sigma(\varphi)$. Contradicting the fact $\vdash_{\mathcal{L}'} \sigma(\varphi)$. \square

Feitosa and Ottaviano [4] define that a conservative mapping from $\mathcal{L} = (A, C)$ into $\mathcal{L}' = (A', C')$ is a function σ such that, for every $x \in A$,

$$x \in C(\emptyset) \text{ if and only if } \sigma(x) \in C'(\emptyset),$$

where A is a set and C is a consequence operator. If we restrict A, A' to formal languages and the syntactical consequence relations $\vdash_{\mathcal{L}}$ and $\vdash_{\mathcal{L}'}$ are associated with C and C' , respectively, then by the proposition [4], we have that the conservative mapping coincides with our definition of preservative translation.

4 General Properties Characterized by the Preservative Translations

In this section, we give some general properties of logical logics, such as the compactness and the decidability, which are characterized by the existence of preservative translations between logical logics.

The compactness theorem of a logical system \mathcal{L} is that any formulas set Φ of \mathcal{L} , if every finite subset of Φ is satisfiable then Φ is satisfiable. The following proposition shows that if there is a preservative translation from \mathcal{L} into \mathcal{L}' then the compactness theorem holds for \mathcal{L}' implies that it holds for \mathcal{L} .

Proposition 5. *Let σ be a preservative translation from \mathcal{L} into \mathcal{L}' . If the compactness theorem holds for \mathcal{L}' , then the compactness theorem holds for \mathcal{L} .*

Proof. For any given formulas set Φ of \mathcal{L} and for every finite Ψ_0 of $\sigma(\Phi)$, there exists a finite subset Φ_0 of Φ such that $\sigma(\Phi_0) = \Psi_0$. By σ is preservative, we have that if Φ_0 is a satisfiable then Ψ_0 is satisfiable. As the compactness theorem holds for \mathcal{L}' , we have that $\sigma(\Phi)$ is satisfiable, and by σ is preservative we have that Φ is satisfiable, i.e., the compactness theorem holds for \mathcal{L} . \square

Since the compactness theorem holds for first-order logic, not for second-order logic, we have the following corollary.

Corollary 1. *There is no preservative translation from second-order logic into first-order logic.* \square

The decidability of a logical system \mathcal{L} is that there exists an effective procedure that, given any formula φ , will decide whether or not it is a theorem of \mathcal{L} . Let $\mathcal{L}, \mathcal{L}'$ be logics whose languages have negation \neg , the following proposition shows that if there is a preservative translation σ from \mathcal{L} into \mathcal{L}' and σ also satisfies the following conditions:

- σ is a recursive function at formulas level, that is, for any formula φ of \mathcal{L} , we can determine $\sigma(\varphi)$ is a formula of \mathcal{L}' .
- σ is distributive with negation, that is, for any formula φ of \mathcal{L} , $\sigma(\neg\varphi) = \neg\sigma(\varphi)$.

Then the decidability of \mathcal{L}' implies that \mathcal{L} is decidable.

Proposition 6. *Let $\mathcal{L}, \mathcal{L}'$ be logics whose languages have negation \neg . If there is a preservative translation from \mathcal{L} into a decidable \mathcal{L}' and σ is recursive at formulas level and distributive with negation, then \mathcal{L} is decidable.*

Proof. For any formula φ of \mathcal{L} , as σ is recursive at formulas level, we can determine $\sigma(\varphi)$. Since \mathcal{L}' is decidable, it is possible to verify that $\sigma(\varphi)$ is or is not a theorem of \mathcal{L}' . Hence, if $\sigma(\varphi)$ is a theorem of \mathcal{L}' but φ is not a theorem of \mathcal{L} then there exists a model \mathfrak{M} and a valuation v of \mathcal{L} such that $(\mathfrak{M}, v) \models_{\mathcal{L}} \neg\varphi$. By σ is preservative and distributive with negation, we have that $(\sigma(\mathfrak{M}), \sigma(v)) \models_{\mathcal{L}'} \neg\sigma(\varphi)$. Contradicting the fact $\sigma(\varphi)$ is a theorem of \mathcal{L}' , i.e., φ is a theorem of \mathcal{L} . If $\sigma(\varphi)$ is not a theorem of \mathcal{L}' , then there exists a model \mathfrak{M}' and a valuation v' such that $(\mathfrak{M}', v') \models_{\mathcal{L}'} \neg\sigma(\varphi)$. By σ is preservative and distributive with negation, there exists a model \mathfrak{M} and a valuation v such that $(\mathfrak{M}, v) \models_{\mathcal{L}} \neg\varphi$, i.e., φ is not a theorem of \mathcal{L} . \square

Since propositional modal logics K, D, T, S4, S5 are decidable but first-order logic is not, we have the following corollary.

Corollary 2. *There is no recursive, distributive and preservative translation from first-order logic into propositional modal logics K, D, T, S4, S5.* \square

5 Conclusion

In this paper, we investigate a subclass of translations, the preservative translations, which preserve the satisfiability and unsatisfiability of formulas. We give the definition of preservative translation, and show that the satisfiability and unsatisfiability of formulas are preserved under such translations. Also, we show that the standard relational translation from propositional modal logic into first-order logic is a preservative translation and the identity translation from intuitionistic propositional logic into classical propositional logic is not. Connections between our definition of preservative translations and other definitions of translations in the literature are given. Finally, the preservation of general properties of logical systems, such as the compactness and the decidability, is also discussed.

Acknowledgements

The work was supported by the National Natural Science Foundation of China under Grant Nos.60496326, 60573063, 60573064, 60773059 and the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z325.

References

1. Carnielli, W.A., Coniglio, M.E., D'Ottaviano, I.M.L.: New dimensions on translation between logics. *Logica Universalis* 3, 1–18 (2009)
2. Epstein, R.L.: *The Semantic Foundations of Logics. Propositional Logics*, vol. 1. Kluwer, Dordrecht (1990)

3. Fara, M., Williamson, T.: Counterparts and actuality. *Mind* 114, 1–30 (2005)
4. Feitosa, H.A., D'Ottaviano, I.M.L.: Conservative translations. *Annals of Pure and Applied Logic* 108, 205–227 (2001)
5. Forbes, G.: Canonical counterpart theory. *Analysis* 42, 33–37 (1982)
6. Hustadt, U., Schmidt, R.A., Georgieva, L.: A survey of decidable first-order fragments and description logics. *Journal of Relational Methods in Computer Science* 1, 251–276 (2004)
7. Kolmogorov, A.N.: On the principle of excluded middle(1925). In: Heijenoort, J. (ed.) *From Frege to Gödel: a Source Book in mathematical logic 1879–1931*, pp. 414–437. Harvard University Press, Cambridge (1977)
8. Kurtonina, N., de Rijke, M.: Expressive of concept expression in first-order description logics. *Artificial Intelligence* 107, 303–333 (1999)
9. Lewis, D.: Counterpart theory and quantified modal logic. *Journal of Philosophy* 65, 113–126 (1968)
10. Ohlbach, H., Nonnengart, A., de Rijke, M., Gabbay, D.: Encoding two-valued non-classical logics in classical logic. In: Robinson, A., Voronkov, A. (eds.) *Handbook of Automated Reasoning*, pp. 1403–1486. Elsevier, Amsterdam (2001)
11. Prawitz, D., Malmnäs, P.E.: A survey of some connections between classical, intuitionistic and minimal logic. In: Schmidt, H., et al. (eds.) *Contributions to Mathematical Logic*, pp. 215–229. North-Holland, Amsterdam (1968)
12. Schmidt, R., Hustadt, U.: The axiomatic translation principle for modal logic. *ACM Transactions on Computational Logic* 8, 1–55 (2007)
13. Shen, Y., Ma, Y., Cao, C., Sui, Y., Wang, J.: Logical properties on translations between logics. *Chinese Journal of Computers* 32, 2091–2098 (2009)
14. Ramachandran, M.: An alternative translation scheme for counterpart theory. *Analysis* 49, 131–141 (1989)

The Description Logic for Relational Databases^{*}

Ma Yue^{1,2}, Shen Yuming^{1,2}, Sui Yuefei¹, and Cao Cungen¹

¹ Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

² Graduate University of Chinese Academy of Science, Beijing, 100039, China

Abstract. Description logics are widely used to express structured data and provide reasoning facility to query and integrate data from different databases. This paper presents a many-sorted description logic \mathcal{MDL} to represent relational databases. We give a translation from relational databases to the description logic \mathcal{MDL} , and show this translation completely and faithfully captures the information in the relational database. Moreover, we show that some relational algebra operations could be expressed in \mathcal{MDL} .

Keywords: Relational database, Description logic, Translation.

1 Introduction

The Relational model is an important theoretic model of database management system. In a relational database, data are manipulated as a collection of relations or tables. In addition, a relational database provides a collection of methods to access the data in tables, and to modify and combine tables by using a set of relational algebra operations.

Description logic is a logic-based formalism of knowledge, focusing on concepts, individuals, roles, and the relationship between them. It is a proper choice to represent relational databases in DL, using its reasoning facility to implement the query of the information in databases, and to integrate different sources of data. There are many papers on represent structured databases in DL. Calvanese et al.[2] proposed a translation from ER-model to a description logic \mathcal{DLR} . In this translation, tuple entities are implemented as constants in the model of \mathcal{DLR} and the sets of entities are regarded as concepts. Moreover, the values of attributes are also described as constants, and there are concepts for attribute values. Unfortunately, in \mathcal{DLR} these two kind of concepts could not be syntactically distinguished. There should be statements in \mathcal{DLR} with form "an attribute value a is an instance of entity-concept R ", which is meaningless in the corresponding databases.

Similar to Calvanese's work, we describe both the sets of attribute values and the sets of tuples in a relational database as concepts in DL. To distinguish these

^{*} The work is supported by the National Natural Science Foundation of China (60496326, 60573063, 60573064 and 60773059), the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z325.

two kind of objects, we propose a many-sorted description logic \mathcal{MDL} , in which the constants of attribute value and the constants of tuples are taken as different types of constant symbols. Similarly, we use different types of concept names representing concepts for attribute values and tuples. Based on the definition of \mathcal{MDL} , we shall present a translation to represent a relational database in \mathcal{MDL} : from a given database, we construct an \mathcal{MDL} model and a knowledge base containing \mathcal{MDL} statements which are satisfied in the \mathcal{MDL} model. We will show that this translation completely and faithfully captures the information in the relational database. Moreover, we show that some relational algebra operations could be expressed in \mathcal{MDL} using concept and role constructors.

This paper is organized as follows: in section 2 we give a brief introduction on relational databases and description logics; in section 3 we propose a many-sorted description logic \mathcal{MDL} , and give its formal definitions ; in section 4 we present the translation from relational database to \mathcal{MDL} , and express relational algebra operations in \mathcal{MDL} ; in the last section we conclude the paper.

2 Preliminaries

This section will give the basic definitions on relational databases and description logics. The relational model is an important theoretic model of database management system. We present a brief formal definitions in relational database and relational algebra [7,8]. Let U be a set of attributes called an universe. For each attribute $A \in U$, we assign a set of values $\text{dom}(A)$ called the domain of A . Let $\text{dom} = \bigcup_{A \in U} \text{dom}(A)$.

Definition. A *relation schema* R on U is a subset of U . A *database schema* D is a set of relation schema. \square

Definition. Let $D = \{R_1, \dots, R_n\}$ be a database schema, R be a relation in D , and X be a set of attributes in U . An X -tuple t is a mapping from X to dom , such that for each $A \in X$, $t(A) \in \text{dom}(A)$. A *relation* r over relation schema R is a finite set of R -tuples. A *database* d over database schema D is a set of relations $\{r_1, \dots, r_n\}$, each r_i is a relation over R_i . We use $\alpha(r)$ to denote the set of attributes of r . Given a database $d = \{r_1, \dots, r_n\}$, each $r_i \in \{r_1, \dots, r_n\}$ is an *atomic relation* in d . \square

Based on the definition of tuples and relations, we now give the definition of the relational algebra. The relational algebra is an algebra on the set of relations, including several primitive operations on relations. We give the formal definitions of these operations as follows:

Definition. Let t be an X -tuple and $Y \subseteq X$. The projection of tuple t on Y , denoted $t[Y]$, is the restriction of t on Y .

Given relations r_1, r_2 and a set of attributes Y , let $\alpha(r_1) = X_1, \alpha(r_2) = X_2$,

- the *projection* of r_1 on Y , denoted $\pi_Y(r_1)$, fulfills:
 - i) $\pi_Y(r_1)$ is defined if $Y \subseteq X_1$. If $\pi_Y(r_1)$ is defined, $\alpha(\pi_Y(r_1)) = Y$,

- ii) $\pi_Y(r_1) = \{t[Y] | t \in r_1\}$.
- the *natural join* of r_1, r_2 , denoted $r_1 \bowtie r_2$, fulfills:
 - i) $\alpha(r_1 \bowtie r_2) = X_1 \cup X_2$,
 - ii) $r_1 \bowtie r_2 = \{t | t \text{ is } X_1 \cup X_2\text{-tuple, such that } t[X_1] \in r_1, t[X_2] \in r_2\}$.
- the *union* of r_1, r_2 , denoted $r_1 \cup r_2$, fulfills:
 - i) $r_1 \cup r_2$ is defined if $X_1 = X_2$. If $r_1 \cup r_2$ is defined, $\alpha(r_1 \cup r_2) = X_1$,
 - ii) $r_1 \cup r_2 = \{t | t \in r_1 \text{ or } t \in r_2\}$.
- the *difference* of r_1, r_2 , denoted $r_1 - r_2$, fulfills:
 - i) $r_1 - r_2$ is defined if $X_1 = X_2$. If $r_1 - r_2$ is defined, $\alpha(r_1 - r_2) = X_1$,
 - ii) $r_1 - r_2 = \{t | t \in r_1 \text{ and } t \notin r_2\}$.
- the *renaming* of r_1 from A to B , denoted $\rho_{B|A}(r_1)$, fulfills:
 - i) $\rho_{B|A}(r_1)$ is defined if $A \in X_1, B \notin X_1$. If $\rho_{B|A}(r_1)$ is defined, $\alpha(\rho_{B|A}(r_1)) = X_1 - \{A\} \cup \{B\}$,
 - ii) $\rho_{B|A}(r_1) = \{t | \text{there exists } t' \in r_1 \text{ such that } t'[A] = t[B], t'[X_1 - \{A\}] = t[X_1 - \{A\}]\}$.
- the *selection* on r_1 by φ , denoted $\theta_\varphi(r_1)$, fulfills:
 - i) φ has the form $A = B | A = v | A \text{ op } v$, where A, B are attributes in $\alpha(r_1)$, $v \in \text{dom}(A)$, **op** is a binary relation over $\text{dom}(A)$.
 - ii) $\alpha(\theta_\varphi(r_1)) = X_1, \theta_\varphi(r_1) = \{t | \varphi \text{ is satisfied on } t\}$. □

From above we give the basic notions of relational database and relational algebra. Now we present a basic introduction of description logic. Description logic is a logic-based formalism for knowledge representing, focusing on describe individual, concept, role and the relationship between them. We give formal definition of a simple description logic \mathcal{ALC} .

Definition (syntax and semantics of \mathcal{ALC}). The language $L_{\mathcal{ALC}}$ for \mathcal{ALC} contains the following primitive symbols:

- object names $\mathbf{c}_0, \mathbf{c}_1, \dots$;
- concept names $\top, \mathbf{C}_0, \mathbf{C}_1, \dots$;
- role names $\mathbf{r}_0, \mathbf{r}_1, \dots$;
- concept constructors: \neg, \sqcap, \exists ;
- the subsumption relation: \sqsubseteq ; and
- logical connectives: \neg, \rightarrow .

All the concept names and \top are atomic concepts; If \mathbf{C} and \mathbf{D} are concepts, \mathbf{c}, \mathbf{d} are object names, φ and ψ are statements, and \mathbf{r} is a role name, then

- (i) $\neg\mathbf{C}, \mathbf{C} \sqcap \mathbf{D}, \mathbf{C} \sqcup \mathbf{D}, \exists \mathbf{r}.\mathbf{C}, \forall \mathbf{r}.\mathbf{C}$ are concepts;
- (ii) $\mathbf{C}(\mathbf{c}), \mathbf{r}(\mathbf{c}, \mathbf{d}), \mathbf{C} \sqsubseteq \mathbf{D}, \neg\varphi, \varphi \rightarrow \psi$ are statements.

A model M is a pair (Δ, I) such that Δ is a non-empty set, and I is an interpretation such that for each concept name \mathbf{C} , $\mathbf{C}^I \subseteq \Delta$; for each role name \mathbf{r} , $\mathbf{r}^I \subseteq \Delta \times \Delta$; for each object name \mathbf{c} , $\mathbf{c}^I \in \Delta$.

The interpretation of concepts are defined as follows:

$$\begin{aligned} \top^I &= \Delta; \\ \mathbf{C}^I &= I(\mathbf{C}); \end{aligned}$$

$$\begin{aligned}
(\neg \mathbf{C})^I &= \Delta - \mathbf{C}^I; \\
(\mathbf{C} \sqcap \mathbf{D})^I &= \mathbf{C}^I \cap \mathbf{D}^I; \\
(\exists r.\mathbf{C})^I &= \{x : \exists y \in \Delta((x, y) \in r^I \& y \in \mathbf{C}^I)\};
\end{aligned}$$

The satisfaction of statements are defined as follows:

$$\begin{aligned}
M \models \mathbf{C}(\mathbf{c}) &\text{ iff } \mathbf{c}^I \in \mathbf{C}^I; \\
M \models \mathbf{r}(\mathbf{c}, \mathbf{d}) &\text{ iff } (\mathbf{c}^I, \mathbf{d}^I) \in r^I; \\
M \models \mathbf{C} \sqsubseteq \mathbf{D} &\text{ iff } \mathbf{C}^I \subseteq \mathbf{D}^I; \\
M \models \neg \varphi &\text{ iff } M \not\models \varphi; \\
M \models \varphi \rightarrow \psi &\text{ iff } M \models \varphi \Rightarrow M \models \psi. \quad \square
\end{aligned}$$

By adding concept and role constructors, we could extend \mathcal{ALC} to some complicated DL systems. In next chapter we present an extended DL for relational database.

3 Description Logic \mathcal{MDL}

In this section we present a description logic \mathcal{MDL} for relational databases, then we propose a translation from relational database to \mathcal{MDL} , and show that the translation preserves the information of tuples, relations in relational database.

Calvanese et al.[2,3] proposed a translation from ER-schema to \mathcal{DLR} . In his translation, a tuple entity is taken as a constant in DL, and entity sets are translated into concepts in DL. For each attribute, there is a special 2-arity role in \mathcal{DLR} , whose first component is an entity and the second component is the attribute value of that entity. We use an similar way to express tuples and relations from relational databases in description logic. Tuples in relational database are interpreted as constants in DL, that is, elements of the domain of DL model. A relations, as a set of tuples with the same attributes set, is interpreted as a concept. Like the translation from ER-schemas to \mathcal{DLR} , we take attribute values as constants, and for each attribute name there is a role associated. Binary relations on attribute values are also interpreted as roles.

Note that there are two kind of constants, one for tuple and the other for attribute value. Assume that there is a concept \mathbf{R} in DL, associated with a relation R in a relational database. For a tuple constant \mathbf{t} , statement $\mathbf{R}(\mathbf{t})$ means that tuple t is an element of relation R . But for a attribute value constant \mathbf{v} , $\mathbf{R}(\mathbf{v})$ is just meaningless in the corresponding relational database, though that statement has no syntax error in the traditional DL system. For the similar reason, the role for the binary relations of attribute name, and the role for the tuple-attribute value association cannot be syntactically distinguished.

All of above show that it is necessary to propose a many-sorted logic to distinguish different types of object syntactically. We now present a many-sorted description logic \mathcal{MDL} , the formal description is as follows:

Definition (syntax of \mathcal{MDL}). The language $L_{\mathcal{MDL}}$ for \mathcal{MDL} contains the following primitive symbols:

- tuple constant names $\mathbf{t}_0, \mathbf{t}_1, \dots$;
- value constant names $\mathbf{v}_0, \mathbf{v}_1, \dots$;
- relation concept names $\top_{\mathbf{R}}, \mathbf{R}_0, \mathbf{R}_1, \dots$;
- value concept names \top_V, V_0, V_1, \dots ;
- attribute role names $\mathbf{a}_0, \mathbf{a}_1, \dots$;
- value role names $\mathbf{op}_0, \mathbf{op}_1, \dots$;
- concept and role constructors: $\neg, \sqcap, \exists, \{\}$;
- the subsumption relation: \sqsubseteq ; and
- logical connectives: \neg, \rightarrow .

All the relation concept names are atomic relation concepts, all the value concepts are atomic value concepts, all the attribute role names are atomic attribute roles. If \mathbf{R} and \mathbf{S} are relation concepts, \mathbf{V}, \mathbf{W} are value concepts, \mathbf{a}, \mathbf{b} are attribute roles, \mathbf{op} is value role, φ and ψ are statements, then

- (i) $\neg\mathbf{R}, \mathbf{R} \sqcap \mathbf{S}, \exists \mathbf{a}.\mathbf{V}$ are relation concepts;
- (ii) $\neg\mathbf{V}, \mathbf{V} \sqcap \mathbf{W}, \exists \mathbf{op}.\mathbf{V}, \{\mathbf{v}_0, \dots, \mathbf{v}_n\}$ are value concepts;
- (iii) $\mathbf{a} \sqcap \mathbf{b}, \neg\mathbf{a}$ are attribute role names;
- (iv) $\mathbf{V}(\mathbf{v}), \mathbf{R}(\mathbf{t}), \mathbf{a}(\mathbf{t}, \mathbf{v}), \mathbf{op}(\mathbf{v}_1, \mathbf{v}_2), \mathbf{V} \sqsubseteq \mathbf{W}, \mathbf{R} \sqsubseteq \mathbf{S}, \neg\varphi, \varphi \rightarrow \psi$ are statements.

□

Definition (semantics of \mathcal{MDL}). A model M is a tuple (Δ, Σ, I) where Δ, Σ are non-empty sets such that $\Delta \cap \Sigma = \emptyset$ and I is an interpretation such that for each relation concept name \mathbf{R} , $\mathbf{R}^I \subseteq \Delta$; for each value concept name \mathbf{V} , $\mathbf{V}^I \subseteq \Sigma$; for each attribute role name \mathbf{a} , $\mathbf{a}^I \subseteq \Delta \times \Sigma$; for each value role name \mathbf{op} , $\mathbf{op}^I \subseteq \Delta \times \Delta$; for each tuple constant name \mathbf{t} , $\mathbf{t}^I \in \Delta$ and for each value constant name \mathbf{v} , $\mathbf{v}^I \in \Sigma$;

The interpretation of concepts and roles are defined as follows:

$$\begin{aligned}
\top_{\mathbf{R}}^I &= \Delta; \\
\top_{\mathbf{V}}^I &= \Sigma; \\
(\neg\mathbf{R})^I &= \Delta - \mathbf{R}^I; \\
(\mathbf{R} \sqcap \mathbf{S})^I &= \mathbf{R}^I \cap \mathbf{S}^I; \\
(\neg\mathbf{V})^I &= \Sigma - \mathbf{V}^I; \\
(\mathbf{V} \sqcap \mathbf{W})^I &= \mathbf{V}^I \cap \mathbf{W}^I; \\
(\exists \mathbf{a}.\mathbf{V})^I &= \{x \mid \text{exists } y \in \Sigma, (x, y) \in \mathbf{a}^I \& y \in \mathbf{V}^I\}; \\
(\exists \mathbf{op}.\mathbf{V})^I &= \{x \mid \text{exists } y \in \Sigma, (x, y) \in \mathbf{op}^I \& y \in \mathbf{V}^I\}; \\
(\{\mathbf{v}_0, \dots, \mathbf{v}_n\})^I &= \{x \mid x = \mathbf{v}_i^I \text{ for some } i\}; \\
(\neg\mathbf{a})^I &= \Delta \times \Sigma - \mathbf{a}^I; \\
(\mathbf{a} \sqcap \mathbf{b})^I &= \mathbf{a}^I \cap \mathbf{b}^I;
\end{aligned}$$

The satisfaction of statements are defined as follows:

$$\begin{aligned}
M \models \mathbf{R}(\mathbf{t}) &\text{ iff } \mathbf{t}^I \in \mathbf{R}^I; \\
M \models \mathbf{V}(\mathbf{v}) &\text{ iff } \mathbf{v}^I \in \mathbf{V}^I;
\end{aligned}$$

$$\begin{aligned}
M \models \mathbf{a}(\mathbf{t}, \mathbf{v}) & \text{ iff } (\mathbf{t}^I, \mathbf{v}^I) \in \mathbf{a}^I; \\
M \models \mathbf{p}(\mathbf{v}_1, \mathbf{v}_2) & \text{ iff } (\mathbf{v}_1^I, \mathbf{v}_2^I) \in \mathbf{p}^I; \\
M \models \mathbf{R} \sqsubseteq \mathbf{S} & \text{ iff } \mathbf{R}^I \subseteq \mathbf{S}^I; \\
M \models \neg\varphi & \text{ iff } M \not\models \varphi; \\
M \models \varphi \rightarrow \psi & \text{ iff } M \models \varphi \Rightarrow M \models \psi. \quad \square
\end{aligned}$$

4 Represent Information from Relational Databases in \mathcal{MDL}

Based on the formal definition of \mathcal{MDL} , We now present a translation σ from relational database to \mathcal{MDL} , which could be divided into syntactical and semantical parts: given a database d , the syntactical part of σ translate d to a collection of \mathcal{MDL} statements $KB(d)$; the semantical part of σ translate d to an \mathcal{MDL} model $\sigma(d)$.

Definition (syntactical part of σ). Given a database $d = \{r_1, \dots, r_n\}$, for each tuple $t \in Ur_i$, the \mathcal{MDL} sub-language L_d for d includes a corresponding tuple constant name \mathbf{t} . Similarly, for each attribute value v , attribute name A , binary relation op and relation r_i ; we have \mathbf{v} as value constant, \mathbf{a} as attribute role, \mathbf{R}_i as relation concept and \mathbf{op} as value role in L_d respectively. In addition, for each attribute A , we have \mathbf{D}_A as a value concept, denotes the concept containing values in the domain of A . For a relation r in database, we use $\sigma(r)$ to denote the corresponding tuple concept in \mathcal{MDL} ; if r is the atomic relation in the original database, then $\sigma(r) = \mathbf{R}$.

We construct $KB(d)$ as follows:

- For each relation $r \in d$ and tuple t such that $t \in r_i$ in relational database, the \mathcal{MDL} statement $\mathbf{R}_i(\mathbf{t}) \in KB(d)$;
- For each tuple t , attribute value v and attribute A such that $t(A) = v$, $\mathbf{a}(\mathbf{t}, \mathbf{v}) \in KB(d)$ and $\mathbf{D}_A(\mathbf{v}) \in KB(d)$;
- For each relation $r \in d$ such that $\alpha(r) = \{A_1, \dots, A_n\}$ and $U - \alpha(r) = \{B_1, \dots, B_m\}$,
 $\mathbf{R}_i \sqsubseteq \exists \mathbf{a}_1. \mathbf{D}_{A_1} \sqcap \dots \sqcap \exists \mathbf{a}_n. \mathbf{D}_{A_n} \sqcap \neg \exists \mathbf{b}_1. \mathbf{D}_{B_1} \sqcap \dots \sqcap \neg \exists \mathbf{b}_m. \mathbf{D}_{B_m} \in KB(d)$.

□

From the definition of $KB(d)$, we know that all the information about how tuples belong to relations and which value associated as an attribute value of a tuple, are translated into $KB(d)$, $KB(d)$ captures the complete static information of a relational database.

Definition (semantical part of σ). Given a database $d = \{r_1, \dots, r_n\}$, we construct an \mathcal{MDL} model $\sigma(d) = (\Delta, \Sigma, I)$, such that:

- $\Delta = \bigcup_{1 \leq i \leq n} r_i$, that is, for each tuple t in database, t is an element of the domain of $\sigma(d)$;

- $\Sigma = \bigcup_{1 \leq i \leq n} \bigcup_{A \in \alpha(r_i)} \text{dom}(A)$, including all the attribute value in the database;
- I is the interpretation fulfills:
 - for each tuple constant \mathbf{t} , $I(\mathbf{t}) = t$, t is the corresponding tuple in relational database;
 - for each value constant \mathbf{v} , $I(\mathbf{v}) = v$;
 - for each tuple concept \mathbf{R} , $I(\mathbf{R}) = r$;
 - for each attribute role \mathbf{a} , $I(\mathbf{a}) = \{(t, v) | t(A) = v\}$
 - for value concept \mathbf{D}_A , $I(\mathbf{D}_A) = \text{dom}(A)$;
 - for each value role \mathbf{op} , $I(\mathbf{op}) = \{(v_1, v_2) | v_1 \text{ op } v_2\}$. □

The translation σ is faithful, that is, all the statements in $KB(d)$ are satisfied in the \mathcal{MDL} model $\sigma(d)$. Formally, we have following theorem:

Theorem. For each $\varphi \in KB(d)$, $\sigma(d) \models \varphi$.

Proof. By induction on the construction of \mathcal{MDL} statements, the detailed proof is omitted here. □

The theorem above shows that we could use \mathcal{MDL} to express the static information of a relational database. We now show some relational algebra operations could also be represented in \mathcal{MDL} .

Firstly, we consider the set-theoretic operations: union \cup and difference $-$. Let r_1, r_2 be relations in relational database, and $\mathbf{R}_1, \mathbf{R}_2$ be the corresponding tuple concept in \mathcal{MDL} . Use the concept constructor \sqcap , we obtain the tuple concept $\mathbf{R}_1 \sqcap \mathbf{R}_2$. It is straightforward to show the relation $r_1 \cup r_2$ is the corresponding relation of tuple concept $\neg(\neg\mathbf{R}_1 \sqcap \neg\mathbf{R}_2)$. Formally we have following proposition:

Proposition. $\sigma(r_1 \cup r_2) = \neg(\neg\mathbf{R}_1 \sqcap \neg\mathbf{R}_2)$. That is, if a tuple t is in relation $r_1 \cup r_2$, then $(\neg(\neg\mathbf{R}_1 \sqcap \neg\mathbf{R}_2))(\mathbf{t}) \in KB(d)$, $\sigma(d) \models (\neg(\neg\mathbf{R}_1 \sqcap \neg\mathbf{R}_2))(\mathbf{t})$. □

Similarly, we have the following proposition:

Proposition. $\sigma(r_1 - r_2) = \mathbf{R}_1 \sqcap \neg\mathbf{R}_2$. □

Note that the above proposition is satisfied only if $\alpha(r_1) = \alpha(r_2)$. If $\alpha(r_1) \neq \alpha(r_2)$, the relation $r_1 - r_2$ is undefined, while $\mathbf{R}_1 \sqcap \neg\mathbf{R}_2 \equiv \mathbf{R}_1$.

Besides of set-theoretic operations, the selection of relations could also be expressed in \mathcal{MDL} . Given a relation r , an attribute $A \in \alpha(r)$ and a value $v \in \text{dom}(A)$, we consider $\theta_{A=v}(r)$. Using enumerate constructor $\{\}$ of concept, we could obtain $\{\mathbf{v}\}$ as a singleton concept of value. It is easy to show the concept $\exists\mathbf{a}.\{\mathbf{v}\}$ corresponds to the set of tuple in relational database, whose element has value v of attribute A . According to the definition of selection operator, $\mathbf{R} \sqcap (\exists\mathbf{a}.\{\mathbf{v}\})$ is the concept corresponding to $\theta_{A=v}(r)$. We could construct concepts for $\theta_{A \text{ op } v}$ and $\theta_{A=B}$ in a similar way. We have following proposition:

Proposition. Selection operation could be expressed in \mathcal{MDL} as follows:

$$\begin{aligned}
 \sigma(\theta_{A=v}(r)) &= \mathbf{R} \sqcap (\exists\mathbf{a}.\{\mathbf{v}\}); \\
 \sigma(\theta_{A \text{ op } v}(r)) &= \mathbf{R} \sqcap (\exists\mathbf{a}.\{\exists\mathbf{op}.\{\mathbf{v}\}\}); \\
 \sigma(\theta_{A=B}(r)) &= \mathbf{R} \sqcap (\exists(\mathbf{a} \sqcap \mathbf{b}).\mathbf{D}_A).
 \end{aligned}$$
□

Notice that the set-theoretical operations and selection are closed to the set of tuples in the original database. That is, there will be no new tuples due to these operations. We now consider the join operation. Given relations r_1, r_2 , tuples in $r_1 \bowtie r_2$ differs from those tuples from the original relations r_1, r_2 : the attribute set of $r_1 \bowtie r_2$ is $\alpha(r_1) \cup \alpha(r_2)$. Since the domain Δ of tuple only includes tuples which are from the original database, We cannot express the new tuple gained from join operation, like $t \in r_1 \bowtie r_2$. Similarly, the rename and projection operations are not closed to tuples. That means without making any extension to \mathcal{MDL} , we cannot express these operations.

5 Conclusion

In this paper we present a many-sorted description logic \mathcal{MDL} for relational databases. Based on \mathcal{MDL} , we present a translation σ from relational database to \mathcal{MDL} , which could be divided into syntactical and semantical parts: given a database d , the syntactical part of σ translate d to a collection of \mathcal{MDL} statements $KB(d)$; the semantical part of σ translate d to an \mathcal{MDL} model $\sigma(d)$. This translation is faithful, which means all the static information of tuples, relations and attribute values in database could be expressed in \mathcal{MDL} . In addition, we show some relational algebra operations could be expressed in \mathcal{MDL} as concepts.

References

- [1] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook. Cambridge University Press, Cambridge (2002)
- [2] Borgida, A., Lenzerini, M., Rosati, R.: Description logics for data bases. In [1], pp.472–494
- [3] Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., Rosati, R.: Description logic framework for information integration. In KR 1998, pp. 2–13 (1998)
- [4] Borgida, A.: Description logics for querying databases. In: DL 1994, Bonn, Germany (1994)
- [5] Borgida, A.: Description logics in data management. IEEE Transactions on Knowledge and Data Engineering 7, 671–682 (1995)
- [6] Codd, E.F.: Relational completeness of database sublanguages. In: Rustin, R. (ed.) Data Base Systems, pp. 65–98. Prentice Hall, Englewood Cliffs (1972)
- [7] Simovici, D.A., Tenney, R.L.: Relational Database Systems. Academic Press, London (1995)
- [8] Kanellakis, P.C.: Elements of relational database theory. Brown U. Technical Report (1988)
- [9] Lenzerini, M.: Description logics and their relationships with databases. In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 32–38. Springer, Heidelberg (1998)

Non-Functional Requirements Elicitation and Incorporation into Class Diagrams*

Xiaoyu Song, Zhenhua Duan, and Cong Tian

Institute of Computing Theory and Technology, and ISN Laboratory, Xidian University, Xi'an,
710071, P.R. China

yming.song@gmail.com, {zhhdian, ctian}@mail.xidian.edu.cn

Abstract. Top-quality software architecture should consider both functional and non-functional aspects of systems and their association. In the the existing literature, considerable efforts have been directed at functional requirement analysis and design, regardless of the non-functional aspects. This disassociation makes architecture comprehension and evolution hard. This paper proposes a strategy on how to elicit non-functional requirements and incorporate them into the design models of functions. We aim at bridging the gap between functionality and non-functionality and constructing high quality software systems.

1 Introduction

The software development is concerned with two different aspects of a system: functional and non-functional ones. The development of functions has made great progress, and it has been supported by many development approaches. In contrast, there are only a few researchers working on non-functions, and even no tools completely support the non-functional development in software engineering. However, non-functional requirements (NFRs) have an important effect on system quality and development costs. The well-known case of the London Ambulance System (LAS) is a good example [1]. The LAS was deactivated, because of several problems related to non-functional requirements. The market is increasing its demands on software that not only implements all of functionalities but also copes with non-functional aspects such as reliability, security, accuracy, safety, performance as well as others [2].

NFRs, in spite of importance, are usually hidden in developers mind. However, it does not mean that software engineers cannot consider information about non-functional requirements [1]. Moreover, NFRs are always linked to functional requirements (FRs). We raise a systematic methodology for software development. It contains the following main processes. First, we capture all requirements from users and customers, and form requirement documents; secondly, we analyze requirement documents and elicit both FRs and NFRs; thirdly, we construct function models based on FRs; fourthly, we incorporate the elicit NFRs into function models. This results in the association of system's

* This research is supported by the NSFC Grant No. 61003078, 60433010, 60873018 and 60910004, DPRPC Grant No. 51315050105, 973 Program Grant No. 2010CB328102 and SRFDP Grant No. 200807010012.

functionality and non-functionality, improving the quality of software and the speed of development.

In this paper, we adopt an object-oriented approach based on UML. It gives an expression to system’s functions. The *use case model* of UML constitutes a suitable description of system’s functionality [3] while *class diagrams* of UML express the design model of system’s functions.

Our work focus on eliciting NFRs and incorporating NFRs into the design models of system’s functions. To elicit NFRs, we first extract the words or phrases regarding non-functional aspects in requirement documents; then we define and enrich the resulted NFRs. The incorporation process is to incorporate NFRs into class diagrams.

The remainder of this paper is constructed as follows: Section 2 introduces the related basic concepts. Section 3 depicts the strategy and the process of the elicitation and integration. Section 4 gives an example. Finally, the conclusion and future work are discussed.

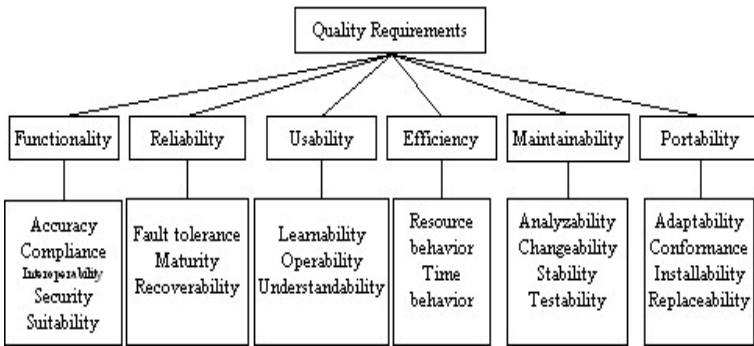


Fig. 1. ISO/IEC 9126 Taxonomy of Quality Requirement

2 Basic Concepts

2.1 Non-Functional Requirements

Software requirements consist of two different requirements, the functional requirements and the non-functional requirements. FRs focus on “what” the system must do while NFRs constrain “how” the system can accomplish the “what”. So it is said that NFRs are always related to FRs in software development.

NFRs are also known as Quality Requirements [4]. A set of ISO/IEC standards are related to software quality, being standards number 9126 (which is in process of substitution by 9126-1, 9126-2 and 9126-3), 14598-1 and 14598-4 the more relevant ones [5]. The ISO/IEC 9126 standard prescribes the top characteristics: functionality, reliability, usability, efficiency, maintainability and portability. And the top characteristics can be future refined into subcharacteristics (see figure 1). The top characteristics of NFRs have a high level of abstraction while the subcharacteristics have a greater level of detailed aspects.

Each NFR should belong to a NFR type. In this work, it is decided that the set of subcharacteristics of NFRs instead of the top characteristics is as the foundation of categorizing NFRs.

2.2 NFR Card

An NFR card is used to record information on an NFR, metaphorized into the belt to connect an NFR and class diagrams.

An NFR card is composed of four parts, NFR symbol, NFR property, NFR behavior and Incorporated Class (see figure 4 and figure 5). An NFR symbol is the name or alias of the meaningful word or phrase referring to NFR, while an NFR type expresses the category an NFR belongs to. The property an NFR possesses is called an NFR property and the behavior satisfying an NFR is called NFR behavior. Incorporated class is what NFRs can be incorporated into. The NFR property plays a role as the class attribute does in a class. Each NFR should contain at least one NFR behavior which implements the NFR. There may be no incorporated Class, because each class in class diagrams may be not related to the NFR. However, it is not to say that the incorporation is failure. In this case, a new class on NFR can be inserted into class diagrams. The detail will be described in Section 3.2.

3 The Proposed Strategy

In this section, we propose a strategy to deal with NFRs from eliciting to incorporating into class diagrams. NFRs can be elicited based on requirement documents, because these documents may contain the words or phrases referring to NFRs. Then an NFR card will be constructed as a belt to connect the elicitation process and the incorporation process. In an NFR card, it is indicated that what properties and behaviors are responsible for satisfying the NFR, and what class in class diagrams is related to the NFR. Once finding out the incorporated class, we insert the NFR into the class. But there may be an exception that the incorporated class does not exist. In this case, we will add a new class which satisfies the NFR to class diagrams. Figure 2 shows the structure for our strategy.

3.1 Eliciting Non-Functional Requirements

Our approach is useful for eliciting non-functional requirements. NFRs are not as clear in stakeholders' minds as functional requirements, so eliciting NFRs calls for the apprehension of the domain and the accumulation of knowledge and skills.

The good approach to elicit NFRs is to read those accomplished requirement documents carefully and identify NFRs with the accumulated experience. These documents record all requirements desired by users and customers and usually contain some words or phrases related to NFRs. The first step is to extract these words or phrases from requirement documents. Then, we refine them into the NFR symbols. One NFR symbol should express an object which possesses non-functional characteristics. The NFR symbols are just the possible ones and they need to be validated by users and customers later on.

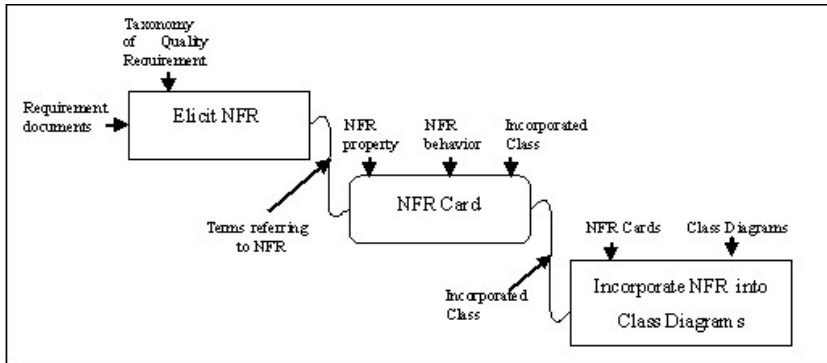


Fig. 2. The structure for our strategy

Each NFR should be placed in an NFR type. As mentioned in Section 2.1, one NFR type is one sub-characteristic of Quality requirements. We categorize all possible NFR symbols to different NFR types, such as accuracy, security and testability.

All NFR properties and NFR behaviors for a possible NFR symbol may not be perceived in the first time. They can be completely captured by interviewing the related stakeholders or sending them questionnaires.

It is very necessary to validate these already elicited NFRs, and their properties and behaviors, because accuracy of these NFRs has a vital impact on the software architecture. It needs the attendance of the related stakeholders to confirm which NFRs elicited are the ones desired by users and customers. Moreover, the correctness and maturity of NFR properties and NFR behaviors should be validated any more.

In order to obtain authentic NFRs, those activities, such as extracting NFRs, capturing NFR properties and behaviors and validating NFRs, should be carried out several times.

The next step is to build the NFR card for each NFR. An NFR card describes the information on an elicited NFR. The NFR symbol is used to name the NFR and a sub-characteristic of quality requirements is marked as the NFR type. In addition, NFR properties and NFR behaviors should be organized and formalized in the NFR card. Those collected NFR properties and behaviors from users and customers may be informal at first. In order to facilitate the incorporation, they have to be formalized.

The format of NFR property is

PropertyName : *propertyType*, *set value* = *propertyValue*

The format of NFR behavior is

BehaviorName : *Parameter1*, ..., *ParameterN*

For instance, there is an informal sentence describing NFR property, "Response time of the transaction must be not longer than 5s".

In accordance with the format above, we can acquire

ResponseTime : *s*, *set value* = 5

3.2 Incorporating NFRs into Class Diagrams

Incorporating Non-functional requirements into class diagrams is the significant contribution of our work. Eliciting NFRs is important but is not enough [1]. We have to incorporate NFRs into the design models of system's functions. As class diagrams give the primary functional representation for a system, the incorporation of NFRs into them will make the system architecture more systematic.

An NFR card plays an important role for the proposed strategy, as output of the elicitation and input of the incorporation. In the elicitation process, the NFR symbol, properties and behaviors have been accomplished. Hence, searching for the incorporated class in class diagrams is the constant activity. Since system non-functionality needs to refer to system functionality and class diagrams are usually used to model system's function, so there may be a relation between class diagrams and NFRs. This is to say, class diagrams may contain the incorporated class that NFRs can be incorporated into. However, it is unassailable to find out the incorporated class for every NFR. The incorporated class for some NFR is not present.

We discuss the incorporation process in two cases: the one case is that the incorporated class exists and the other case is that the incorporated class does not exist.

Suppose that the incorporated class exists and it has been appended to the NFR Card. Based on the incorporated class, the NFR properties and behaviors will be inserted into the class as new attributes and new operations. In order to distinguish the new attributes and operations from the old ones in a class, we attach the postfix {NFR-Type [NFR-Name]} to the new attributes and operations, for instance, Generate-Alarm () {Security [Credit Card]}.

We propose to use the following processes to incorporate NFRs into class diagrams.

1. Pick up an NFR card and search for the class that the NFR can be incorporated into. Since the NFR symbol is identified from requirement documents and class diagrams are also constructed based on these documents, it is possible that some class contains the NFR symbol. We consider the NFR symbol as the starting point to search for the incorporated class, because the symbol from requirement documents may be related to both the functional aspects and the non-functional aspects. Of course, the NFR properties and behaviors may be also helpful. When found out, the class would be marked as the incorporated class in an NFR card.

2. Incorporate the NFR into the incorporated class. Each NFR property is translated into a new attribute while each NFR behavior is translated into a new operation. The NFR symbol combines with NFR type to form the postfix {NFR-Type[NFR-symbol]}. Both NFR property and NFR behavior must be formalized because it is convenient for the automation of the incorporation (see figure 3).

Some NFRs are not directly related to any class. We call those requirements global non-functional requirements (GNFRs). Incorporating GNFRs into class diagrams is also absolutely necessary for software development. In this case, every GNFR will be translated into a new class with NFR symbol as class name. Furthermore, the NFR properties and behaviors are also translated into class attributes and operations as mentioned above.

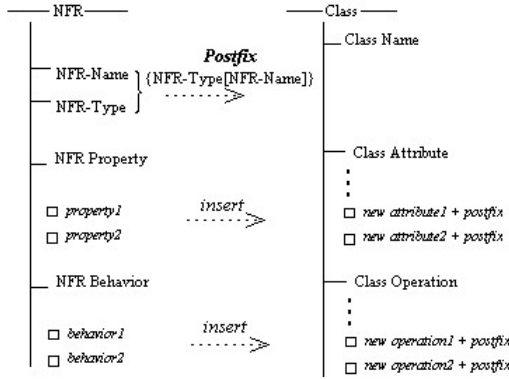


Fig. 3. Incorporate NFR into class diagram

After incorporating NFRs into class diagrams, we propose to adjust class diagrams, because the addition of NFRs can lead to some changes of the structure of class diagrams. However, for the limitation of space, we do not discuss it in detail in the paper.

4 An Example for the Proposed Strategy

The example used to validate the proposed strategy is about the development of a credit card system. Here is a segment of requirement documents for this system. "In the Credit Card system, it is necessary to protect against the vicious access to the system. Security of the transaction must be satisfied. ... Moreover, the response time of the transaction is not longer than 5s".

We start to analyze the above sentence to identify the non-functional requirements for this system. After analyzing thoroughly, we find the phrases "security of the transaction" and "response time of the transaction" are related with NFRs. So we obtain two NFRs: i) security of the transaction; ii) response time of the transaction.

As each NFR has an NFR card, we have to construct two NFR cards. In the example, it is accidental that both of the NFRs are based on the same object *transaction*, so the word *transaction* is considered as NFR symbol for the two NFRs. *Security* and *Time behavior* are two NFRs Types.

The next step is to confirm NFR properties and NFR behaviors. This step depends on not only software engineers but also the other stakeholders through interviewing them or sending them questionnaires. For the NFR with the type *Security*, we discover that "alarm" is used to notify the proper authority of all vicious accesses to the credit card. So there is one NFR behavior "send alarm" to implement the NFR. For the NFR with the type *Time behavior*, the property *Response Time* and the behavior *halt the transaction* are asked to satisfy the requirement about response time of the system.

Validating NFR properties and NFR behaviors is necessary, and it requires all related stakeholders to work together. In our work, the result of NFRs validation demonstrates that the identified NFR property and behaviors are correct.

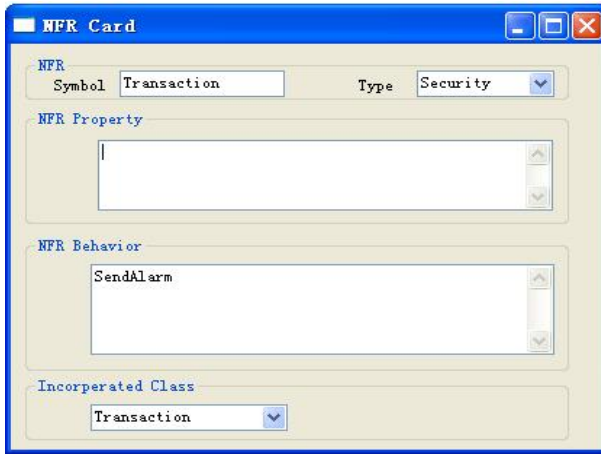


Fig. 4. The NFR Card with type Security

Accomplishing the NFR Cards is the next activity. It has to be done to formalize the NFR properties and behaviors as introduced in Section 3.1. Figure 4 and 5 portray two NFR cards.

The elicited NFRs will be incorporated into class diagrams. Figure 6 shows a partial class diagram for the Credit Card system. There is a class whose name is the same as the NFR symbol in this diagram and there is a close relation between the class and two NFRs, so we consider class *Transaction* as the incorporated class. until now, the two NFR Cards have been constructed successfully.

For the NFR with *Security* type, the behavior *SendAlarm* is translated into a class operation *SendAlarm()*. And the postfix $\{\text{Security} [\text{Transaction}]\}$ will be attached to the operation. For the NFR with *Time* behavior type, the property *Respose Time* is translated into a class attribute *ResponseTime* while the behavior *halt the transaction* is translated into a class operation *HaltTransaction()*. Their postfixes are $\{\text{Security} [\text{Time behavior}]\}$; Figure 7 shows the class *Transaction* after the NFR integration.

5 Related Works

In spite of its importance, NFRs have surprisingly received little attention in the literature, and they are poorly understood in contrast with other less critical aspects of the software development [6]. Recently, some researches concerning NFRs have made progress. There are two main approaches in the software development concerning with NFRs. On one hand, the majority of the studies on NFRs use a product-oriented approach, which is mainly concerned with how much a software is in accordance with the set of NFRs that it should satisfy [7] [8] [9] [10]. On the other hand, there are also a little studies proposing to use a process-oriented approach in order to explicitly deal with NFRs. Unlike the product-oriented approach, this approach is concerned with making NFRs a relevant and important part of the software development process. Chung's

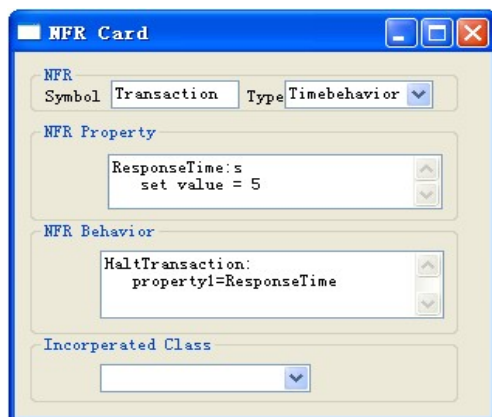


Fig. 5. The NFR Card with type Time behavior

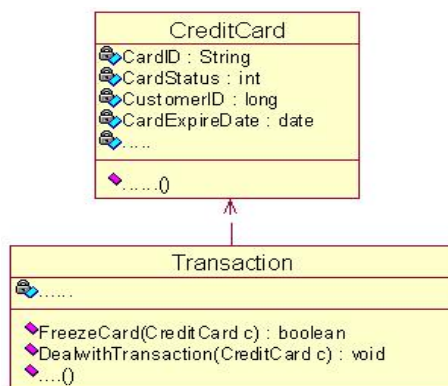


Fig. 6. A partial class diagram of the Credit Card system

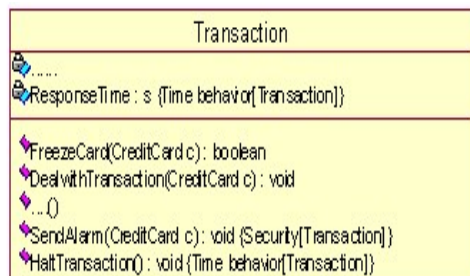


Fig. 7. Class Transaction after NFR integration

Framework [6] is one of complete works. In his work, NFRs are viewed as goals that are decomposed into subgoals until one finds that all the necessary actions and information are already represented at the leaf levels of the graphs. Although Chung's work aims to represent NFRs and their conflicts, it does not consider the association with functional requirements. In our work, we believe that the systematic elicitation and integration of NFRs into design models may be helpful in software development.

6 Conclusion

Errors resulting from NFRs are the most expensive and difficult to correct [11], and not dealing or improperly dealing with NFRs can lead to more expensive software and a longer time-to-the market [12]. So it is strongly demanded to satisfy all NFRs for a system and deal with both FRs and NFRs and their association through the entire development process.

Our current work fills the gap between system functionality and system non-functionality in software development, and also applies NFRs to one design model of system's functions. We raise a systematic methodology for software development. The paper mainly focus on the strategy to elicit NFRs and incorporate NFRs into class diagrams. The result shows that integrating NFR into the design of system's functions can consummate system architecture, improve system quality and reduce failure risk.

However, we only consider the incorporation between NFRs and class diagrams. In the future, we will extend the strategy of the incorporation to the other design models, such as sequence diagrams and state machine diagram. We have developed a prototype to support the proposed strategy. However, it has not achieved automation of the incorporation process and still requires software engineer's attendance. Further, we will improve the tool so that our strategy can be better supported.

References

1. Cysneiros, L.M., Leite, J.C.S.P., Neto, J.S.M.: A Framework for Integrating Non-Functional Requirements into Conceptual Models. In: Requirements Eng., vol. 6, pp. 97–115 (2001) @ 2001 Springer-Verlag London Limited
2. Mala, G.S.A., Uma, G.V.: Requirement Preference for Actors of Usecase from Domain Model. In: Hoffmann, A., Kang, B.-h., Richards, D., Tsumoto, S. (eds.) PKAW 2006. LNCS (LNAI), vol. 4303, pp. 238–243. Springer, Heidelberg (2006)
3. Kim, H.-K., Chung, Y.-K.: Automatic Translation from Requirements Model into Use Cases Modeling on UML. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005. LNCS, vol. 3482, pp. 769–777. Springer, Heidelberg (2005)
4. Chung, L.: Representing and Using Non-Functional Requirements: A Process Oriented Approach. Ph.D. Thesis, Dept. of Comp. Science. University of Toronto, Also tech. Rep. DKBS-TR-91-1 (June 1993)
5. ISO/IEC Standards 9126 (Information Technology - Software Product Evaluation - Quality Characteristics and Guidelines for their use, 1991) and 14598 (Information Technology - Software Product Evaluation: Part 1, General Overview; Part 4, Process for Acquirers; 1999)
6. Chung, L., Nixon, B., Yu, E., Mylopoulos, J.: Non-Functional Requirements in Software Engineering. Kluwer Academic Publishers, Dordrecht (2000)

7. Fenton, N.E., Pfleeger, S.L.: *Software Metrics. A Rigorous and Practical Approach*. International Thomson Computer Press (1997)
8. Kirner, T.G., Davis, A.M.: *Nonfunctional Requirements of Real-Time Systems*. *Advances in Computers* 42, 1–37 (1996)
9. Lyu, M.R. (ed.): *Handbook of Software Reliability Engineering*. McGraw-Hill, New York (1996)
10. Musa, J., Lannino, A., Okumoto, K.: *Software Reliability: Measurement, Prediction, Application*. McGraw-Hill, New York (1987)
11. Brooks Jr., F.P.: *No Silver Bullet. Essences and Accidents of Software Engineering*. *IEEE Computer* (4), 10–19 (1987)
12. Cysneiros, L.M., Leite, J.C.S.P.: *Using UML to Reflect Non-Functional Requirements*. In: *Proc.of the CASCON 2001, Toronto (November 2001)*

Architectures Integrating Case-Based Reasoning and Bayesian Networks for Clinical Decision Support

Tore Bruland, Agnar Aamodt, and Helge Langseth

The Norwegian University of Science and Technology (NTNU)
NO-7491 Trondheim, Norway

Abstract. In this paper we discuss different architectures for reasoning under uncertainty related to our ongoing research into building a medical decision support system. The uncertainty in the medical domain can be divided into a well understood part and a less understood part. This motivates the use of a hybrid decision support system, and in particular, we argue that a Bayesian network should be used for those parts of the domain that are well understood and can be explicitly modeled, whereas a case-based reasoning system should be employed to reason in parts of the domain where no such model is available. Four architectures that combine Bayesian networks and case-based reasoning are proposed, and our working hypothesis is that these hybrid systems each will perform better than either framework will do on its own.

1 Introduction

The field of knowledge-based systems has over the years become a mature field. This is characterized by the availability of a set of methods for knowledge representation, inference, and reasoning that are well understood in terms of scope, strengths, and limitations. Numerous applications have been built that are in daily use, and hence have proven the various methods' value for intelligent decision support systems and other applications. As the individual method areas get more explored and better understood, the identification of limits and strengths opens up for integration of individual methods into combined reasoning systems.

The history of knowledge-based decision support systems, e.g. expert systems, started with rule-based systems. They were followed by systems that tried to open up the “if-then” association to look for what underlying knowledge, in terms of “deeper relationships” such as causal knowledge, could explain the rule implications [1]. Cognitive theories in the form of semantic networks, frames, and scripts formed the theoretical basis for many of these model-based systems. Statistical and probabilistic theories formed another method path. As the availability of data has increased over the recent years, and methods and algorithms for probabilistic reasoning have significantly evolved, probabilistic models, and in particular those based on Bayesian theory in one way or the other, have come to dominate the model-based method field [2]. Bayesian Networks (BN) is the

most prominent among these. It is particularly interesting in that it combines a qualitative model part and a quantitative model part [3].

Both rules and deeper models represent knowledge as generalized abstractions. A good knowledge model is therefore dependent on a human domain expert to construct the model, or on methods that can generalize the model from data. In either case, details about actual observations in the world are abstracted away in the model building phase, without knowing whether some of this specific information could be useful in the problem solving phase. The third and most recent basic type of reasoning in the history of knowledge-based systems addresses this problem by representing each problem instance as a unique piece of situation-specific knowledge, to be retrieved and reused for solving similar problems later [4]. Since its birth in the early 80s, the field of case-based reasoning (CBR) has grown to become a major contributor to decision support methods in academia as well as for industrial applications [5,6,7]. Increased availability of data on electronic form has also contributed to the growth of this field.

Although some early attempts have been made to discuss possible combinations of the two, including our own [8], our current research agenda represents a much larger and more comprehensive effort. Our focus in the work presented here is on improved support for clinical decision making. We are cooperating with the Medical Faculty of our university and the St. Olavs Hospital in Trondheim. More specifically we are working with the European Research Center for Palliative Care, located in Trondheim, in order to improve the assessment, classification and treatment of pain for patients in the palliative phase [9].

Decision making in medicine is to a large degree characterized by uncertain and incomplete information. Still, clinicians are generally able to make good judgments based on the information they have. Decision making under uncertainty – in the wide sense of the term – is therefore our setting for analysing the properties of BN and CBR, aimed at how they can be integrated to achieve synergy effects.

In the following chapter, decision making under uncertainty and the essentials of BN and CBR are characterized. Related research on combined methods are summarized in chapter 3. We discuss relevant combinations of the two methods in chapter 4, by outlining four specific architectures that utilize different properties of the two methods. In chapter 5 we give an example that illustrates one of the architectures, within a simplified, non-medical “toy” domain. The last chapter summarizes the results so far and points out future work.

2 Decision-Making under Uncertainty and Incompleteness

Our motivation for integrating BN and CBR is that they both contribute to improved decision making under incomplete information and with uncertain knowledge. They are both advocated as methods that to some extent address problems in *weak theory domains*. A weak theory domain is a domain in which relationships between important concepts are uncertain [10]. Statements are more or

less plausible, and stronger or weaker supported, rather than true or false. Examples of weak theory domains are medical diagnosis, law, corporate planning, and most engineering domains. A counter-example is a mathematical domain, or a technical artifact built from well-established knowledge of electricity or mechanics.

So, theory strength is one dimension of uncertainty characterization. Another dimension is the *knowledge completeness* of the domain. The fact that a domain has a weak theory does not always imply that there is little knowledge available. Although it may seem somewhat contradictory, weak theory domains need to compensate for lack of strong knowledge with larger amounts of knowledge, which jointly can provide a strengthening or weakening of an hypothesis being evaluated. Inferences in these domains are abductive (in the sense of "inference to the best explanation") rather than deductive, which is a characteristic of strong theories [11]. Three main knowledge types are typically combined in medical diagnosis and treatment: Physiological and pathological theories of various strengths, evidence-based clinical trials from controlled experiments, and person-centric experiences in diagnosing and treating patients [12].

General knowledge, with varying degrees of theory strength, can often be modeled by statistical distributions. The type of uncertainty that deals with assigning a probability of a particular state given a known distribution is referred to as *aleatory uncertainty*. This is a type of uncertainty that fits perfectly with the Bayesian Networks method. Another type of uncertainty, referred to as *epistemic uncertainty*, refers to a more general lack of knowledge, whether stronger or weaker, and are linked to cognitive mechanisms of processing knowledge [13]. Case-based reasoning, on the other hand, has nothing to offer for aleatory uncertainty, but is able to utilize situation-specific experiences as one type of epistemic knowledge.

For decision making under uncertainty, it is important to work with a framework that fits the domain, the available knowledge, the types of uncertainty, and the types of decisions to be made. The strongest theories in the medical sciences are often supported by randomized clinical trials, whereas weak theories lack this basis, and are just as often based on episodic knowledge and previous examples of successful and unsuccessful patient treatments. We are advocating the use of Bayesian Networks to model aleatory uncertainty and some aspects of epistemic uncertainty, and case-based reasoning to handle epistemic uncertainty related to episodic knowledge. We will achieve effects beyond what is possible with one method alone by combining them into a hybrid representation and reasoning framework, and a set of more specific architectures. This is the research hypothesis that guides our work.

Bayesian Networks constitute a modelling framework particularly made for decision making under aleatory uncertainty. Syntactically, a Bayesian network consists of a set of nodes, where each node represents a random variable in the domain, and where there are directed links between pairs of variables. Together, the nodes and arcs define a directed acyclic graph structure. Mathematically, the links and absence of links make assertions about conditional independence

statements in the domain, but for ease of modelling, it is often beneficial to consider a link as carrying information about a causal mechanism [14].

Bayesian Networks can be used for causal inferences (reasoning along the directions of the arc), and for diagnostic inference (reasoning backwards wrt. the causal influences). Recently, there has also been an increased interest in using Bayesian Networks to generate explanations of their inferences (see for instance [15] for an overview).

In case-based reasoning, a collection of past situations or events, i.e. concrete episodes that have happened, make up the knowledge. The concrete episodes are referred to as cases, and the cases - represented in some structural representation language - are stored in a case base, which is a CBR system's knowledge base. The knowledge in a CBR system is therefore situation-specific, as opposed to the generalized knowledge in a BN. A case has two main parts: A problem description and a problem solution. Sometimes a case also includes an outcome part, i.e. the result of having applied the solution to the problem. A CBR system also assigns numerical weights to the features, according to how important a particular feature type or feature value is for characterizing a particular case. In the four-step CBR cycle [5], the RETRIEVE step starts with a problem description and ends when a matching case has been found. REUSE takes that case and tries to build a solution of the new problem, either the easiest way by just copying the solution in the retrieved case over to the new problem, or by making some adaptations to better fit the current problem. In the REVISE step the solution proposed by the system is evaluated in some way, and possibly updated, before RETAIN decides what of this problem solving session should be learned, by updating the case base. A core principle of CBR is the notion of *partial matching*, which means that two cases match to a higher or lesser degree, rather than either match or do not match. Hence, the basic inference method in a CBR system is *similarity assessment*.

On this basis, CBR should be viewed as a method to deal with uncertainty along two dimensions. First, the capturing of domain knowledge as a set of specific experienced situations, rather than general associations, implicitly reflects a degree of uncertainty and incompleteness in the general theories or models of the domain. Second, the similarity assessment procedure that enables the partial matching is a method for reasoning with uncertainty. Uncertainty is captured in the individual feature weights as well as in the computation of total similarity between two cases.

3 Related Research

There is not a large volume of research that describes a combination of BN and CBR methods. Below we have identified five articles of relevance to our architectures, which are presented with a brief description of how they combine BN and CBR.

The earlier Creek system, in which general domain knowledge was represented as a semantic network [16], was extended with a Bayesian component and exemplified by finding the cause of a "car does not start" problem [8]. The semantic

network contains causal links with uncertainty and the probabilistic reasoning is performed by a Bayesian Network. The nodes and causal relations are shared between the semantic network and the BN. The cases are present as variables (on/off) in the BN, and Creek uses the BN to choose relevant cases to apply the similarity measure on (Bayesian case retrieval). The BN is a preprocessing step in the RETRIEVE phase. The Bayesian Network can also calculate causal relations that are used in the adapt method in the REUSE phase.

Tran and Schönwälder [17] describe a distributed CBR system used to find solutions in the communication system fault domain. The problem description is a set of symptoms, S , and the problem solution contains a fault hypothesis, H . Their reasoning process contains two steps: ranking and selection. The ranking step (RETRIEVE phase) finds the most similar cases with their BN relations $S_i|H_j$. The selection step (REUSE phase), use the BN relations $S_i|H_j$ from the cases to build a Bayesian Network. The most probable hypothesis from the BN is chosen.

Gomes [18] presents a computer aided software engineering tool that helps software engineers reuse previous designs. The system combines CBR, BN and WordNet. The cases in the system have a problem description part that contains a number of WordNet synonym sets. The cases are nodes in the Bayesian Network, as are also the synonym sets from the problem description. The synonym sets from the problem description are used to find all the parents from WordNet's *hypernym* relation (is-a), and all the parents are inserted into the BN. The conditional probability tables are built with formulas depending on how many parents the node has. The RETRIEVE phase is performed in three steps as follows: a) the query case description is used to activate (turn on) the synonym sets in the Bayesian Net, b) the BN nodes are calculated and the most relevant cases are found, and c) their probabilities are used to rank the cases.

Bayesian Case Reconstruction (BCR) [19] is a system for design of screening experiments for Macromolecular Crystallization. The domain knowledge is incomplete, the case base is incomplete, there are a large number of variables with a large number of values, and there are limitation on time, material and human resources. BCR is used to expand the coverage of the case base library. The BN is constructed from the domain experts and the content of the case library. The RETRIEVE phase selects the most probable cases, and they are disassembled in order to form new solutions. The Bayesian network contains the causal relations from the domain model that are well understood. In the REUSE phase, the BN is used to find the most probable new solutions. The result is a plausible solution, only.

Another system that combine BN and CBR is used to choose the optimal parameters for an algorithm used in different domains [20]. The case description contains features for an algorithm used on a domain and the case solution is a Bayesian Net. The BN is learned and evaluated through experiments in the domains with the algorithms using different parameter settings. The RETRIEVE phase selects the most similar cases. The REUSE phase is used to calculate a reliability measure in addition to calculate the most probable arguments from the

BN. The most reliable cases are those who have a high number of experiments and a large number of variations in the parameters used.

4 Different CBR and BN Architectures

Case-based reasoning and Bayesian Networks can be combined in the following ways:

- In parallel
- In the sequence BN-CBR
- In the sequence CBR-BN

In the parallel way, both methods use all of the input variables and then produce a classification independently. The results are compared by a “select the best result” algorithm, and the best classification is chosen. Our focus is on integrated approaches, represented by the two sequential combinations. BN and CBR are connected in such a way that the first system in the sequence computes something that the second system needs. The variable types used in the problem domain are as follows:

- I_i is input variable number i . For illustration purpose, see figures 1-4, the variables I_1, I_2 , and I_3 , are used by the BN only and the variables I_5, I_6 are used by the CBR system only. Input variable I_4 is used by both systems.
- A_j is mediating variable number j . The mediating variables represent concepts in the domain model. An expert of the domain can also be a part of the classification process and he can set evidence on a mediating variable.
- D is a variable that is derived by inference from domain knowledge. It is the main output from the BN in the BN-CBR sequence. It can be the solution of a case, as an intermediate result in the CBR-BN sequence architecture.
- C is a classification variable and it can be calculated by a BN or be the final solution of a case.

The user creates a query description of the problem with the input variables.

Two specializations of each sequence type have been developed. The BN-CBR-1 architecture is shown in [1](#). The case identifiers are present as variables in the Bayesian network and they have the binary values on/off that indicates if the case is activated or not. The derived variable in the variable set D is causing the cases to be activated. These D s are *derived* features that are obtained from the input variables by inference based on domain knowledge. Hence, the BN has a filtering role in the RETRIEVE phase of the CBR system. The similarity measures are only applied on the filtered cases. The systems are loosely coupled in this architecture, because the information used in the BN is hidden from the CBR system. An example from the BN-CBR-1 architecture is given in [Section 5](#). The BN-CBR-1 architecture was found in two of the related research articles [8,18](#).

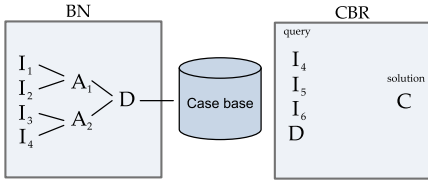


Fig. 1. BN-CBR-1 Architecture

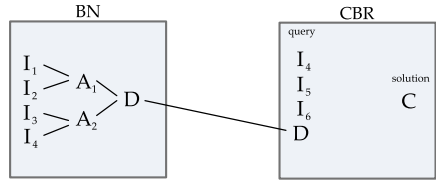


Fig. 2. BN-CBR-2 Architecture

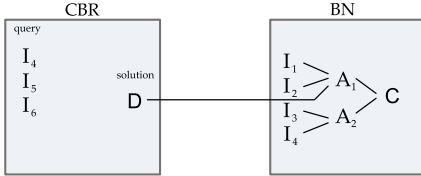


Fig. 3. CBR-BN-1 Architecture

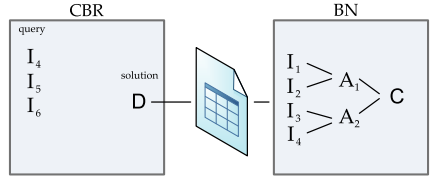


Fig. 4. CBR-BN-2 Architecture

The BN-CBR-2 architecture is shown in Figure 2 and here the systems are tightly coupled. The user states all the input variables, and variable [1..4] are set as evidence in the BN, and the expert set his evidence in the BN. The probabilities of the network are calculated and the D is placed in the case description together with I_1, I_2, I_3, I_4 . The RETRIEVE phase in CBR is performed, resulting in a ranked list of similar case solutions (classification variable). A variant of the BN-CBR-2 architecture was found among the related research articles [8]. In that work, as interpreted by our current framework, the CBR system is the master and the BN the slave. The domain knowledge is represented by the Bayesian network and a semantic net where the variables in the BN are shared with the semantic net. In our approach, the CBR system uses the Bayesian Network in several steps of the reasoning process. For example, the activate step in RETRIEVE (the steps are from the explanation engine [16]) sets some evidence in the BN that activates the relevant cases (BN-CBR-1 architecture). The explain step in RETRIEVE finds the most similar cases. The focus step in RETRIEVE sets new evidence from the case in the BN and finds new casual probabilities that can strengthen information in the semantic net. The BN can also be used in the REUSE phase.

The CBR-BN-1 architecture is shown in Figure 3 and it shows two tightly coupled systems. The CBR system finds a n-best list with the input variables I_4, I_5, I_6 , and the case solution contains the derived variable D . The variable D , the input variables I_1, I_2, I_3, I_4 are set as evidence in the BN, before the posterior probabilities of the classification variable C are calculated. There are two ways to look at the CBR-BN-1 architecture. The first is where the CBR system is a preprocessing step for the BN. The second is where the BN is used in the REUSE phase of CBR. In the first approach, the preprocessing step can be used on a part of the BN model that is unknown. Here an expert can create

cases that replace this unknown BN model. The cases must contain D variables with probabilistic values. Some can also be given by an expert of the domain that is present in the classification process. After the variables are inserted as evidence in the BN, the classification variable is calculated. If the C values are in range of each other there is a possibility of more than one probable class. If the best C value is under a threshold there is no probable class. In the second way of the CBR-BN-1 architecture, the REUSE phase can contain reasoning under uncertainty. The CBR system finds the most similar cases and the REUSE phase can use the BN in order to adapt the case. The classification variable is available in the BN. The CBR-BN-1 architecture was found in one of the related research articles [17].

The CBR-BN-2 architecture is shown in Figure 4. The CBR system uses the input variables I_4, I_5, I_6 to find a solution that contains the most suitable BN model. The BN model is loaded and the input variables I_1, I_2, I_3, I_4 are set as evidence. Afterwards, the classification variable C is calculated. The different BN models has common evidence and classification nodes, but other nodes, causal links, and the conditional probability tables can be different in each model. The information used in the CBR system is hidden from the BN system, and therefore the systems are loosely coupled in this architecture. The CBR-BN-2 architecture was found in one of the related research articles [20].

5 Implementation and Example System

We are currently in the process of analyzing previous recorded clinical data, but this is a time-consuming process, and we do not yet have sufficient results for experimentation with the above architectures. Instead of a medical example we are studying the architectures through a simple movie recommendation system. The sole purpose of the experiment is therefore to study the feasibility of a cooperation between BN and CB methods as suggested by one of the architectures. In the following example the BN-CBR-1 architecture is illustrated.

The system evaluates movies based on comparing the user of the system to a set of previous users, and recommends the favorite movie of the previous user that is most similar to the current user. To make sure that any recommendation is suitable for the current user, a filtering process that removes films that are either unsuitable due to age constraints or excessive violence/nudity must be undertaken. The task of generating a recommendation therefore consists of two subtasks: *i*) Finding the movies that are appropriate for the current user. For this task we have a good understanding of the mechanisms, which makes it suitable for the BN method. *ii*) Among those movies, choose the one that she will most probably enjoy. For this task we have no explicit domain model representing what users like to type of movies, making it fit for a CBR method.

A case consists of a description of a user, in the problem description part, and her favorite movie, in the problem solution part. The user description contains personal features (like **Gender**, **Age**, **Occupation**, and **Favorite Movie Genre**). We have no guarantee that the favorite movies of previous users are suitable

for the current user; still only the appropriate movies in the system should be made available for her. This is ensured by letting the BN take charge of the case activation. The input to the BN is each film's `Age Limit`, `Nudity Level`, and `Violence level` together with the current user's `Age`. The output from the BN is the variable `Suitable` with the values `yes` and `no`. The age categories in the BN are `kid`, `small youth`, `big youth`, and `adult`. The first task for the BN is to let the age groups see movies with the correct age limit only. The second task is to restrict the age groups access to movies with nudity and violence. The kids are not allowed to see any movie with nudity and violence. The small youth is allowed to see a little violence, the big youth can see a little violence and nudity. The adult has no restrictions. Assume, for instance, that a 12 year old girl with drama as her favorite genre approaches the recommender system. Only cases recommending movies appropriate for that age, nudity, and violence level are activated. Next, the CBR system uses a local similarity measure that uses a taxonomy, and the result is a list of drama movies free of unsuitable age limits, nudity, and violence.

Our integrated system is implemented with the software components Smile, jColibri, and MyCBR. The CBR development environment jColibri (from the University of Madrid) integrates the Bayesian network software Smile (from the University of Pittsburgh) and local similarity measure functions from MyCBR (developed at DFKI in Kaiserslautern).

Our small experimental study has shown that an integration of CBR and BN according to the properties of each individual method, as captured by the BN-CBR1 architecture is feasible. The system has been implemented to include all the four architectures, and the detailed study of the other three are now in process.

6 Conclusion and Further Plans

We have presented a framework for reasoning under uncertainty by combining BN and CBR, and described four architectures. So far, we have created a simple application using Smile, jColibri, and MyCBR. The BN-CBR-1 architecture can be a preprocessing step to CBR or a part of the similarity measure for uncertain information. BN-CBR-2 and CBR-BN-1 are tightly coupled architectures. Here the uncertain causal relations are present in BN and CBR. BN's strength is to reason under uncertainty with a well understood model, although not requiring a strong theory. CBR's strength is to reason under uncertainty with a model that is less understood. Based on past research, and the current state of our research, it is reasonable to claim that the combination of the strengths of BN and CBR perform better than BN and CBR on their own. However, we still have to provide experimental evidence for this.

In our ongoing and future work, our group will elaborate on how to combine BN and CBR in all the architectures. We will move from our toy domain into medical decision support in palliative care as soon as a sufficient amount of data and knowledge is available.

Acknowledgements

The reported research is funded by the TLCPC project, Norwegian Research Foundation under contract no NFR-183362.

References

1. Hamscher, W., Console, L., de Kleer, J. (eds.): Readings in model-based diagnosis. Morgan Kaufmann Publishers Inc., San Francisco (1992)
2. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
3. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs, 2nd edn. Springer, Heidelberg (2007)
4. Kolodner, J.L.: Case-based reasoning. Morgan Kaufmann, San Francisco (1993)
5. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications* 7(1), 39–59 (1994)
6. Watson, I.: Applying case-based reasoning: techniques for enterprise systems. Morgan Kaufmann Publishers Inc., San Francisco (1998)
7. Aha, D.W., Marling, C., Watson, I.D.: Case-based reasoning; a special issue on state-of-the-art. *The Knowledge Engineering Review* 20(03) (2005)
8. Aamodt, A., Langseth, H.: Integrating Bayesian Networks into Knowledge-Intensive CBR. In: *AAAI Workshop on Case-Based Reasoning Integrations* (1998)
9. Hjernstad, M., Fainsinger, R., Kaasa, S., et al.: Assessment and classification of cancer pain. *Current Opinion in Supportive and Palliative Care* 3(1), 24 (2009)
10. Porter, B.: Similarity Assessment: computation vs. representation. In: *Procs. of DARPA CBR Workshop*, p. 82. Morgan Kaufmann Publishers, San Francisco (1989)
11. Patel, V., Arocha, J., Zhang, J.: Thinking and reasoning in medicine (2004)
12. Schmidt, R., Montani, S., Bellazzi, R., Portinale, L., Gierl, L.: Cased-based reasoning for medical knowledge-based systems. *International Journal of Medical Informatics* 64(2-3), 355–367 (2001)
13. Lindgaard, G., Pyper, C., Frize, M., Walker, R.: Does Bayes have it? Decision Support Systems in diagnostic medicine. *International Journal of Industrial Ergonomics* 39(3), 524–532 (2009)
14. Pearl, J.: *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge (2000)
15. Lacave, C., Díez, F.: A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* 17(02), 107–127 (2003)
16. Aamodt, A.: Explanation-driven Case-Based Reasoning. *Topics in case-based reasoning*, 274–288 (1994)
17. Tran, H., Schönwälder, J.: Fault Resolution in Case-Based Reasoning. In: Ho, T.-B., Zhou, Z.-H. (eds.) *PRICAI 2008. LNCS (LNAI)*, vol. 5351, p. 429. Springer, Heidelberg (2008)
18. Gomes, P.: Software design retrieval using Bayesian Networks and WordNet. *LNCS*, pp. 184–197. Springer, Heidelberg (2004)
19. Hennessy, D., Buchanan, B., Rosenberg, J.: Bayesian Case Reconstruction. *Lecture notes in computer science*, pp. 148–158. Springer, Heidelberg (2002)
20. Pavón, R., Díaz, F., Laza, R., Luzón, V.: Automatic parameter tuning with a Bayesian case-based reasoning system. A case of study. *Expert Systems With Applications* 36(2P2), 3407–3420 (2009)

Event Extraction for Legal Case Building and Reasoning

Nikolaos Lagos, Frederique Segond, Stefania Castellani, and Jacki O'Neill

Xerox Research Centre Europe 6, chemin de Maupertuis, 38240 Meylan, France
{Nikolaos.Lagos, Frederique.Segond, Stefania.Castellani,
Jacki.Oneill}@xrce.xerox.com

Abstract. We are interested in developing tools to support the activities of lawyers in corporate litigation. In current applications, information such as characters that have played a significant role in a case, events in which they have participated, people they have been in contact, etc., have to be manually identified. There is little in the way of support to help them identify the relevant information in the first place. In this paper, we describe an approach to semi-automatically extracting such information from the collection of documents the lawyers are searching. Our approach is based on Natural Language Processing techniques and it enables the use of entity related information corresponding to the relations among the key players of a case, extracted in the form of events.

Keywords: Legal case building and reasoning, natural language processing, information extraction, e-discovery, event extraction, knowledge-based event analysis.

1 Introduction

We are interested in developing tools to support the activities of lawyers in corporate litigation, that is, the process of bringing and pursuing lawsuits. Typically corporate litigation involves the processing and analysis of large volumes of documents with the aim of finding evidence for or against the litigation claims. Identifying the important information is time consuming and costly and in recent years there has been a move to bring into play language processing technologies. Litigation involves a number of stages with different support requirements, from preservation and collection of all documents with possible relevance to the case; to review – a filtering process reducing the document set to those (non-confidential documents) answering specific criteria of relevance; and case construction – where arguments around facts and evidence are put together for presentation in court. The primary goal of the searching and browsing facilities offered in current litigation tools is to find relevant documents - often based on keyword/boolean based search techniques. Although this has proved to be relatively useful in the first stages of litigation, e.g. to identify responsive documents, during case construction the emphasis shifts from finding documents to finding entities and actionable information ([1], [2]) derived from these entities.

We are developing a system to help lawyers process the large document collections associated with a legal case and extract information from them to build the case. The idea is to provide some forms of semi-automatic support for the lawyers working to identify, for example, characters that have played a role in a case, events they have participated in, who they have been in contact with, etc. This kind of search is an important part of the work and tools currently on the market allow the users to store information on relevant characters and events. However, in current applications as users identify relevant information they must manually enter it in the tools database. There is little in the way of support to help them identify the relevant information in the first place. In this paper we describe how information might be semi-automatically extracted from the collection of documents the lawyers are searching. Our approach is based on Natural Language Processing (NLP) techniques and it enables the use of entity related information corresponding to the relations among the key players of a case, extracted in the form of events. Events are viewed as temporally bounded objects that have entities important within the application domain (e.g. persons and organisations) as participants. We chose a semi-automatic approach because case building requires deep semantic understanding of the events described in documents, thus people are integral to the process, but we also believe that information analysis can valuably support their work.

2 Legal Case Building and Reasoning

The e-discovery phase of litigation (preservation, collection and review) has long been a focus point for applying search ([3], [4]) and other technologies (e.g. [5]) in an attempt to help the lawyers manage the enormous document sets. The cost of this area has made it a particular focus for the implementation of technology as even small improvements can produce major savings. However, we believe that linguistic technologies could also benefit later stages of review, making it is easier to find information and construct cases [6]. One way of supporting such activities is to develop effective search mechanisms that aid in the discovery of relevant characters and events.

Even after responsiveness review large volumes of documents often remain, few of which actually contain information important to the case. Thus lawyers have to peruse many, often deadly dull, documents in search of anything that might stand as evidence. In addition, any single document often only contains part of the information which may be used to construct a fact – it is often the contents of a set of documents that constitutes evidence for any fact. Since the document set is large, it is usually divided between groups of paralegals and lawyers – either passing through iterations of relevance or being equally divided amongst the case team. In either case a potential problem is that something is not seen as relevant when first uncovered because the information which will make it relevant has not yet been found (or has been found by someone else). On finding partial information lawyers must currently choose whether to follow that line of enquiry themselves (when they may not have the relevant documents) or hope someone else turns up and notices the additional information [7]. Currently this distribution of information is managed through duplication and extra

effort. Thus technologies which could help lawyers better find, explore and manage the information within the entire document set could be useful.

As an example of information searching in this context, let's consider a scenario inspired by the TREC Legal Track collection [8]. One of the issues explores the proposition that cigarettes were sold to people from defendant D, with the misinterpretation, promoted by D that they were not doing any harm while D knew that they were. A sub-issue relates to proving that D denied fraudulently that nicotine is addictive. A tool that would help the user to construct the case around the above issue should support the search for information such as:

- what tests were carried out about the addictiveness of nicotine? who conducted them? when? with what results?
- when were the test results published? who saw them?
- what meetings did key people attend after the tests? who else participated in them?
- what publicity did the company release after the production of the tests?

The following sections illustrate our approach towards the extraction of information for answering these kinds of questions.

3 Event-Based Information Model for Litigation

People and organisations are typical examples of characters that may have a role in a legal case. However, depending on the litigation domain, other kinds of characters may need to be extracted. For example, in the tobacco case the following are also important: chemical substances, e.g. nicotine; products, e.g. cigarettes; and monetary expressions. The role of the characters in a case is determined, among other factors, by the events in which they participate. For instance, the role of an executive officer (EO) who publicly states something relevant to the subject of a case is more central to the case than that of other EO not involved. Naturally that is a two way relationship. The events that a key character participates in may be important for the case, but also the participants of a key event may be key characters. One of the core requirements is therefore identifying other factors (in addition to the participants) that make an event important. These include:

- the topic of an event, if any – for instance, in our example identifying that a person stated something about nicotine;
- the role of a character in the event – this enables, for example, the cases where an EO states something to be distinguished from the ones where he/she is considered in the topic of a statement;
- the relative time of an event in the chronology of the case – e.g. has the EO made a statement after contradicting results were communicated to him;
- the location that the event took place – for example did tests on tobacco take place in the company's laboratories indicating knowledge of the results from the company itself?

Events are extracted from the collection of documents associated to a legal case. They may describe situations, e.g. meetings, actions, e.g. studying, or even statuses, e.g. belong to. The events identified will depend on the matters and the domain that the legal case covers. For example, in our scenario events related to nicotine will need to be extracted (c.f. section 4.2).

Additionally, we have identified a number of classes of relations among people and organisations that we believe to be of interest to lawyers, during case construction, independently from the litigation domain. Those classes correspond to events or event abstractions and include the following:

- Role-based events such as “is employed by” (i.e. employee-employer relation).
- Interaction-based events, such as “meets”, which corresponds to the act of an entity interacting with another entity (i.e. person or of type organisation).
- Reference events such as “says”, correspond to the act of an entity referring to another entity through a written or spoken message.
- Cognitive events such as “knows” which indicate possible knowledge of a topic or entity. For example the publication of a study or writing of an email indicates the authors’ knowledge of the contents.

4 Knowledge-Based System for Event Extraction and Analysis

In order to manipulate the information described above, a system that combines event extraction, integration, and inference is required. The architecture of such a system is illustrated in Figure 1.

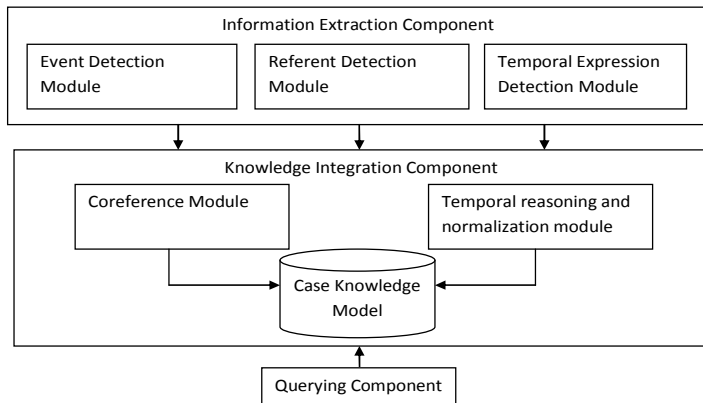


Fig. 1. Overall architecture of the knowledge-based system

The main components are: an Information Extraction Component, used to extract events, named entities along with their attributes, and temporal expressions; a Knowledge Integration Component that integrates the data extracted and infers additional

information; and a Querying Component enabling user interaction. The remaining of this section details the first two components.

4.1 Information Extraction Component

The Information Extraction Component is based on the Xerox Incremental Parser (XIP) [9]. XIP combines the following five linguistic processing layers: preprocessing (tokenization, morphological analyzer and part of speech tagging); named entity extraction; chunking; extraction of dependencies between words on the basis of sub-tree patterns over chunks sequences, and a combination of those dependencies with boolean operators to generate new dependencies or to modify or delete existing dependencies. XIP is altogether robust - able to process large collections of any documents' type, able to extract different types of relations between words and groups of words - and efficient (about 2000 words per second on an average PC).

XIP's event recognition module is also used, including a named entity extraction sub-module that detects and "semantically" categorizes proper nouns related to events. Event detection in our system is based on the approach described in [10] where an event description is considered as "a predicate (verb, adjective and predicative noun) related to its arguments and modifiers". For example, in our scenario it is important to find studies that talk about nicotine. Thus, we should be able to identify events referring to nicotine, as shown in the example in Figure 2.

<p>"The 1983 internal study conducted by Person X allegedly concluded that nicotine was an addictive substance." ----- PERSON (Person X) OBJ-N_PRDEP_CONT-ACT_KNOW (conduct,study) COORDINATE (study,conclude) COORDINATE_ACTOR_NENTITY (conduct,Person X) COORDINATE_ACTOR_NENTITY (conclude,Person X) CORE-DRIVE_SAY (conclude,be) COORDINATE_ACTOR (be,nicotine)</p>

Fig. 2. Named Entity and Event Recognition by XIP

4.2 Knowledge Integration Component

The case knowledge model (or ontology), used to describe the data to be extracted for supporting case building and reasoning activities, has three different layers. The first layer supports the integration with the indexing tools focusing on low-level features (such as text zones); the second represents concepts useful in the investigation domain (such as people, organizations, and relations); and the last allows case specific information to be included (e.g. for the tobacco case, chemical elements and substances). Based on the previous analysis, the following information has to be explicitly represented:

- Concepts defining organizations, people, documents, locations, dates/times and events (where an event is used to define states and transitions between states);
- Temporal and spatial features of events (when and where an event occurred);
- Event classes (e.g. X talks to Y and X emails Y can be abstracted as X CONTACT Y with CONTACT being an event class, X its agent and Y its patient). Motivated by our scenarios, key events in our context include cognitive, interaction, reference, and role-based ones.

To integrate information for actors involved in the litigation case, the coreference module is used. The module is able to identify the same entity occurring in the text several times with similar naming variations, using among others a string-edit distance algorithm, and also to track pronominal coreference (e.g. he, she). Distance between a co-reference and the nearest named entity is often used for disambiguating these cases, but this is not the only existing method. In traditional coreference, information such as birth date, birth place, etc. can be used to aid in identifying a chain of objects (i.e. referents) that refer to the same entity. In litigation though, the most important features that can be used include the name, the professional title and the social network of the referent. We have used the first two features.

In order to generate a timeline, static diary-like structures anchored to specific instants are needed, as well as temporally-based relations between events. All events are defined therefore in terms of the temporal interval attached to them. All intervals are identified based on some ordered pair of instants and a set of relations to other intervals. Conversely, any given interval or chain of intervals allows us to identify two instants, corresponding to the interval's or chain of intervals' beginning and end time points. As in [11] we define these time points as unique and stand for corresponding points on the timeline. A temporal graph is computed within each document using Allen's temporal relations with a forward reasoning engine [12] while all instants are normalised by changing their granularity to that of a day (that granularity was selected as the most suitable according to our objectives and the data being available). Posing queries over temporally related sequences of events thus becomes possible, bearing in mind that the temporal information may be over or under-specified.

4.3 Example

In our scenario the users of such a system may want to search for chief executive officers of Company C that said something about smoking after the results of nicotine related tests were released. Figure 3 demonstrates the extraction and integration process.

The first sentence with the coreference module enables to collect the facts indicating that a CEO of Company C has stated something about smoking while the temporal reasoning module enables the discovery of the fact that the statement was done after Company C ordered Person X to withdraw a relevant paper (Person X as a chief scientist involved in the nicotine addiction tests, is assumed to be one of the case's key people).

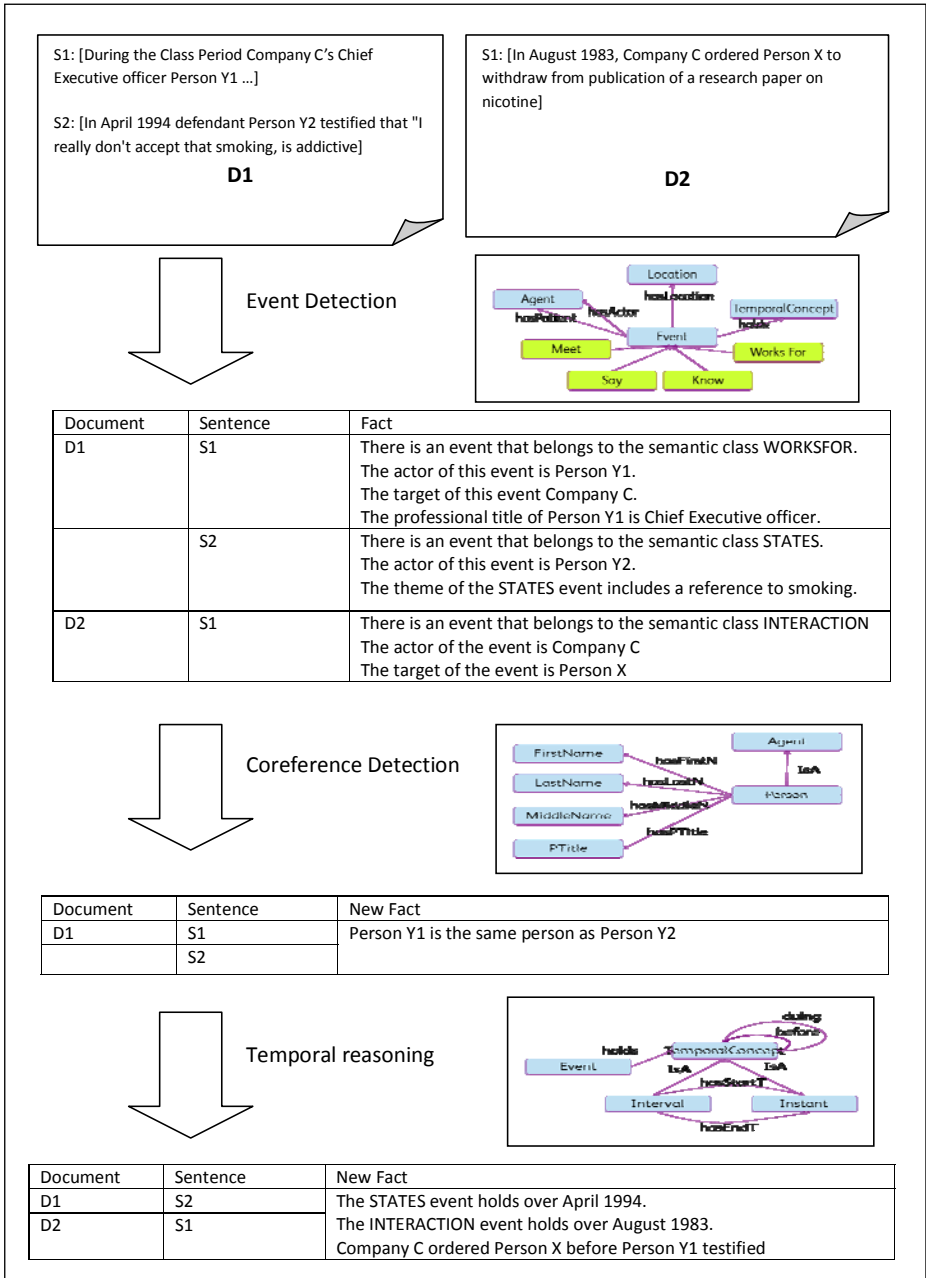


Fig. 3. Example of extraction and integration process

5 Related Work

Our work is closely related to ongoing research in the NLP community on extraction of events, temporal information, entities, and links among them. In particular, there are several evaluation campaigns, e.g. the TREC Entity Track [13], the Automatic Content Extraction (ACE) [14], on event extraction, the TempEval task evaluation, focusing on the discovery of temporal event chains [15], and SemEval 2010, on the identification of event participants [16]. This research forms a complementary and essential part of our work, since the accuracy of events extraction is an important factor to the acceptance of our system.

Other areas such as Knowledge Representation have mainly focused on the modeling and inference aspects. More particularly for legal data a number of models that explicitly represent events as first-class objects have been developed including the DOLCE-CLO [17], LRI-Core [18], and LKIF [19]. All those models focus on legal texts and deontic knowledge, while some of them are based on fundamental ontologies that do not recommend event subclasses specialized to the litigation domain, as in our case.

There is also work that integrates the two fields described above to achieve efficient legal document retrieval. [20] has applied event extraction for legal case retrieval. An event in that case is defined as any eventuality (event, state and attribute) related to illocutionary expressions existing in legal texts and therefore there has a different focus to ours. [21] also use an NLP system to define the rhetorical structure of legal texts and identify the parts in which the illocutionary expressions are present. [22] use Wikipedia to enrich the knowledge about entities and their relations for improving document retrieval in e-discovery.

A source of inspiration for all those works, including ours, is the TREC Legal Track [8] that, however, focuses on document retrieval rather than fine grained information extraction.

6 Conclusions and Future Work

In this paper we describe an NLP-based semi-automatic approach and system to support litigation case construction. We show that role, interaction, reference, and cognitive events –an event being a temporally bounded object having key entities as participants- can be used to represent relations among key characters of a case. We claim and demonstrate with an example that information integration from disparate events requires coreference resolution to identify objects describing the same entity and temporal reasoning to temporally qualify events and infer new information based on their relative chronological order. As event and participants' extraction is an integral part of our work we plan to continue improving our extraction accuracy and extending it to cross-document event extraction and tracking. In addition, we are interested in the development of user interaction components that will facilitate friendly navigation of the extracted and inferred information, as well as integration with other components of a legal case management system.

References

1. Noel, L., Azemard, G.: From Semantic Web Data to Inform-Action: a Means to an End. In: ACM Computer Human Interaction, Florence, Italy (2008)
2. Sheth, A., Arpinar, B., Kashyap, V.: Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships, Technical Report, LSDIS Lab, Computer Science, Univ. of Georgia, Athens GA (2002)
3. Sedona Conference WG1: The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery, Vol. 8 (2007)
4. Baron, J.R., Thompson, P.: The Search Problem Posed By Large Heterogeneous Data Sets In Litigation: Possible Future Approaches To Research. In: 11th International Conference on Artificial Intelligence and Law, pp. 141–147. ACM, New York (2007)
5. O’Neill, J., Privault, C., Renders, J.-M., Ciriza, V., Bauduin, G.: DISCO: Intelligent Help for Document Review. In: Global E-Discovery/E-Disclosure Workshop – A Pre-Conference Workshop at the 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain (2009)
6. Benedetti, V., Castellani, S., Grasso, A., Martin, D., O’Neill, J.: Towards an Expanded Model of Litigation. In: DESI II, Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings, London, UK (2008)
7. Attfield, S., Blandford, A., De Gabrielle, S.: E-discovery Viewed as Integrated Human-Computer Sensemaking: The Challenge of ‘Frames’. In: DESI II, Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings, London, UK (2008)
8. Oard, D.W., Hedin, B., Tomlinson, S., Baron, J.R.: Overview of the TREC 2008 Legal Track. In: 17th TREC, Gaithersburg, Maryland, USA (2008)
9. Ait-Mokhtar, S., Chanod, J.P., Roux, C.: Robustness Beyond Shallowness: Incremental Deep Parsing. *J. Nat. Lang. Eng.* 8(2-3), 121–144 (2002)
10. Capet, P., Delevallade, T., Nakamura, T., Tarsitano, C., Sandor, A., Voyatzi, S.: A Risk Assessment System with Automatic Extraction of Event Types. In: IIP2008 - 5th International Conference on Intelligent Information Processing. LNCS, vol. 288, pp. 220–229. Springer, Boston (2008)
11. Fikes, R., Zhou, Q.: A Reusable Time Ontology. AAAI Technical Report WS-02-11 (2002)
12. Hagege, C., Tannier, X.: XTM: A Robust Temporal Processor. In: Gelbukh, A. (ed.) CILing 2008. LNCS, vol. 4919, pp. 231–240. Springer, Heidelberg (2008)
13. Balog, K., de Vries, A.P., Serdyukov, P., Thomas, P., Westerveld, T.: Overview of the TREC 2009 entity track. In: 18th Text REtrieval Conference (2010)
14. NIST: The ACE 2005, Evaluation Plan (2005), <http://www.itl.nist.gov/iad/mig/tests/ace/ace05/doc/ace05-evalplan.v3.pdf>
15. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: SemEval-2007 Task 15: TempEval Temporal Relation Identification. In: ACL Workshop on SemEval (2007)
16. Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., Palmer, M.: SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In: The NAACL-HLT 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW 2009), Boulder, Colorado, USA (2009)

17. Gangemi, A., Pisanelli, D.M., Steve, G.: A Formal Ontology Framework to Represent Norm Dynamics. In: Proceedings of the 2nd International Workshop on Legal Ontologies, LEGONT (2001)
18. Breuker, J., Hoekstra, R.: Core Concepts of Law: Taking Common-Sense Seriously. In: Proceedings of Formal Ontologies in Information Systems (FOIS 2004), pp. 210–221. IOS Press, Amsterdam (2004)
19. Hoekstra, R., Breuker, J., Bello, M.D., Boer, A.: The LKIF Core Ontology of Basic Legal Concepts. In: Casanovas, P., Biasiotti, M.A., Francesconi, E., Sagri, M.T. (eds.) Proceedings of the Workshop on Legal Ontologies and Artificial Intelligence Techniques (LOAIT 2007), Stanford, CA, USA (2007)
20. Maxwell, K.T., Oberlander, J., Lavrenko, V.: Evaluation of Semantic Events for Legal Case Retrieval. In: Proceedings of the WSDM 2009 Workshop on Exploiting Semantic Annotations in Information Retrieval (ESAIR), pp. 39–41. ACM, New York (2009)
21. Weber-Lee, R., Barcia, R.M., Costa, M.C., Filho, I.W., Hoeschl, H.C., Bueno, T.C., Martins, A., Pacheco, R.C.: A Large Case-Based Reasoner for Legal Cases. In: Leake, D.B., Plaza, E. (eds.) ICCBR 1997. LNCS, vol. 1266, pp. 190–199. Springer, Heidelberg (1997)
22. Ka Kan, L., Lam, W.: Enhance Legal Retrieval Applications with Automatically Induced Knowledge Base. In: 11th International Conference on Artificial Intelligence and Law, California, USA (2007)

Applications of CBR in Oil Well Drilling: A General Overview

Samad Valipour Shokouhi^{1,3}, Agnar Aamodt², and Pål Skalle¹

¹ Department of Petroleum Technology (IPT)

² Department of Computer and Information Science (IDI)
Norwegian University of Science and Technology (NTNU)

³ Verdande Technology AS: Stiklestadveien 1- Trondheim, Norway
valipour@ntnu.no, samad@verdandetechnology.com,
agnar.aamodt@idi.ntnu.no, pal.skalle@ntnu.no

Abstract. This overview of different applications of CBR in petroleum engineering is based on a survey and comparative evaluation of different successful applications of CBR. The number of papers and research groups is indicative of importance, need, and growth of CBR in different industries. Application-oriented research in the area of case based reasoning has moved mature research results into practical applications. In this paper we present the evolving story of CBR applied in petroleum engineering especially in drilling engineering. Drilling engineering contains several potential domains of interest, in which CBR can be employed successfully.

Keywords: Case-based reasoning, oil well drilling.

1 Introduction

Case-based reasoning (CBR) is defined as the branch of artificial intelligence (AI) concerned with solving problems by reuse of past experiences. Case-based reasoning (CBR) is an approach to problem solving and decision making where new problems are solved by finding one or more similar previously solved problems, called cases, and re-using them in the new problem situation [1]. CBR may be used on its own, or integrated with other reasoning modalities to provide more accurate results by compensating the shortcomings of one approach through use of the strengths of another [2].

The aim of the study reported here is to show what possible benefits CBR can provide to the oil and gas drilling industry. The number of publications on the application of CBR in drilling operations indicates that this is a potential method to reduce cost of drilling, and increase safety of the drilling operation, by using previous experiences, hidden in reports and/or known by experts.

Oil and gas are the main energy sources in many countries. To supply world oil consumption, new wells are continuously demanded. Such needs have motivated and inspired people around the world to employ artificial intelligence in drilling

operations. Oil well drilling is a complex operation. Problems frequently occur when drilling several kilometers through different geological formations. Each well may experience both similar and new problems during the drilling operation. Offshore drilling of an oil well is also an expensive operation, costing typically 250,000 US\$ per day per rig. Access to experts for the purpose of solving problem and knowledge acquisition is limited.

2 History of CBR from Academia to Industry

CBR enables utilization of specific knowledge of previously experienced, concrete problem situations. A CBR system requires a good supply of cases in its case database. The retrieval task starts with a problem description, and ends when a best matching previous case has been found. A new problem is solved by finding a similar past case, and reusing it in the new problem situation. Sometimes a modification of the solution is done to adapt the previous solution to the unsolved case. It is important to emphasize that CBR also is an approach to incremental and sustained learning; learning is the last step in a CBR cycle [1], [3]. A CBR system can also enhance its reasoning power through the explicit representation and use of generalized knowledge about a specific domain. A classical example is the CASEY system, a medical application to diagnose heart failures [4]. Later, other frameworks for building knowledge-based systems that integrate CBR with rule-based reasoning (RBR) and model-based reasoning (MBR) were introduced by other groups such as [5] and [6].

The CBR approach was initiated roughly 35 years ago, assuming the work of Schank and Abelson [7] could be considered the origins of CBR. Several academic studies were triggered, including some with ambitions of commercialization of their own applications in the future. One of the early successful and influential applications was at Lockheed, a US aerospace company [8]. Modern aircrafts contain parts made of composite materials which must be cured in large industrial autoclaves. These parts have different characteristics requiring different autoclave heating and cooling profiles. This is complicated by the fact that many parts need to, for economical reasons, be placed together in a single large autoclave, and fact that the parts interact to alter the heating and cooling characteristics of the autoclave. Operators of Lockheed's autoclaves relied upon previous successful parts layouts to inform how to layout the autoclave. They were inspired to develop CLAVIER, the system to assist autoclave operators to reuse previously successful loadings. New successful layouts provided by operators were added to a library to improve performance of CLAVIER. The system retrieved or adapted successful layouts in 90 % of the time. The results indicated that the developed system had capability to solve problems. Note that mistakes were so costly in the domain of application, that CLAVIER was successfully served to other companies[8], [9].

There is a growing trend to employ new approaches in oil well drilling to reduce operational costs by using previous experiences gained from either previously drilled wells or on wells being currently drilled. CBR has been widely applied in the drilling industry with different focuses, described next. However, it is somewhat applied in this industry in the same proven manner as in other domains, e.g. as in aerospace through CLAVIER.

Applications of CBR in the oil and gas industry are shown on the right side of Fig.1. It further demonstrates the evolution path of the projects that we studied. They vary from methodology to field evaluated phase. The last phase, field evaluated, is synonymous with systems that have been commercialized like CLAVIER.

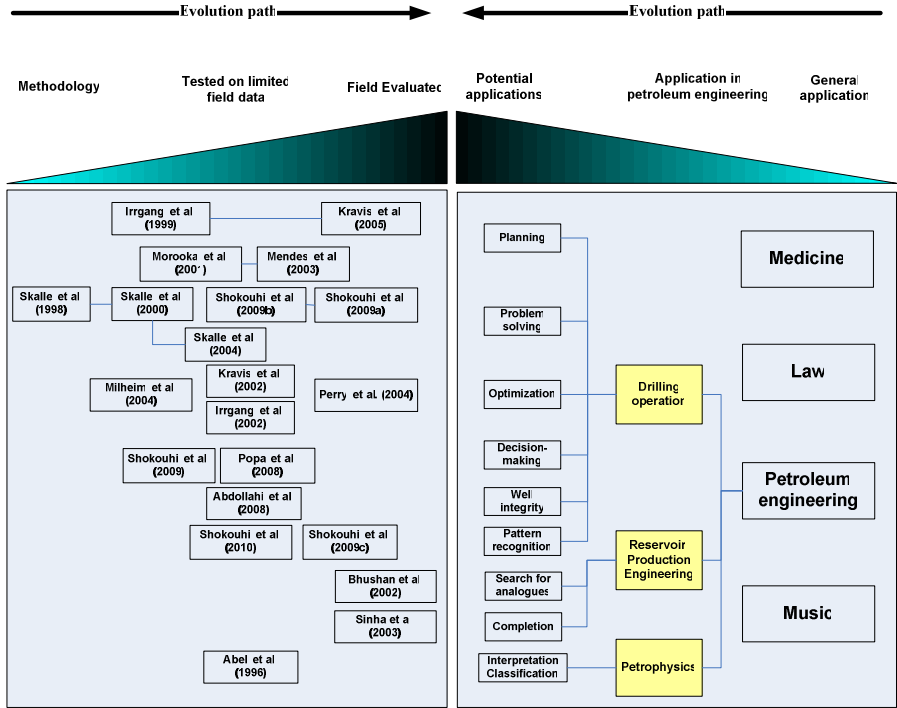


Fig. 1. Summary of CBR used in this study

In section 3 we review applications of CBR in drilling operations. Admittedly, the list may have been somewhat flavoured by the research in our group, but other well-known work has also been included to the degree their documentation has been easily available. Section 4 explains the applications of CBR in other domains of petroleum engineering. The last section summarizes and concludes on the CBR’s state of the art in petroleum engineering.

3 Applications in Drilling Operation

A good implementation of CBR will inevitably lead to facilitating solution of repetitive problems. Advanced technologies and equipments are invented and employed in the drilling industry to reduce cost of drilling operations. Drilling an offshore well may typically take one month involving high investments. Generally speaking, the

drilling industry is a technology dependent industry. Therefore, any sorts of tools or equipments that can improve the drilling operation are essential and are demanded during planning and during plan execution. Case-based reasoning has shown to provide effective support for several tasks related to drilling. Optimization of drilling plans can be achieved through CBR. The CBR method is also used for solving operational problems which require a good description of problematic situations. The information gathered via the problem analysis process may be used in order to make decision. Cases in a CBR system can be kept to predict upcoming situation through sequences.

The potential applications of CBR in drilling operations, mentioned above, were demonstrated by different groups around the world and will be presented in the following sections more in detail.

3.1 Planning

Planning a well in an optimal manner is a complex task and highly experienced engineers are needed. As well as using all the information from other disciplines, such as seismic, the experience and analysis of neighboring wells is essential for a good well plan; ‘‘good planning- few problems’’.

CBR has been tried in the oil well drilling domain by different groups around the world. One of the first applications of CBR was documented by CSIRO [10] and later refined by [11]. The technique was applied to derive alternate drilling plans based on previously drilled wells. Each well was represented as one case. A case structure has three levels. The three levels are: groups of cases, groups of attributes and groups of defined drilling phases and operations. The proposed system, Genesis, can use multiple cases at varying levels of generalization. Moreover, it uses automated tools to extract knowledge and indexes for the case base from text data.

Mendes et al. presented an application of CBR in offshore well design. The result of that work was a formalization of the methodology for planning of an oil well in a case-based reasoning context. They used fuzzy set theory for the indexing and matching of index features [12].

Mendes and his colleagues implemented a genetic algorithm to determine the proper trajectory and all the pertinent information for drilling. The initial population is the retrieved cases via CBR. It should be noted that the proposed well trajectories have to be designed by other well-known simulators starting from the well created by the genetic algorithm [13]. The current system is a continuation of past work [14], which was moved from research into a real world experimental setting.

3.2 Operation Problem Solving

Problems frequently occur when drilling several kilometers through different geological formations. The sensors installed up at the surface and downhole in the drillstring help to run the drilling process smoothly. Data via sensors are transmitted and translated to a viewable format to be interpreted by the drilling crews. In addition to real-time data, documents such as daily drilling reports and end of well reports are reported for every single well. They are valuable resources in which the solutions for

most of the past problems are expressed. Description of situations is made through both reports, consisting of description of the problems and proposed solution, or through integration of the real-time data and reports. In most CBR approaches, an abnormal situation is considered a place to make a case. It means that whenever a new problem occurs, a case is built and stored in the case database, to be used for the reasoning routine. In addition, some non-problem cases are often also stored. In particular, it is used to store "good" cases that are somewhat similar to "bad" cases, in order to better discriminate between problem and non-problem. In brief, real episodes, previously (re)solved situations, are the core of the reasoning process, in which problematic situations are explained.

Skalle along with his colleagues [15] pioneered the employment of the CBR method in the drilling industry. The work was at the conceptual and design levels. *Stuck pipe* incidents from an operator for six years were statistically analyzed. Statistical analyses led them to select parameters for cases and build a knowledge model. In this paper there was not any case matching assessment and basically it was a statistical analyses with a focus on development of the knowledge model. Two years later, they implemented the system for prevention of unwanted events in the domain of offshore oil well drilling. They introduced how to make a case in oil well drilling, mostly based on static parameters. Static parameters do not change much between measurements [16]. Their focus was on lost circulation, which means that some of the drilling fluid that always fills the gap between the drillstring and the well wall gets lost into fractures in the geological formation. They built fifty cases on the basis of information from one North Sea operator. A general domain model was used to match non-identical features that were related in the model. The integrated reasoning method is referred to as knowledge-intensive CBR (KiCBR) [6]. The CREEK framework for building knowledge-based systems that integrate CBR with model-based reasoning (MBR) was described more in detail and implemented in the drilling domain by Skalle's group [17].

In 2009, Shokouhi and his colleagues utilized a newly developed version of CREEK to integrate CBR and MBR. In this work static parameters along with dynamic parameters i.e., they can change more often, were used. Hole cleaning episodes were tagged as the problematic situation in this research work. To evaluate the case matching process, cases were categorized and labeled with respect to their downtime. It showed that KiCBR improved the retrieval process and increased the accuracy more than case based reasoning alone. It also presented how to determine the most probable root causes of poor hole cleaning episodes on basis of the knowledge model. They found how integration of MBR and CBR could improve the case matching process [18]. In late 2009, two types of KiCBR were introduced and compared to other reasoning approaches such as plain CBR and plain MBR. The aim was to obtain the best approach of reasoning in terms of the accuracy of the case matching process. The semantic network for the drilling domain was created and all the entities are binary linked. KiCBR tried to expand the set of features in the input cases through the semantic network model. The study showed that the integration of different reasoning methods improved the reasoning better than plain CBR and MBR alone [19].

3.3 Optimization and Execution of Drilling Process

Optimization of the drilling process is another application of CBR in the drilling industry. A well is being drilled in an efficient way if all the information and knowledge about drilling is utilized. Drilling performance optimization requires all the related knowledge to identify and diagnose barriers to drilling performance.

Milheim and Gaebler implemented an experience transfer tool (heuristic simulation approach) in the oil well drilling domain, based on data sets of 22 actual wells [20]. The accumulated data are treated statistically and fitted to a model based on combining human thought, artificial intelligence and heuristic problem solving. The paper presents the methodology through transformation of 22 sets of well data into a heuristic data set (activated data set). The work had a great potential to be implemented into any geological domain or specific types of drilling process.

Kravis et al. developed software for assessment of overall well quality. By means of a CBR technique, previous, analogous wells or aspects of a well are selected through similarity matching, and adapted to new wells [21]. A comprehensive set of quality measures has been derived and tested on a global database containing wells from all over the world.

Perry et al. [22] describes the development of a case-based knowledge for drilling performance optimization. A system was designed to represent all pertinent information to be used. Project documents, well summary documents, and technical lesson documents are three levels of documents in the knowledgebase hierarchy. The last one, technical lesson documents contain "case" where the lessons were learned from the analysis of the particular drilling application. This knowledge-base system enables clients, e.g. engineers, to work smarter by identifying and implementing proven solutions to the drilling problems at varying phases; planning phase, implementation phase, and post well phase.

3.4 Decision-Making

Every decision-making process provides a final option and requires identifying all options beforehand. In this regard, [23] presented a specific application of CBR to determine the optimum cleaning technique for sanded/ seized failures. These failures occur when unconsolidated reservoir sands flow into a well and cause the pump to become stuck. To correct the situation they needed to decide one out of three options; *bail*, *washback* or *foam*. The job length and job costs for each method were significantly different. They presented an application of CBR for planning and execution of well interventions, i.e., production operations, in order to improve the decision-making process. In this paper a CBR and a rule-based system are integrated. Rules (IF statements) are used for adaptation of the most common solution proposed by the CBR system. A large database for reasoning assessment was built. Data from almost 5000 well interventions over a period of three years were collected and analyzed. A small subset of historical cases was taken from the database to evaluate the proposed solution with the actual results. According to the similarity assessment, 80 % of the cases were correctly assigned the successful cleaning method. The system presented by Popa was under development and has not been implemented in the field using the revise and retain steps.

Another research work was launched by [24] that showed the procedure of the case building process and determination of root causes of poor hole cleaning. Three main groups were chosen and a CBR system was used to distinguish between them. However, discrimination among these three groups is a difficult task. They presented some examples of analyses in which the system could enhance the decision-making process by retrieving cases from the correct groups.

3.5 Well Integrity

Abdollahi et al. opened a new window for the application of CBR in the petroleum engineering domain. They explained the applicability of CBR for diagnosis of well integrity problems in order to reduce the risk of uncontrolled release of formation fluids into the well through the lifecycle of a well. Well leakages, well control issues and well collisions are named as well integrity. Abdollahi's work focused on well leakages and smartly identified causes of the leakages versus well activities. Pre-defined rules were used just for determining root causes of the leakage problems. They defined three most phases in which well leakages may occur. The three phases are: installation testing, operations (production / injection), routine testing (in-flow test for BSV and ASV). A causal model is established related to well leakages. Out of 18 cases, 12 solved and 6 unsolved cases were built and used in case matching assessment. All the cases were categorized into five groups according to the main cause of leakage. It was inferred that pre-defined rules could integrate with CBR to obtain causes of well leakages [25].

3.6 Pattern Recognition

In most CBR approaches, an abnormal situation is considered to make a case. It means that whenever a new problem occurs, a case is built and stored in the case database, to be used for the reasoning routine. One of the issues for the case building routine is to determine the severity of problems. One criticism made to CBR is the subjectivity in the case definition. The objective and advantage of each case for being stored in the case database is not straightforward task. Moreover, building cases is a time consuming process. To reduce this factor, the methodology of a semi-automatic case building and case discrimination process to make a robust case-based reasoning system was introduced and implemented [26]. All cases regardless of their severity of problems are captured. It means that the case database contains diverse cases from high to low risk. It helps to diminish subjectiveness of case building process. Past cases can be retrieved and evaluated sequentially. As the number of cases increases it is necessary to prioritize which cases should be enter into the case base immediately and which should be stored for later inclusion or discard. Shokouhi et al. [27] presented an intelligent system for prediction through sequences. As most problems during drilling operation are depth dependant, the system keeps all the cases and experiences in each defined depth interval to compose sequences of cases. Each sequence is composed of previous, present and next case. The work demonstrated that minor problems might turn into the high risk problems later on. The prediction was done and the methodology showed its ability through the good results which were obtained.

4 Applications in Other Domains of Petroleum Engineering

Over the last few years CBR has also been applied in other domains of petroleum engineering. The paper closes with a summary of related work in reservoir engineering, production engineering and petrophysics.

4.1 Applications in Reservoir Engineering

A standard database search engine returns the results whenever the search criteria meet exactly the matches. A CBR system determines the similarity to search for analogues on basis of matching attributes that are not exactly similar. Reservoirs characteristics are not exactly matched. In 2002, [28] applied CBR to globally search for reservoir analogues as an important step in planning new fields. A knowledge sharing tool was developed, called the *Smart Reservoir Prospector* (SRP). The results are accessed in a web-based system. It allows users in any Shell operating unit to access the detailed information in milli-seconds. The similarity between reservoirs computes through a set of attributes. Moreover, using reservoir analogues can provide benefits at all stages of the hydrocarbon exploration and production lifecycle, such as benchmarking scope, sharing knowledge, understanding uncertainties, finding peers, making decision, and applying lessons learned.

4.2 Applications in Petrophysics

A CBR system coupled with a database system was developed to support the interpretation and classification of new rock samples [5]. To provide petrographic analyses, the system achieves its reasoning power through the set of previous cases combined with some other source of knowledge about a certain domain. Information to build cases was provided through optical and electronic microscope analysis, isotopic and chemical analysis and petrophysics. The system was applied in one type of reservoir rocks i.e., sandstone. An interesting extension of this work would be to interpret other kinds of reservoir rocks.

4.3 Applications in Well Completion

A CBR framework was developed in Schlumberger to assess the applicability of seven lift methods for different drilling operations such as land and platform [29]. It works through decoupling the well design into a high-level or conceptual design phase and allowing for interactions between phases. Similar tools were developed for assessment of other completion methods as well.

5 Summary and Conclusion

In brief, CBR is a recent methodology compared to other computer science branches. Through this review work, it has been pointed out that the CBR methodology in the oil and gas industry has been employed by several groups all around the world. In this paper we present the evolving story of CBR applied in petroleum engineering, especially in the oil well drilling domain. The focus of this paper is to present and evaluate a number of different research efforts that employ the CBR in an attempt to improve the drilling operations.

The main point deduced from the above applications is that access to the data and information is a major problem in this business. This general overview may help leaders and manager to be more positive about the CBR technique and give “limit-less” access to the data and information. A list of potential topics of case based reasoning is covered for employing in research-oriented or industrial-oriented groups.

The study indicates that the integration of different reasoning methods improves the reasoning and the retrieval process substantially.

Acknowledgment

The authors would like to express appreciation to people from NTNU and Verdande Technology AS, for their help and cooperation in this work.

References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: Fundamental Issues, Methodological Variations, and System Approaches. *Artificial Intelligence Communications* 7(1), 39–59 (1994)
2. Marling, C., Rissland, E., Aamodt, A.: Integrations with case-based reasoning. *Knowledge Engineering Review* 20(03), 241–245 (2005)
3. Kolodner, J.: *Case-Based Reasoning*. Morgan Kaufmann Publishers Inc., San Francisco (1993)
4. Koton, P.: Reasoning about evidence in causal explanations. In: *Proceedings of the Seventh National Conference on Artificial Intelligence (AAAI 1988)*, pp. 256–261. AAAI Press, Menlo Park (1988)
5. Abel, M., Reategui, E.B., Castilho, J.M.V.: Using Case-Based Reasoning In A System That Supports Petrographic Analysis. In: *Artificial Intelligence in the Petroleum Industry: Symbolic and Computational Applications II*, ch. 7 (1996)
6. Aamodt, A.: Knowledge-Intensive Case-Based Reasoning in CREEK. In: Funk, P., González Calero, P.A. (eds.) *ECCBR 2004*. LNCS (LNAI), vol. 3155, pp. 1–15. Springer, Heidelberg (2004)
7. Schank, R.C., Abelson, R.P.: *Scripts, Plans, Goals and Understanding*. Erlbaum, Hillsdale (1977)
8. Mark, W.S.: Case-Based Reasoning for Autoclave Management. In: *Proceedings of the Case-Based Reasoning Workshop* (1989)
9. Watson, I., Marir, F.: Case-Based Reasoning: A Review. *The Knowledge Engineering Review* 9(4), 355–381 (1994)
10. Irrgang, R., Damski, C., Kravis, S., Maidla, E., Millheim, K.: A Case-Based System to Cut Drilling Costs. In: *SPE 56504*, Presented at the SPE Annual Technical Conference and Exhibition Held in Houston, Texas (1999)
11. Kravis, S., Irrgang, R.: A Case Based System for Oil and Gas Well Design with Risk Assessment. *Applied Intelligence* 23(1), 39–53 (2005)
12. Mendes, J.R.P., Guilherme, I.R., Morooka, C.K.: Case-based system: indexing and retrieval with fuzzy hypercube. In: *Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, Vancouver, vol. 7 (2001)
13. Mendes, J.R.P., Morooka, C.K., Guilherme, I.R.: Case-based reasoning in offshore well design. *Journal of Petroleum Science and Engineering* 40, 47–60 (2003)

14. Morooka, C.K., Guilhermeh, I.R., Mendesa, J.R.P.: Development of intelligent systems for well drilling and petroleum production. *Journal of Petroleum Science and Engineering* 32(2-4), 191–199 (2001)
15. Skalle, P., Aamodt, A., Sveen, J.: Case-Based Reasoning, a method for gaining experience and giving advice on how to avoid and how to free stuck drill strings. In: *Proceedings of IADC Middle East Drilling Conference, Dubai (November 1998)*
16. Skalle, P., Sveen, J., Aamodt, A.: Improved Efficiency of Oil Well Drilling through Case-Based Reasoning. In: Mizoguchi, R., Slaney, J.K. (eds.) *PRICAI 2000. LNCS, vol. 1886. Springer, Heidelberg (2000)*
17. Skalle, P., Aamodt, A.: Knowledge-based Decision Support in Oil Well Drilling. In: *Proceedings of the ICIIP, International Conference on Intelligent Information Systems, Beijing, October 21-23 (2004)*
18. Shokouhi, S.V., Aamodt, A., Skalle, P., Sørmo, F.: Determining Root Causes of Drilling Problems by Combining Cases and General Knowledge. In: McGinty, L., Wilson, D.C. (eds.) *ICCBR 2009. LNCS, vol. 5650, pp. 509–523. Springer, Heidelberg (2009); ISSN-0302-9743*
19. Shokouhi, S.V., Aamodt, A., Skalle, P., Sørmo, F.: Comparing Two Types of Knowledge-Intensive CBR for Optimized Oil Well Drilling. In: *Proceedings of the 4th Indian International Conference on Artificial Intelligence (IICAI 2009), Tumkur, India, December 16-18, pp. 722–737 (2009)*
20. Millheim, K.K., Gaebler, T.: Virtual Experience Simulation for Drilling - The Concept. In: *52803-MS, SPE/IADC Drilling Conference, Amsterdam, March 9-11 (1999)*
21. Kravis, S., Irrgang, R., Phatak, A., Martins, A., Nakagawa, E.: Drilling Parameter Selection for Well Quality Enhancement in Deepwater Environments. In: *SPE 77358-MS, SPE Annual Technical Conference and Exhibition, San Antonio, September 29 - October 2 (2002)*
22. Perry, P.B., Curry, D.A., Kerridge, J.D., Lawton, J., Bowden, D., Flett, A.N.: A Case Based Knowledge Repository for Drilling Optimization. In: *IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition, Malaysia, September 13-15 (2004)*
23. Popa, A., Popa, C., Malamma, M., Hicks, J.: Case-Based Reasoning Approach for Well Failure Diagnostics and Planning. In: *SPE 114229-MS, SPE Western Regional and Pacific Section AAPG Joint Meeting, Bakersfield, California, USA, March 29 - April 2 (2008)*
24. Shokouhi, S.V., Skalle, P.: Enhancing Decision Making in Critical Drilling Operations. Paper SPE 120290, Prepared for Presentation at the SPE Middle East Oil & Gas Show and Conference Held in the Bahrain, March 15-18 (2009)
25. Abdollahi, J., Carlsen, I.M., Randhol, P., Tenold, E., Haga, H.B., Jakobsen, T.: A Case-Based Approach to Understand the Complexity of Causal Connections Related to Well Integrity Problems. In: *IADC/SPE 111129-MS, Presented at the IADC/SPE Drilling Conference, Orlando, Florida, USA, March 4-6 (2008)*
26. Shokouhi, S.V., Aamodt, A., Skalle, P.: A Semi-Automatic Method for Case Acquisition in CBR, A Study in Oil Well Drilling. In: *Proceedings of the Tenth IASTED International Conference on Artificial Intelligence and Applications, AIA-2010, Innsbruck, Austria, February 15-17, pp. 263–270. ACTA Press (2010)*
27. Shokouhi, S.V., Aamodt, A., Skalle, P., Sørmo, F.: Integration of Real-Time Data and Past Experiences for Reducing Operational Problems. In: *IPTC 13969, Proceedings of the International Petroleum Technology Conference held in Doha, Qatar, December 7-9 (2009)*
28. Bhushan, V., Hopkinson, S.C.: A Novel Approach to Identify Reservoir Analogues. In: *European Petroleum Conference, Aberdeen, United Kingdom, October 29-31 (2002)*
29. Sinha, S., Yan, M., Jalali, Y.: A Methodology for Integrated Well Planning. In: *SPE 85315-MS, SPE/IADC Middle East Drilling Technology Conference and Exhibition, Abu Dhabi, United Arab Emirates, October 20-22 (2003)*

Associated Clustering and Classification Method for Electric Power Load Forecasting

Quansheng Dou^{1,2}, Kailei Fu², Haiyan Zhu², Ping Jiang², and Zhongzhi Shi¹

¹ Key Laboratory of Intelligent Information Processing; Institute of Computing Technology;
Chinese Academy of Sciences; Beijing; 100080

² School of Computer Science and Technology; Shandong Institute of Business and Technology,
Yantai 26400
douqs@ics.ict.ac.cn

Abstract. In the process of power load forecasting, electricity experts always divide the forecasting situation into several categories, and the same category uses the same forecasting model. There exists such a situation that some load curve which domain experts consider belonging to the same category has shown the different characteristics, but some load curve which belongs to different category seems very similar, and usually able to gather into a category by clustering. For this problem, the definition of associated matrix was proposed in this paper, and based on this conception the associated clustering-classification algorithm was proposed, We applied this algorithm to data sample classification for power load prediction, the experiment showed that the classification results obtained by our method were more reliable.

Keywords: Electricity Load Prediction; Classification; Wavelet Anlysis.

1 Introduction

Load forecasting is an important component for power system energy management system. Precise load forecasting helps the electric utility to make unit commitment decisions, reduce spinning reserve capacity and schedule device maintenance plan properly. Besides playing a key role in reducing the generation cost, it is also essential to the reliability of power systems. The system operators use the load forecasting result as a basis of off-line network analysis to determine if the system is vulnerable. If so, corrective actions should be prepared, such as load shedding, power purchases and bringing peaking units on line.

Classification and clustering are two important research areas of data mining. To map data into some given classes, classification depends on prior knowledge, and clustering is to make samples in the same cluster similar enough, while samples belonging to different clusters should have enough difference. Recently, [1]~[2] use granularity computation to solve classification problems, and with the improvement of granularity computation theory these methods will develop further. [3]~[4] use ant colony optimization etc. to search the classification rules, these algorithms are

the combination of data mining and intelligence computation. [5] proposes a new classification algorithm based on the combination of supported vector machine and non-supervisor clustering, and gets better results when it is applied in web page classification. [6] systematically summarizes the clustering method. All these researches represent the newest development in this area.

In the process of power load forecasting, electricity experts always divide the forecasting situation into several categories, the same category uses the same forecasting model. There exists such a situation that some load curve which domain experts consider belonging to the same category has shown the different characteristics, but some load curve which belongs to different category seems very similar, and usually able to gather into a category by clustering. In other words, the prior knowledge is very likely uncoordinated with similarity measure. [7] analyzed this issue by granularity theory and put forward classification algorithm based on information granularity principle. This has strong theoretical and practical significance. Aimed at the above problem, this paper proposes the Associate Clustering-Classification Algorithm to ensure the consistency of classification and clustering. The algorithm in sample classification of power system load forecasting is applied, and better results are obtained. The detail of the Associate clustering-classification method will be described in the following.

2 Associated Clustering and Classification Method

Let $U = \{x_1, x_2, \dots, x_k\}$ be a sample set, δ be a cluster operation, and it forms into a Cluster Genealogy G under action of δ . We cut Cluster Genealogy G , divide U into independent subset, and get $\delta(U) = \{G_1, \dots, G_m\}$. We classify U with Prior knowledge and obtain classification $C = \{C_1, \dots, C_n\}$. For $\forall C_i \in C \ i = 1, 2, \dots, n$

is divided into m sub-sets by $\delta(U)$, we have $C_i = \bigcup_{j=1}^m C_{ij}$ where m is the number

of sub-sets in $\delta(U)$, and $C_{ij} = C_i \cap G_j$.

Definition 1. Suppose U is a sample space, δ is a cluster operation, $G = \delta(U) = \{G_1, \dots, G_m\}$ and $C = \{C_1, \dots, C_n\}$ is classification of U . We call

matrix $\Lambda = \{\lambda_{ij}\}_{n \times m}$ an Associated Matrix of classification C based on G . Here,

$\lambda_{ij} = \frac{|C_{ij}|}{|C_i|}$ and $C_{ij} = C_i \cap G_j$, $|C_{ij}|$ and $|C_i|$ represent the number of elements in the collection C_{ij} and C_i . We call each row R_i $i = 1, \dots, n$ of matrix Λ an Associated Vector.

Definition 2. On the basis of definition 1, let $R_i = (r_{i1}, r_{i2}, \dots, r_{im})$ $i = 1, \dots, n$, where R_i is the Associated Vector. If there are s components which satisfy $r_{ij} \geq \frac{1}{m}$, $j = 1, 2, \dots, s$, then vector R_i is called s items dominant or multi-term dominant, and if vector R_i has only one r_{ij} which satisfies $r_{ij} > \frac{1}{m}$, then R_i is called one term dominant. If $r_{ij} > 0.8$ and makes R_i one term dominant, then R_i is called one term sufficiently dominant.

Algorithm 1. Associated Clustering and Classification Algorithm

Step 1. Classify C according to prior-knowledge, and get the initial classification $C = \{C_1, C_2, \dots, C_n\}$.

Step 2. Implement cluster operation according to Euclidean distance on U and get cluster genealogy G . Cut from top of G , and get two branches, each of which forms one class. Get the Associated Matrix:

$$\Lambda = \begin{pmatrix} \lambda_{1G1} & \lambda_{1G2} \\ \dots & \dots \\ \lambda_{nG1} & \lambda_{nG2} \end{pmatrix} \tag{1}$$

Step 3. At some time, suppose Associated Matrix

$$\Lambda = \begin{pmatrix} \lambda_{1G1} & \lambda_{1G2} \dots & \dots \lambda_{1G1} \\ \dots & & \dots \\ \lambda_{nG1} & \lambda_{nG2} \dots & \dots \lambda_{nG1} \end{pmatrix} \tag{2}$$

Inspect each column $(\lambda_{1G_j}, \lambda_{2G_j}, \dots, \lambda_{nG_j})^T$ of Λ , If there are two or more components λ larger than μ , and $|G_j| > \tau |U|$, where $0 < \tau < \mu < 1$ is the threshold parameter, $|G_j|$ and $|U|$ represent the number of elements of G_j and U respectively. Cutting at the top of the cluster genealogy G_j , form the new branch, and revise the Associated Matrix. Implement step3 repeatedly until there is only one $\lambda_{iG_k} > \mu$ in each row of Λ or $|G_j| < \tau |U|$.

Step 4. For each column $(\lambda_{1G_j}, \lambda_{2G_j}, \dots, \lambda_{nG_j})^T$ of the matrix Λ , if there are two or more components λ sufficiently dominant in their rows, set these components be $\lambda_{hG_j}, \lambda_{h+1G_j}, \dots, \lambda_{pG_j}$, then combine $\lambda_{hG_j}, \lambda_{h+1G_j}, \dots, \lambda_{pG_j}$ into one category. Revise matrix Λ .

Step 5. Analyze each row $Row_i = (\lambda_{iG_1}, \lambda_{iG_2}, \dots, \lambda_{iG_m})$ of matrix Λ . If Row_i is one term sufficiently dominant, then take C_i as one class individually. Otherwise, set the threshold κ , Suppose there are l components larger than κ in Row_i , and they are $\lambda_{i_s}, \lambda_{i_{s+1}}, \dots, \lambda_{i_{s+l-1}}$ respectively. C_i Will be divided into l categories based

on G_s, \dots, G_{s+l-1} . For $\forall x \in C_i - \bigcup_{r=0}^{l-1} G_{s+r}$, according to the principle of minimum

distance to the collection center, add them into some class of C_i .

Step 1 and 2 of the algorithm implement classification and clustering operation on the sample collection. The emphasis is that the prior-knowledge used by classification and the measure function used by clustering are essentially the same, otherwise, it is not worth harmonizing.

Step3 of the algorithm can ensure that there is only one classification C_i , whose most samples appear in a clustering G_j . If there are two or more classifications, whose most elements are in the same clustering called G_j , cut up G_j on the top of the clustering genealogy. Finally, if there are still two or more classifications whose most elements are in the same clustering G_j , the number of samples in G_j must be below a certain size. These classifications were combined into one class in step4.

In step5, if a row Row_i of the Associated Matrix is one term sufficiently dominant, C_i should be set as one class individually. Steps 3 and 4 have ensured that there can be no more than two categories C_i whose majority of samples appear in a clustering G_j . If Row_i is not a one term sufficiently dominant vector, and most samples of C_i distribute in G_s, \dots, G_{s+l-1} , C_i should be divided into l classes according to G_s, \dots, G_{s+l-1} . Samples out of G_s, \dots, G_{s+l-1} in C_i should be added to a certain classification in C_i with the principle of minimum distance.

By analyzing the performance process of the above algorithm, it's easy to see that when the algorithm is finished, Associated Matrix Λ has k rows which are one term sufficiently dominant and $s - k$ rows which are multi-term dominant. Here, $k \geq 0$ and $s \leq n$, n is the number of classifications obtained by priori knowledge. And no columns in Λ can make all rows sufficiently dominant. The above can be shown in the following formula (3):

$$\Lambda = \left(\begin{array}{cccc} 1 & 0 & \dots & \\ 0 & 1 & \dots & \\ \dots & \dots & \dots & \\ \lambda_{i1} & 0 & \lambda_{ij} & \dots & 0 \\ 0 & \dots & \lambda_{i+1,j} & \lambda_{i+1,m} & \dots \\ \dots & \dots & & & \end{array} \right) \begin{array}{l} k \\ k + 1 \\ s \end{array} \tag{3}$$

Two kinds of standards are involved here. One is the priori knowledge of domain experts. Because many complex factors affect the change of power load, and some reasons which cause power load changing is not clear, the priori knowledge used by experts on power, often just reflect the variation of load roughly. The other is that characteristics of load change can be objectively identified by clustering, but the reasons why the samples cluster into a class are not yet determined. This makes the clustering method can not be directly used on predicting. For this reason, the next best thing is to take a relatively compromise. Associated clustering-classification algorithm is an exactly compromise method between classification and clustering.

3 Description of the Problem of Power System Load Forecasting

Load forecasting is a traditional research field of power system [9]~[11]. In the process of power load forecasting, electricity experts always divide the forecasting situation into several categories, the same category uses the same forecasting model. So a reasonable classification is the basis for effective forecast. Generally, domain experts classify the samples relying on their experience. In this paper, 96-points data samples of a Chinese power company in recent years were classified by the expert's experience and associated clustering-classification algorithm described in the previous section. Here, the forecasting models used by the different classification methods are the same.

First of all, the samples are classified. For different categories, Daubechies wavelets are used to extract the feature of load data.

Let $\{p(t)\} \ t = 1, \dots, 96$ be the load value of 96 points one day. Let $C_0(t) = p(t)$, wavelet decomposition is shown as follows:

$$\left\{ \begin{array}{l} C_0 = p(t) \\ C_j[k] = C_{j-1} \bar{h}[2k] \\ D_j[k] = C_{j-1} \bar{g}[2k] \end{array} \right\} \quad j = 1, 2, \dots, L \quad (4)$$

In the formula $\bar{h}[-k] = h[k]$, $\bar{g}[k] = g[-k]$, $g[k] = (-1)^{k-1} h[k-1]$. $h[k]$ is the low-pass filter, $g[k]$ is the high-pass filter, and L is the decomposition level.

$C_j[k]$, $D_j[k] \ j = 1, 2, \dots, L$ are low-pass signal (features) and high-pass signal (noise) of the j -layer wavelet transform respectively. By the wavelet transform, the 96-points time series are broken down into two parts, feature and noise. The dimension of the low-pass signal C_{j+1} and high-pass signal D_{j+1} obtained in each of the decomposition is half of the dimension of C_j . Let C_0 be the 96 points load data initially, the dimensions of C_3, D_3, D_2, D_1 are 12, 12, 24, 48 after three wavelet decomposition. So the dimension of $\{C_3, D_3, D_2, D_1\}$ remains 96. Here, the previous 12 components C_3 contain the overall volatility information $\{p(t)\}$, i.e. the characteristic component while D_3, D_2 , and D_1 are high-frequency information, i.e. the noise component of $\{p(t)\}$ at different spatial scales. $\{p(t)\}$ can be obtained by reconstruction of vector $\{C_3, D_3, D_2, D_1\}$.

We can obtain the temperature information from the meteorological station and analyze the relationship between the temperature and the feature components. As the temperature changes, the feature component values show a certain discipline. We can regress this law and get the polynomial relations between the temperature and features components. So the feature components can be forecast according to the change of real sense temperature.

It is impossible to determine the relationship between temperature and noise with regression approach, because noise components show splattering distribution to the temperature. As described above, noise component is constituted by the high frequency information on different scales of space obtained by three-layer wavelet decomposition to 96 points data. Its vector length is 84. We use the following method to determine the noise components:

Let $D_i = \{d_{i1}, d_{i2}, \dots, d_{i84}\}$, $i=1,2,\dots,q$ be the noise component of 96 point

data for one day. Let $f(d) = \sum_{i=1}^q \sum_{j=1}^{84} |d - d_{ij}|$, where q is the total number of samples

for the classification. The noise component \bar{d}_j , $j=1,2,\dots,84$ can be determined by

solving the optimization problem of $\min f(d)$.

As mentioned above, we can predict the feature $\{C_3\}$ through the temperature.

Reconstruction of C_3 and $\{D_3, D_2, D_1\}$ will get the electricity load forecasting value on that day.

4 Forecasting Results of Different Classification Methods

Classify the samples by date type according to prior knowledge, denote as $C = \{C_1, C_2, \dots, C_n\}$. For each of the 96 points historical load data denoted as

$(p_0, p_1, \dots, p_{95})$, let $\Delta p = (\frac{p_1 - p_0}{p_0}, \frac{p_2 - p_1}{p_1}, \dots, \frac{p_{95} - p_{94}}{p_{94}})$. Cluster Δp by

Euclidean distance to form cluster genealogy, and reclassify the above classification with associated clustering-classification algorithm. The original classification

$\{C_1, \dots, C_n\}$ is reclassified to get the classification $C' = \{C'_1, \dots, C'_m\}$. Analyze the final associated matrix

$$\begin{pmatrix} 1 & 0 & \dots & & \\ 0 & 1 & \dots & & \\ \dots & \dots & & & \\ \hline \lambda_{i1} & 0 & \lambda_{ij} & \dots & 0 \\ 0 & \dots & \lambda_{i+1,j} & \lambda_{i+1,m} & \dots \\ \dots & \dots & & & \end{pmatrix}$$

For all the rows which are one term sufficiently dominant in associated matrix Λ . Classification results based on priori knowledge are about the same as the results of clustering, and the method used for forecasting is consistent with that mentioned above.

For all the rows as $(0, \dots, \lambda_{ij}, \dots, 0, \dots, \lambda_{im}, \dots, 0)$, which are multi-terms dominant in Associated Matrix Λ , suppose $\lambda'_{is}, \lambda'_{is+1}, \dots, \lambda'_{is+l-1}$ are dominant items. There exists classification $C'_{is}, C'_{is+1}, \dots, C'_{is+l-1}$ in the new classification set C' corresponding with them. Extract feature and noise in these classification respectively, and regress relationship between the feature and temperature, and reconstruct with the corresponding high-frequency signal to get predictor $P'_{is}, P'_{is+1}, \dots, P'_{is+l-1}$. Because the reason why C_i is classified into $C'_{is}, C'_{is+1}, \dots, C'_{is+l-1}$ is unknown, a specific forecast can only start from a priori knowledge. Forecast should be taken as follows on the corresponding situation of the row:

$$P_i(t) = \sum_{j=s}^{s+l-1} \lambda_{ij} P_{ij} \tag{5}$$

To verify the effectiveness of the method described in this paper, we use the historical data of the previous two years as the learning sample, and respectively predict the load of next year by different classification methods. In the process of load forecasting, for some special holidays such as the Chinese Spring Festival and New Year's Day, it has no sense to regress for lacking of historical data, and the forecast can only be made by historical trends.

Table 1 lists several groups of statistical results of the experiment. In table 1, statistics type A is the percentage of the data points whose errors are less than 1%. B is the percentage of points whose errors are between 1% and 3%. C is the percentage of points whose errors are larger than 3%. D is the average of root-mean-square error between predictions and the actual data.

Table 1. Predictions based on different classification. Error described as: A, B, C, D are equal to the percentage of points whose errors are less than 1%, the percentage of points whose errors are between 1% and 3% ,the percentage of points whose errors are larger than 3% and the average of error respectively, in the forecast data points.

Year	Error type	Forecasting result of classification according to the prior knowledge	Forecasting result of Associated Clustering and Classification
2007	A	68% (23827 points)	76% (26630points)
	B	17% (5957points)	14% (4906points)
	C	15% (5256points)	10% (3504points)
	D	3.02%	2.41%
2008	A	77% (27055points)	81% (28460points)
	B	15% (5270points)	13% (4568points)
	C	8% (2811points)	6% (2108points)
	D	2.16%	1.88%
The first half of 2009	A	75% (13032points)	80% (13901points)
	B	16% (2780points)	15% (2606points)
	C	9% (1564points)	5% (869points)
	D	2.25%	1.82%

From Table 1, we can see that forecasting results based on the new classification method are significantly better than the original classification based on experience. It is not difficult to see that through the above analysis, the load data often have different characteristics objectively in the classification based on priori knowledge. Considering classification whose features is different from samples as one classification to regress is the main reason for the large error of regression curve. The associated clustering-classification algorithm in this paper avoids this problem to a certain extent, and forecast accuracy has been improved significantly. It also validated that the method proposed in this paper better solved the inconsistent problem between the priori knowledge and the similarity measure function.

5 Conclusion

Classification and clustering are two important research areas of data mining; however in the process of power load forecasting, the classification results based on priori knowledge and the clustering results are not consistent. For this problem and the practical application background of power system, the definition of associated matrix has been proposed in this paper, and based on this concept, the associated clustering-classification algorithm has been proposed. We applied this algorithm to data sample classification for power load prediction, the experiment showed that the classification results obtained by our method were more reliable.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No.60970088,60775035), the National High-Tech Research and Development Plan of China (No. 2007AA01Z132), National Basic Research Priorities Programme(No. 2007CB311004) and National Science and Technology Support Plan (No.2006BAC08B06), Dean Foundation of Graduate University of Chinese Academy of Sciences(O85101JM03).

References

- [1] Yao, T., Yao, Y.Y.: Granular computing approach to machine learning. In: Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age, November 18-22, vol. 2. Orchid Country Club, Singapore (2002)
- [2] Yao, Y.Y., Yao, J.T.: Induction of classification rules by granular computing. In: Proceedings of The Third International Conference on Rough Sets and Current Trends in Computing, pp. 331–338 (2002)
- [3] Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an Ant Colony Optimization Algorithm. *IEEE Trans. on Evolutionary Computation* (2002)
- [4] Sousa, T., Silva, A., Neves, A.: A particle swarm data miner. In: Pires, F.M., Abreu, S. (eds.) *EPIA 2003. LNCS (LNAI)*, vol. 2902, pp. 43–53. Springer, Heidelberg (2003)
- [5] Li, X.l., Liu, J.m., Shi, Z.z.: A Chinese Web Page Classifier Based on Support Vector Machine and Unsupervised Clustering. *Chinese Journal of Computers* 24(1) (2001)
- [6] Xu, R.: Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks* 16(3) (May 2005)
- [7] Bu, D.b., Bai, S., Li, G.j.: Principle of Granularity in Clustering and Classification. *Chinese Journal of Computers* 25(8) (2002)
- [8] Zhang, J., Song, R., Yu, W.X., Xia, S., Hu, W.-D.: Construction of Hierarchical Classifiers Based on the Confusion Matrix and Fisher's Principle. *Journal of Software* 16(9), 1000–9825 (2005)/ 16(09) 1560
- [9] Alfares, H.K., Nazeeruddin, M.: Electric load forecasting: literature survey and classification of methods. *International Journal of Systems Science* 33(1), 23–34 (12) (2002)
- [10] Metaxiotis, K., Kagiannas, A., Askounis, D., Psarras, J.: Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. *Energy Conversion and Management* 44(9), 1525–1534 (2003)
- [11] Kang, C.q., Xia, Q., Zhang, B.m.: Review of power system load forecasting and its development. *Automation of Electric Power Systems* 28(17) (2004)

Two Improvement Strategies for PSO

Quansheng Dou^{1,2}, Shasha Liu², Ping Jiang², Xiuhua Zhou², and Zhongzhi Shi¹

¹ Key Laboratory of Intelligent Information Processing; Institute of Computing Technology;
Chinese Academy of Sciences; Beijing; 100080, China

² School of Information and Electronics Engineering; Shandong Institute of Business and
Technology, Yantai 264005, China
douqs@ics.ict.ac.cn

Abstract. This paper proposed an improved particle swarm optimization algorithm (IPSO) to solve continuous function optimization problems. Two improvement strategies named “Vector correction strategy” and “Jump out of local optimum strategy” were employed in our improved algorithm. The algorithm was tested using 25 newly proposed benchmark instances in Congress on Evolutionary Computation 2005 (CEC2005). The experimental results show that the search efficiency and the ability of jumping out from the local optimum of the IPSO have been significantly improved, and the improvement strategies are effective.

Keywords: Particle Swarm Optimization; Convergence Property; Improved Strategy.

1 Introduction

Particle Swarm Optimization (PSO) method was proposed by Kennedy and Eberhart in 1995 as an evolutionary computing technology [1] [2], which is widely used in various types of optimization problems [3]-[7]. In solving optimization problems, a potential solution of each optimization problem is seen as a "particle" in the search space. An ordered triple of numbers (x_i, v_i, p_i) corresponds to each particle; the location of particle's each iterations is decided by the following formula:

$$v_{t+1} = \omega v_t + c_1 r_1 (p_i - x_t) + c_2 r_2 (p_g - x_t)$$
$$x_{t+1} = x_t + v_{t+1}$$

Where x_i represents the current position of the particle, v_i represents the current speed of the particle, p_i is the best location of particle i (individual optimum), p_g is the best location of all the particles in the swarm have passed (swarm optimum); c_1 and c_2 are positive constants, r_1 and r_2 are random numbers which obey normal distribution in the interval $[0,1]$; ω is an inertia parameter. The advantage of PSO is that it is easy to implement and there are few parameters to adjust. PSO has been successfully applied in

many areas: function optimization, artificial neural network training, fuzzy system control, etc.. However, for some complex problem, e.g. 25 benchmark problems in the CEC2005 (The 2005 IEEE Congress on Evolutionary Computation), the searching results of PSO method are not satisfactory. In this paper, two improvement strategies were proposed in order to improve the performance of PSO, the effectiveness of the improvement strategy is proved by 25 benchmark problem provided by CEC2005[8] [9], the following will describe the relevant content.

2 Two Improvement Strategies for PSO

The reasons for the algorithm to fall into local optimum are described in the following:

- i). Because the velocity of particles in the swarm decays too fast, the step-length of particles revision decreases rapidly, which make the search efficiency too low or search stagnation;
- ii). For some complex issues with a strong deceptive, once the search fall into local optimum, even if the particle's velocity does not completely decay, the probability of that particles hitting a better one than the existing optimal point is still smaller, and thus cause the search stagnation.

To overcome the above two points, we make the following improvements on the PSO method:

- i). Vector revision strategy

In traditional PSO method, at each step t , particle's position is updated by the following program

for $i=1$ to Dims

$$v_{t+1}^{(i)} = wv_t^{(i)} + c_1r_1(p_i^{(i)} - x_t^{(i)}) + c_2r_2(p_g^{(i)} - x_t^{(i)})$$

$$x_{t+1}^{(i)} = x_t^{(i)} + v_{t+1}^{(i)}$$

end

Here, Dims is the dimensions of the search space, each components of vector is revised respectively. Yet if there is some relativity between the components of vector, the efficiency of such a revision method is low. To improve the search efficiency, this paper revises the particle as a whole vector according to a probability. For each step of the algorithm, particle's position is updated by the following program

if $Rand > \alpha$

for $i=1$ to Dims

$$v_{t+1}^{(i)} = wv_t^{(i)} + c_1r_1(p_i^{(i)} - x_t^{(i)}) + c_2r_2(p_g^{(i)} - x_t^{(i)})$$

$$x_{t+1}^{(i)} = x_t^{(i)} + v_{t+1}^{(i)}$$

end

else

$$\vec{v}_{t+1} = w\vec{v}_t + c_1r_1(\vec{p}_i - \vec{v}_t) + c_2r_2(\vec{p}_g - \vec{v}_t)$$

$$\vec{x}_{t+1} = \vec{x}_t + \vec{v}_{t+1}$$

end if

Where *Rand* is a random number between 0 and 1, α is a given threshold, such a strategy could speed up the efficiency of that individual converge to a swarm optimum. Especially when the various dimensional components of the issue are interrelated, the search efficiency improves clearly. Here, it is very important to set the threshold α . If the value of α is too small, the number of particles in the swarm revised as a vector is too large, while raising the efficiency, also increase the possibility of "premature". Otherwise, search efficiency will be affected. It will be better to set $\alpha = 0.945$ in our experiment.

(2) Jump out of local optimum strategy

In order to make particles jump out of local optimum effectively to avoid "premature" in PSO search process, this paper has developed the following strategies:

if $\text{Immovable} > \gamma$ or $\text{Sigm} < \delta$

Save global best into Bestgroup;

if $\text{Rand} > \theta$ and *Bestgroup* not Null

global best = Randomly select different particle from BestGroup

else

Initial population;

Initial global best;

end if;

end if

In the above program, *Immovable* is the iterative times of that swarm optimum p_g keep unchanged, γ , δ , θ are the specified thresholds respectively. *BestGroup* is an array which is used to save swarm optimum keeps unchanged or lower revision efficiency after *Immovable* iterations. Here *Sigm* is used to determine the revision efficiency of the swarm optimum p_g :

$$\text{sigm} = \frac{\sum_{i=1}^M \text{rate}_i}{M}$$

Where $rate$ is an M dimensional vector. Each component $rate_i, i=1, M$ was given a relatively large value initially. When algorithm finds new swarm optimum, $rate$ will be updated as follows:

$$rate(2, \dots, M) = rate(1, \dots, M-1) \text{ and } rate_1 = \frac{g_best_{i+1} - g_best_i}{S(i+1) - S(i)}$$

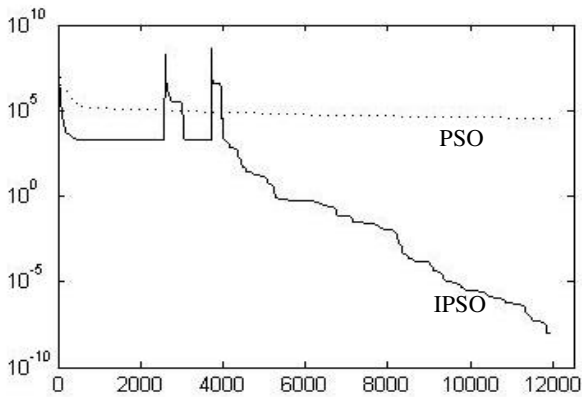
In the above equation g_best_{i+1} , g_best_i are the swarm optimums what the algorithms have searched twice in succession. $S(i+1)$, $S(i)$ are the time for getting the two swarm optimum respectively.

We can see from the above strategy that when the swarm optimum remains invariable or changes in value less than a certain threshold for a long time, the algorithm according to a probability choose a particle from the *BestGroup* as new swarm optimum or re-initialize the swarm and select a new swarm optimum from swarm. Because the particles in the *BestGroup* have been the “swarm optimum”, when it is selected again, the original distribution pattern of the swarm is broken, the velocity of particles will be compensated indirectly, and the probability of the algorithm achieving the goal is still very high. For some complex multi-peak problems, search can be achieved possibly only when the particle jump out of a local optimum. Therefore, when the swarm optimum remains invariable or changes in value less than a certain threshold value for a long time, the algorithm according to a probability re-initializes the group and generates a new optimal group to jump out of the attraction of local optimum.

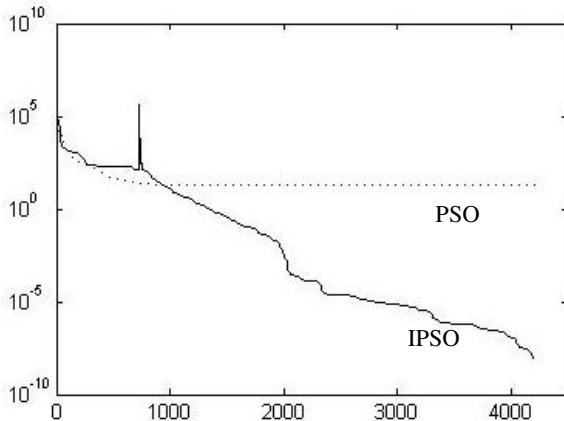
In this paper, the PSO with the above improved strategies is named Improved Particle Swarm Optimization (IPSO), in the following Fig.1a, b are examples of PSO method before and after improvement. Both f_3 (Shifted Rotated High Conditioned Elliptic Function) and f_{12} (Schwefel’s Problem 2.13) are two test functions in CEC2005. f_3 which is generated from moving and rotating Elliptic Function has a unique extreme point, f_{12} is obtained by moving the Schwefel’s Problem, and has multiple local optimal point and one global optimal point. Dotted and solid lines are the search curve of PSO and IPSO respectively in the figure. In the Fig.1a, As a result of rotation and movement, although the test function has only one optimal point, it is still very difficult to find its global optimum point. Because f_3 is rotated by the following formula:

$$\vec{x} = \vec{x} \times M$$

Where M is the specified matrix, so the components of position vector x correlate between each other. It can be seen from the Fig1.a that the implementation curve of IPSO has two "mutations", it is the results triggered by "Jump out of local optimum strategy". After two jumps, the groups eventually converge to the global optimum point, however the PSO is trapped in "premature" state and the search fails. Before the "jump out of local optimum" strategy triggers, only the "Vector correction strategy" plays a role in the IPSO. It is not difficult to see from Fig.4a that the convergence speed of IPSO is faster than PSO. Fig.4b experienced one time of jumping out, since f_{12} isn't rotated, every components of vector is independent, the role of "Vector correction strategy" does not seem great.



a: Shifted Rotated High Conditioned Elliptic Function (f_3)



b: Schwefel's Problem 2.13(f_{12})

Fig. 1. Example of IPSO jumping out of local optimum point

Due to limited space, the content of the experiment is not described in this section, the details of experiment will be described in the following sections.

3 Experiment and Test Results

We adopted 25 benchmark problems [8]proposed in CEC 2005 where $f_1 - f_6$ are single-peak functions and $f_7 - f_{25}$ are multi-peak functions, these functions are obtained through the shift, rotation and other operations based on the traditional benchmark functions. PSO method performed poorly to these 25 test problems. Literature [9] compared 11 algorithms which have been published in CEC2005.This paper used the same comparison method as the literature [9]: For all benchmark functions, the dimensions of search space was Dims=10 and Dims=30, the maximum number of iterations can be taken as $\text{Dims} \times 10^5$, each test problem was carried out 25 times respectively, in order to measure the implementation of the algorithm results, let

$$FEs = \frac{\text{mean}(\#\text{fevals}) \times \#\text{all runs}}{\#\text{successful runs}}$$

Here, $\#\text{fevals}$ is the iteration times for each successful runs, $\#\text{all runs}$ is the total implementation times of the algorithm (25), $\#\text{successful runs}$ is the number of successful runs. The smaller the value of FEs is, the higher the search performance of the algorithm. The following Table 2 lists the comparison between the best results in literature [9] and the running results with IPSO in the conditions that the dimensions of search space are 10 and 30 respectively.

For all the algorithms in literature [9], function $f_8, f_{13}, f_{14}, f_{16}-f_{25}$ were not completely searched within the allotted time. IPSO completed the search of f_{13} twice in 25 times implementation when Dims=10, and the corresponding FEs was listed in the table. For the comparable 13 functions, IPSO gets 8 best outcomes for the function in the cases when dimension is 10. When Dims=30, There are seven functions obtain the best search results, otherwise the PSO without the improvement have not completed the search except f_1, f_2, f_4 . It sufficiently proved the effectiveness of the improved strategies.

Table 2. The results of IPSO method compared with literature [9] in the conditions that Dims=10 and Dims=30. The number in parentheses in the table is the number of algorithms which complete search at least once in literature [9]. The second sub-column of IPSO column represents the ratio of *FES* of IPSO to the Best *FES* of previous column. The number in square bracket is the value of the corresponding *FES* in the corresponding row of f_{13} .

Fun	Dims=10					Dims=30				
	Best in [9]		Best FEs	IPSO		Best in [9]		IPSO		
	Method Name	Succ. Rate		Succ. Rate	$\frac{FES}{B_FES}$	Method Name	Succ. Rate	Best FEs	Succ. Rate	$\frac{FES}{B_FES}$
1	K-PCX	100%(11)	1000	100%	0.45	K-PCX	100%(7)	2700	100%	0.30
2	K-PCX	100%(11)	2400	100%	0.33	K-PCX	100%(9)	12000	100%	0.30
3	G-CMA-ES	100%(7)	6500	100%	2.6	G-CMA-ES	100%(4)	43000	100%	5.9
4	G-CMA-ES	100%(10)	2900	100%	0.41	G-CMA-ES	40%(4)	59000	100%	0.44
5	G-CMA-ES	100%(7)	5900	100%	0.40	G-CMA-ES	100%(2)	66000	100%	0.36
6	K-PCX	100%(8)	7100	100%	0.34	G-CMA-ES	100%(6)	60000	100%	0.25
7	G-CMA-ES	4700(9)	4700	100%	1.22	G-CMA-ES	100%(10)	6100	100%	1.75
9	L-SaDE	100%(7)	17000	100%	0.76	L-SaDE	100%(3)	99000	100%	0.85
10	K-PCX	92%(2)	55000	80%	1.4	K-PCX	56%(2)	450000	44%	2.2
11	DE	48%(3)	190000	28%	3.79	G-CMA-ES	4%(1)	500000	-	-
12	K-PCX	56%(3)	8200	100%	0.95	K-PCX	20%(3)	180000	100%	0.70
13	-	-	-	8%	[1.07e+6]	-	-	-	-	-
15	L-SaDE	92%(3)	33000	16%	8.16	-	-	-	-	-

4 Conclusion

This paper proposes an improved particle swarm optimization algorithm (IPSO) to solve continuous function optimization problems. Two improvement strategies named “Vector correction strategy” and “Jump out of local optimum strategy” were employed in our improved algorithm. The algorithm was tested using 25 newly proposed benchmark instances in Congress on Evolutionary Computation 2005 (CEC2005). For these benchmark problems, the problem definition files, codes and evaluation criteria are available in <http://www.ntu.edu.sg/home/EPNSugan>. The performance of IPSO was compared with the 11 algorithms published in CEC2005. The experimental results show that the search efficiency and the ability of jumping out from the local optimum of the IPSO have been significantly improved, and the improvement strategies are effective.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (No.60970088,60775035), the National High-Tech Research and Development Plan of China (No. 2007AA01Z132), National Basic Research Priorities Programme(No. 2007CB311004) and National Science and Technology Support Plan (No.2006BAC08B06), Dean Foundation of Graduate University of Chinese Academy of Sciences(O85101JM03).

References

- [1] Eberhart, R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the sixth International Symposium on Micro Machine and Human Science, Nagoya Japan, pp. 39–43 (1995)
- [2] Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. Proc IEEE International Conference on Neural Networks. In: IEEE Service Center, Piscataway, NJ, vol. IV, pp. 1942–1948
- [3] Clerc, M.: TRIBES-Aparameter Free Particle Swarm Optimizer, <http://clerc.maurice.free.fr/PSO2002-08-10/2003-10-08>
- [4] Salman, A.: Discrete Particle Swarm Optimization for Heterogeneous Task Assignment Problem. In: Proceedings of World Multiconference on Systemics, Cybernetics and Informatics (SCI 2001) (2001)
- [5] Clerc, M.: Discrete Particle Swarm Optimization: A Fuzzy Combinatorial Black Box, http://clerc.maurice.free.fr/PSO/Fuzzy_Discrete_PSO/Fuzzy_DPSO.htm. 2000-04-01/2003-10-08
- [6] Hirotaka, Y., Kenichi, K.: A particle Swarm Optimization for Reactive Power and Voltage Control Considering Voltage Stability. In: IEEE International Conference on Intelligent System Applications to Power Systems, Rio de Janeiro (1999)
- [7] Voss, M.S., Feng, X.: Arma Model Selection Using Particle Swarm Optimization and Aic Criteria. In: 15th Triennial World Congress, Barcelon, Spain (2002)
- [8] Suganthan, P.N., Hansen, N., Liang, J.J., Deb, K.: Problem Definitions and Evaluation Criteria for the CEC 2005, Special Session on Real-Parameter Optimization (2006), http://www3.ntu.edu.sg/home/EPNSugan/index_files/CEC-05/CEC05.htm
- [9] Hansen, N.: Compilation of Results on the 2005 CEC Benchmark Function Set (2006), http://www3.ntu.edu.sg/home/epnsugan/index_files/CEC-05/comp_areresults.pdf

Mining Temporal Patterns of Technical Term Usages in Bibliographical Data

Hidenao Abe and Shusaku Tsumoto

Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
abe@med.shimane-u.ac.jp, tsumoto@computer.org

Abstract. In text mining framework, data-driven indices are used as importance indices of words and phrases. Although the values of these indices are influenced by usages of terms, many conventional emergent term detection methods did not treat these indices explicitly. In order to detect research keys in academic researches, we propose a method based on temporal patterns of technical terms by using several data-driven indices and their temporal clusters. The method consists of an automatic term extraction method in given documents, three importance indices from text mining studies, and temporal patterns based on results of temporal clustering. Then, we assign abstracted sense of the temporal patterns of the terms based on their linear trends of centroids. Empirical studies show that the three importance indices are applied to the titles of four annual conferences about data mining field as sets of documents. After extracting the temporal patterns of automatically extracted terms, we compared the emergent patterns and one of the keyword of this article between the four conferences.

Keywords: Text Mining, Trend Detection, TF-IDF, Jaccard's Matching Coefficient, Temporal Clustering, Linear Regression.

1 Introduction

In recent years, the accumulation of document data has been more general, according to the development of information systems in every field such as business, academics, and medicine. The amount of stored data has increased year by year. Document data includes valuable qualitative information to not only domain experts in the fields but also novice users on particular domains. However, detecting adequate important words or/and phrases, which are related to attractive topics in each field, is one of skilful techniques. Hence, the topic to support the detection has been attracted attentions in data mining and knowledge discovery fields. As for one solution to realize such detection, emergent term detection (ETD) methods have been developed [1][2].

However, because the frequency of the words were used in earlier methods, detection was difficult as long as each word that became an object did not appear. These methods use particular importance index to measure the statuses of the words. Although the indices are calculated with the words appearance in each temporal set of documents, and the values changes according to their usages, most conventional methods do not consider the usages of the terms and importance indices separately. This causes difficulties in text mining applications, such as limitations on the extensionality of time direction, time consuming post-processing, and generality expansions. After considering these problems, we focus on temporal behaviors of importance indices of phrases and their temporal patterns.

In this paper, we propose an integrated for detecting temporal patterns of technical terms based on data-driven importance indices by combining automatic term extraction methods, importance indices of the terms, and trend analysis methods in Section 2. After implementing this framework, we performed an experiment to extract temporal patterns of technical terms. In this experiment, by considering the sets of terms extracted from the titles of four data mining relating conferences as examples, their temporal patterns based on three data-driven importance indices are presented in Section 3. With referring to the result, we discuss about the characteristic terms of the conferences. Finally, in Section 4, we summarize this paper.

2 An Integrated Framework for Detecting Temporal Patterns of Technical Terms Based on Importance Indices

In this section, we describe a framework for detecting various temporal trends of technical terms as temporal patterns of each importance index consisting of the following three components:

1. Technical term extraction in a corpus
2. Importance indices calculation
3. Temporal pattern extraction

There are some conventional methods of extracting technical terms in a corpus on the basis of each particular importance index [2]. Although these methods calculate each index in order to extract technical terms, information about the importance of each term is lost by cutting off the information with a threshold value. We suggest separating term determination and temporal trend detection based on importance indices. By separating these phases, we can calculate different types of importance indices in order to obtain a dataset consisting of the values of these indices for each term. Subsequently, we can apply many types of temporal analysis methods to the dataset based on statistical analysis, clustering, and machine learning algorithms. An overview of the proposed method is illustrated in Figure 1.

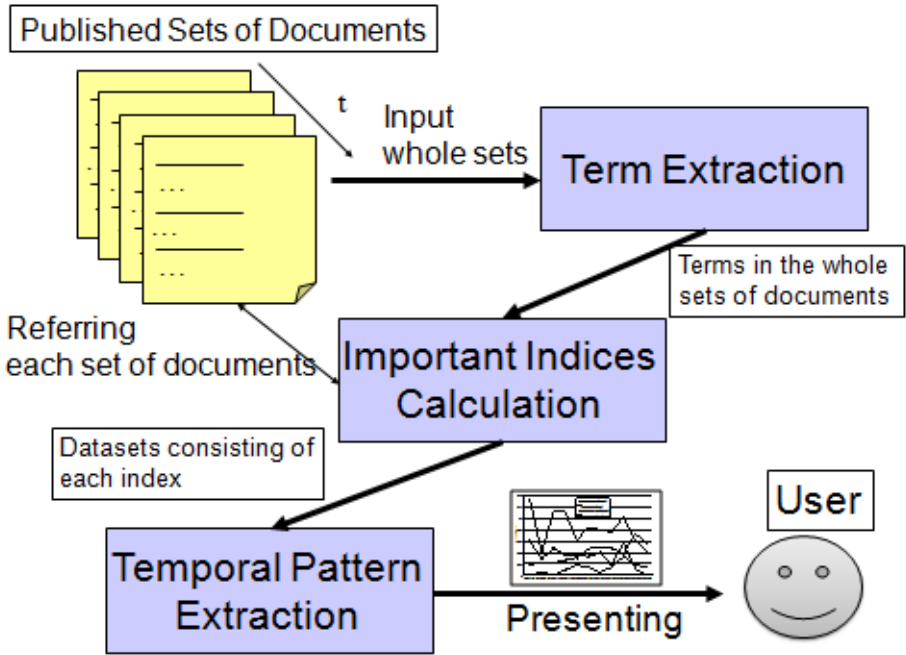


Fig. 1. An overview of the proposed remarkable temporal trend detection method

First, the system determines terms in a given corpus. There are two reasons why we introduce term extraction methods before calculating importance indices. One is that the cost of building a dictionary for each particular domain is very expensive task. The other is that new concepts need to be detected in a given temporal corpus. Especially, a new concept is often described in the document for which the character is needed at the right time in using the combination of existing words.

After determining terms in the given corpus, the system calculates multiple importance indices of the terms for the documents of each period. Further, in the proposed method, we can assume the degrees of co-occurrence such as the χ^2 statistics for terms consisting of multiple words to be the importance indices in our method.

In the proposed method, we suggest treating these indices explicitly as a temporal dataset. The features of this dataset consist of the values of prepared indices for each period.

Figure 2 shows an example of the dataset consisting of an importance index for each year.

Then, the framework provides the choice of some adequate trend extraction method to the dataset. In order to extract useful temporal patterns, there are

Term	Jacc 1996	Jacc 1997	Jacc 1998	Jacc 1999	Jacc 2000	Jacc 2001	Jacc 2002	Jacc 2003	Jacc 2004	Jacc 2005
output feedback	0	0	0	0	0	0	0	0	0	0
H/sub infinity	0	0	0.012876	0	0.00885	0	0	0	0.005405	0.003623
resource allocation	0.006060606	0	0	0	0	0	0	0	0	0
image sequences	0	0	0	0	0	0	0	0.004785	0	0
multiagent systems	0	0	0	0	0	0	0.004975	0	0	0
feature extraction	0	0.005649718	0	0.004484	0	0	0	0	0	0
images using	0	0	0	0	0	0.004673	0	0	0	0
human-robot interaction	0	0	0	0	0.004425	0	0	0	0	0
evolutionary algorithm	0	0.005649718	0	0.004484	0	0	0	0	0.002703	0.003623
deadlock avoidance	0	0	0	0	0.004425	0	0	0	0	0
ambient intelligence	0	0	0	0	0	0	0	0	0	0.003623
feature selection	0	0	0	0	0	0	0	0	0.002703	0
data mining	0	0	0	0	0.004425	0	0	0	0.002703	0

Fig. 2. Example of a dataset consisting of an importance index

so many conventional methods as surveyed in the literatures [34]. By applying an adequate time-series analysis method, users can find out valuable patterns by processing the values in rows in Figure 2.

3 Experiment: Extracting Temporal Patterns of Technical Terms by Using Temporal Clustering

In this experiment, we show the results temporal patterns by using the implementation of the method described in Section 2. As the input of temporal documents, we used the annual sets of the titles of the following four academic conferences [1]; KDD, PKDD, PAKDD, and ICDM.

We determine technical terms by using the term extraction method [6] for each entire set of documents.

Subsequently, the values of tf-idf, Jaccard coefficient, and Odds are calculated for each term in the annual documents. To the datasets consisting of temporal values of the importance indices, we extract temporal patterns by using k-means clustering. Then, we apply the meanings of the clusters based on their linear trends calculated by the linear regression technique for the timeline.

3.1 Extracting Technical Terms

We use the titles of the four data mining related conferences as temporal sets of documents. The description of the sets of the documents is shown in Table 1.

As for the sets of documents, we assume each title of the articles to be one document. Note that we do not use any stemming technique because we want to consider the detailed differences in the terms.

By using the term extraction method with simple stop word detection for English, we extract technical terms as shown in Table 2. After merging all of titles of each conference into one set of the documents, these terms were extracted for each set of the titles.

¹ These titles are the part of the collection by DBLP [5].

² The implementation of this term extraction method is distributed in <http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html> (in Japanese).

Table 1. Description of the numbers of the titles

	KDD		PKDD		PAKDD		ICDM	
	# of titles	# of words	# of titles	# of words	# of titles	# of words	# of titles	# of words
1994	40	349						
1995	56	466						
1996	74	615						
1997	65	535	43	350				
1998	68	572	56	484	51	412		
1999	93	727	82	686	72	628		
2000	94	826	86	730	52	423		
2001	110	942	45	388	63	528	109	908
2002	140	1,190	43	349	62	515	121	1,036
2003	108	842	44	340	60	520	127	1,073
2004	133	1,084	64	504	83	698	105	840
2005	113	868	76	626	101	882	150	1,161
2006	139	1,068	67	497	128	1,159	317	2,793
2007	131	1,065	67	537	196	1,863	213	1,779
2008	134	1,126	110	832	136	1,224	264	2,225
TOTAL	1,498	12,275	783	6,323	1,004	8,852	1,406	11,815

Table 2. Description of the numbers of the extracted terms

	KDD	PKDD	PAKDD	ICDM
# of extracted terms	3,232	1,653	2,203	3,033

3.2 Extracting Temporal Patterns by Using K-Means Clustering

In order to extract temporal patterns of each importance index, we used k-means clustering. We set up the numbers of one percent of the terms as the maximum number of clusters k for each dataset. Then, the system obtained the clusters with minimizing the sum of squared error within clusters. By iterating less than 500 times, the system obtains the clusters by using Euclidian distance between instances consisting of the values³ of the same index.

Table 3 shows the result of the SSE of k-means clustering. As shown in this table, the SSE values of Jaccard coefficient are higher than the other two indices: tf-idf and odds. Since we were not selected the terms with two or more words, the values of Jaccard coefficient of the terms with just one word, which are 0 or 1, are not suitable to make clusters.

3.3 Details of a Temporal Pattern of the Technical Terms

As shown in Table 4, there are several kind of clusters based on the averaged linear trends. The centroid terms mean the terms that are the nearest location

³ The system also normalized the values for each year.

Table 3. The sum of squared errors of the clustering for the technical terms in the titles of the four conferences

Conf. Name	SSE (tf-idf)	SSE (Jaccard)	SSE (Odds)
KDD	46.71	689.44	8.87
PKDD	58.76	432.21	18.17
PAKDD	35.13	325.53	10.01
ICDM	21.05	286.91	4.93

to the centroids. Then, by using the averaged degree and the averaged intercept of each term, we attempt to determine the following three trends:

- Popular
 - the averaged degree is positive, and the intercept is also positive.
- Emergent
 - the averaged degree is positive, and the intercept is negative.
- Subsiding
 - the averaged degree is negative, and the intercept is positive.

Since the terms assigned as the centroid have the highest FLR score in each pattern, the term is frequently used in the cluster by comparing to the other terms. As for the centroids of the degree and the intercept, they are the same as the average of each cluster, because the calculation of the centroid is assumed as the least-square method.

The emergent temporal patterns of the tf-idf index are visualized in Figure 3. According to the meanings based on the linear trend, the patterns #5, #6, and #8 of KDD have the emergent patterns. The emergent patterns that are #4 for PKDD, #1, #2, and #4 for PAKDD, and #4 for ICDM are also visualized.

Although these conferences share some emergent and subsiding terms based on the temporal patterns, characteristic terms can be also determined. The centroids of terms assigned as the emergent patterns⁴ express the research topics that have attract the attentions of researchers.

The emergent terms in KDD, they are related to web data and graphs. As for PKDD, the phrases ‘feature selection’ determine as emergent phrases only for this conference. The mining techniques that are related to items and text are also determined in PAKDD and ICDM. These terms indicate some characteristics of these conferences, relating to people who have been contributed for each conference.

By comparing these patterns of the indices, we can understand not only the remarkable terms but also similarity and dissimilarity of the conferences.

⁴ The emergent terms are emphasized in Table 4.

Table 4. Whole of the temporal patterns as the k-means clustering centroids on the three data-driven indices

KDD	Cluster No.	tf-idf				Jaccard Coefficient				Odds			
		Term	Avg. Deg.	Avg. Int.	Term	Avg. Deg.	Avg. Int.	Term	Avg. Deg.	Avg. Int.			
KDD	1	sequence using data mining	0.007	0.039	graph mining	0.000	0.005	sequence using data mining	-0.000	0.0001			
	2	data mining	0.759	15.348	machine learning	0.006	-0.021	mining	-0.0060	0.3012			
	3	database mining	-0.088	1.271	databases	0.018	0.351	database mining	-0.0008	0.0085			
	4	web usage mining	0.022	0.255	pattern discovery	0.002	0.014	web usage mining	0.0000	0.0004			
	5	web data	0.094	-0.273	graphs	0.025	-0.026	web data	0.0001	-0.0003			
	6	relational data	0.132	-0.444	latent	0.020	-0.054	relational data	0.0002	-0.0004			
	7	web mining	-0.001	0.448	constraints	-0.003	0.122	web mining	0.0000	0.0014			
	8	graph mining	0.140	-0.558	prediction models	0.007	-0.017	graph mining	0.0002	-0.0006			
	9	bayesian network	-0.069	0.987	interactive exploration	0.025	-0.097	bayesian network	-0.0004	0.0049			
	10	data streams	0.045	0.004	rule induction	-0.009	0.092	data streams	0.0001	0.0004			
	11	knowledge discovery	0.519	4.485	predictive modeling	0.009	0.034	data mining	-0.0093	0.1430			
	12	mining knowledge	-0.055	0.898	mining	0.022	0.627	mining knowledge	-0.0003	0.0030			
	13	high-dimensional data	-0.029	0.798	data mining	-0.014	0.176	high-dimensional data	-0.0002	0.0039			
	14	distributed data mining	-0.017	0.543	learning bayesian networks	-0.003	0.027	distributed data mining	-0.0001	0.0015			
	15	data sets	0.354	1.968	scale space exploration	0.005	0.075	databases	0.0003	0.0185			
	16				knowledge discovery	-0.008	0.135						
	17				efficient algorithms	-0.020	0.232						
	18				bayesian networks	-0.025	0.275						
	19				abstract	-0.005	0.128						
	20				categorical datasets	0.001	0.062						
PKDD	1	classification learning	0.004	0.096	spatial data	0.002	0.004	classification learning	0.0000	0.0007			
	2	knowledge discovery	-0.168	1.932	document collections	-0.033	0.262	data mining	-0.0136	0.1382			
	3	data mining	-0.104	10.324	feature selection	-0.004	0.110	learning	-0.0017	0.1188			
	4	feature selection	0.195	-0.559	learning	0.007	0.668	pattern discovery	0.0004	-0.0012			
	5	spatial data	-0.116	1.195	supervised learning	-0.028	0.252	spatial data	-0.0007	0.0059			
	6	data clustering	-0.062	0.840	applications	-0.013	0.268	data clustering	-0.0002	0.0027			
	7	data streams	0.089	0.046	knowledge discovery	0.022	0.018	data analysis	0.0002	0.0017			
	8	relational learning	0.041	0.735	rule discovery	0.006	0.067	databases	0.0000	0.0071			
	9	web	0.073	0.270	data mining	-0.008	0.082	web	0.0002	0.0016			
	10				time series	0.009	0.072						
PAKDD	1	hierarchical clustering based	0.143	-0.230	text mining	0.001	0.003	hierarchical clustering based	0.0002	-0.0004			
	2	data mining based	0.004	0.101	decision trees	0.010	0.034	data mining based	0.0000	0.0005			
	3	mining association rules	-0.122	1.201	density-based clustering	-0.012	0.090	databases	-0.0006	0.0053			
	4	text classification	0.220	-0.525	machine learning	0.011	-0.015	text classification	0.0002	-0.0005			
	5	frequent pattern mining	0.263	-0.709	association rules	0.034	-0.079	mining frequent	0.0004	-0.0009			
	6	mining structured association patterns	-0.044	0.949	data mining	0.004	0.005	knowledge discovery	-0.0006	0.0069			
	7	data mining	1.365	3.439	continuous features	0.050	-0.137	data mining	0.0003	0.0272			
	8	knowledge discovery	0.570	1.739	databases	0.033	-0.005	algorithm	-0.0052	0.0933			
	9	clustering	2.382	8.213	mixed similarity measure	-0.032	0.283	clustering	-0.0012	0.1575			
	10	text mining	-0.020	0.790	rule extraction	0.014	-0.036	text mining	-0.0003	0.0033			
	11	data clustering	0.030	0.597	applications	-0.017	0.202	data clustering	-0.0001	0.0031			
	12				model	0.073	0.092						
	13				sequential patterns	-0.037	0.267						
	14				clustering	0.011	0.782						
	15				feature selection	-0.015	0.260						
	16				bayesian classifiers	-0.008	0.167						
ICDM	1	using data mining	0.070	-0.061	data clustering	0.003	0.003	using data mining	-0.0001	0.0000			
	2	data clustering	-0.271	1.847	feature selection	-0.045	0.398	data clustering	-0.0006	0.0035			
	3	data mining approach	-0.154	1.407	sequence modeling	-0.027	0.211	medical data mining	-0.0004	0.0031			
	4	text mining	0.332	-0.005	data mining	-0.017	0.112	text classification	0.0007	0.0013			
	5	text classification based	0.206	0.010	data streams	0.051	-0.065	text classification based	0.0003	0.0000			
	6	mining	0.476	23.930	text classification	0.013	0.004	data mining	-0.0001	0.0408			
	7	web mining	-0.407	2.537	mining	0.010	0.814	web mining	-0.0009	0.0053			
	8	data mining	0.284	8.233	event sequences	-0.079	0.450	mining	-0.0110	0.2284			
	9	spatial data mining	0.065	0.468	link prediction	0.050	0.002	data mining approach	0.0001	0.0014			
	10				association rules	0.037	0.426						
	11				change	0.011	0.102						

3.4 Visualizing the Trend of a Key Word of This Article in the Different Conferences

Figure 4 shows the trend of ‘text mining’, which is included in the titles of the different four conferences, by using the tf-idf values. Their tf-idf values are increased in around 2000 and 2007 respectively. The later peak is not observed in the titles of PKDD. The trend shows that the technical topics related to this term are different in each peak. Since a novel technique itself is attractive in the earlier period, the technique tends to apply other topics by using the technique in the later periods. The trend of ‘text mining’ also shows that the technique was paid attentions in the earlier period, and the technique was applied to the other objects such as stream data.

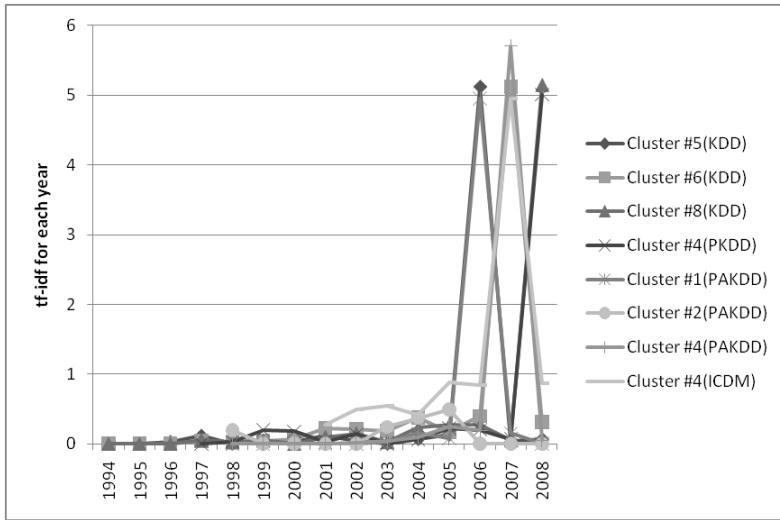


Fig. 3. The emergent temporal patterns of tf-idf through the four conferences

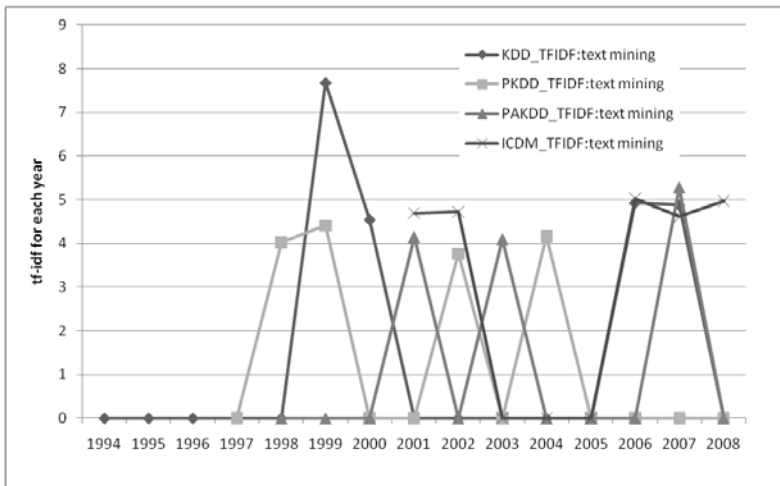


Fig. 4. The tf-idf values of 'text mining' in the titles of the four conferences

4 Conclusion

In this paper, we proposed a framework to detect temporal patterns of the usages of technical terms appeared as the temporal behaviors of the importance indices. We implemented the framework with the automatic term extraction, the three importance indices, and temporal pattern detection by using k-means clustering.

The empirical results show that the temporal patterns of the importance indices can detect the trends of each term, according to their values for each annual set of the titles of the four academic conferences. Regarding the results, we detected not only the emergent temporal patterns in the conferences, but also the difference of the research topics between the conferences by comparing the temporal patterns and their representative terms. By focusing on the trend of one keyword of this article, ‘text mining’, we show the trend of this technical topic and the difference of the trends in the different conferences.

In the future, we will apply other term extraction methods, importance indices, and trend detection method. As for importance indices, we are planning to apply evaluation metrics of information retrieval studies, probability of occurrence of the terms, and statistics values of the terms. To extract the temporal patterns, we will introduce temporal pattern recognition methods [7], which can consider time differences between sequences with the same meaning. Then, we will apply this framework to other documents from various domains.

References

1. Lent, B., Agrawal, R., Srikant, R.: Discovering trends in text databases. In: KDD 1997: Proceedings of the third ACM SIGKDD international conference on Knowledge discovery in data mining, pp. 227–230. AAAI Press, Menlo Park (1997)
2. Kontostathis, A., Galitsky, L., Pottenger, W.M., Roy, S., Phelps, D.J.: A survey of emerging trend detection in textual data mining. *A Comprehensive Survey of Text Mining* (2003)
3. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In: An Edited Volume, *Data mining in Time Series Databases*, pp. 1–22. World Scientific, Singapore (2003)
4. Liao, T.W.: Clustering of time series data: a survey. *Pattern Recognition* 38, 1857–1874 (2005)
5. The dblp computer science bibliography, <http://www.informatik.uni-trier.de/~ley/db/>
6. Nakagawa, H.: Automatic term recognition based on statistics of compound nouns. *Terminology* 6(2), 195–210 (2000)
7. Ohsaki, M., Abe, H., Yamaguchi, T.: Numerical time-series pattern extraction based on irregular piecewise aggregate approximation and gradient specification. *New Generation Comput.* 25(3), 213–222 (2007)

Automated Empirical Selection of Rule Induction Methods Based on Recursive Iteration of Resampling Methods

Shusaku Tsumoto, Shoji Hirano, and Hidenao Abe

Department of Medical Informatics, Faculty of Medicine,
Shimane University
89-1 Enya-cho Izumo 693-8501 Japan
{tsumoto,hirano,abe}@med.shimane-u.ac.jp

Abstract. One of the most important problems in rule induction methods is how to estimate which method is the best to use in an applied domain. While some methods are useful in some domains, they are not useful in other domains. Therefore it is very difficult to choose one of these methods. For this purpose, we introduce multiple testing based on recursive iteration of resampling methods for rule-induction (MULT-RECITE-R). We applied this MULT-RECITE-R method to monk datasets in UCI data repository. The results show that this method gives the best selection of estimation methods in almost the all cases.

1 Introduction

One of the most important problems in rule induction methods [1,5,6,8] is how to estimate which method is the best to use in an applied domain. While some methods are useful in some domains, they are not useful in other domains. Therefore it is very difficult to choose one of these methods.

In order to solve this problem, we introduce multiple testing based on recursive iteration of resampling methods for rule induction methods (MULT-RECITE-R). MULT-RECITE-R consists of the following four procedures: First, it randomly splits training samples(S_0) into two parts, one for new training samples(S_1) and the other for new test samples(T_1) using a given resampling method(R). Second, S_1 are recursively split into training samples(S_2) and test samples(T_2) using the same resampling strategy(R). Then rule induction methods are applied to S_2 , results are tested and given metrics(S_2 metrics) are calculated by T_2 for each rule induction methods. This second procedure, *as the inner loop*, is repeated for finite times estimated from precision set by users and the statistics of metrics are obtained. Third, in the same way, rules are induced from S_1 and metrics(S_1 metrics) are calculated by T_1 for each rule induction methods. Then S_1 metrics are compared with S_2 metrics. If the difference between both results are not statistically significant, then it is counted as a success. The second and the third procedure, *as the outer loop*, are iterated for certain times estimated from precision preset by users, which gives a total success rate which shows how many

times of total repetitions S_2 metrics predict S_1 metrics. Finally, fourth, the above results are interpreted in the following way. If a success rate is high, then this estimation method is expected to be well-performed, and the induction method which gives the best metric is selected as the most suitable induction method. If a success rate is low, then this estimation is expected not to be a good evaluation method. So a list of machine learning methods ordered by S_1 metrics is returned as an output.

We applied this MULT-RECITE-R method to monk datasets in UCI repository [7]. The results show that this method gives the best selection of methods in almost the all cases.

The paper is organized as follows: Section 2 and 3 present the strategy of MULT-RECITE-R and its algorithm. Section 4 gives experimental results. Finally, we conclude this paper in Section 5.

2 Strategy of MULT-RECITE-R

There are many reports on rule induction methods and their performance in the community of machine learning [11]. However, since each performance is different in each paper, it is very difficult to determine which method should be selected.

Each of these methods has interesting characteristics of induced rules. For example, CN2 induces a decision list subsection, while ID3 calculate a decision tree. Strangely, comparison of these features of induced rules are used as secondary, because of the difficulties in evaluation, although classification accuracy or error rate are as the primary comparison index. However, as to classification accuracy, it is pointed out that these performances may depend on applied domains [9,10], although it is easy to apply statistical methods to testing significance. Actually, it is hard and controversial to determine what factor should be applied to evaluation of rule induction methods, which remains to be an open question in machine learning.

Since our objective is to develop a method which empirically selects rule induction methods, we use accuracy as a metric for statistical evaluation in this paper [1].

The next important thing is that one may want to evaluate these rule induction methods without domain knowledge in case when domain-specific knowledge may not be applicable.

Therefore, since one of the most characteristics of resampling methods is that they are domain-independent [3,4,10], one way for evaluation is to select one method from considerable resampling methods, that is to say, to select the best rule induction method by using subsets of training samples. For example, let us consider when we have training samples, say $\{1,2,3,4,5,6,7,8,9,10\}$. Then, first, they are split it into new training samples, say $\{1,3,5,7,9\}$, and new test samples, $\{2,4,6,8,10\}$. Using new training samples, rule induction methods are applied and the results are compared with the result by the new test samples. Then the method which gives the best metric, such as the best classification rate,

¹ It is notable that our MULT-RECITE-R can be applied to any numeric metrics.

will be selected. For example, let the accuracy of the induced decision tree be equal to 0.97, and the accuracy of the rule to be equal to 0.82. Then induction of decision tree is selected as the best method. It may depend on splitting, so these procedures should be repeated for certain times, say 100 times. Several statistics of the given metrics are calculated over these 100 trials, such as average, variance, and t -statistics.

In this method, we implicitly assume that the "matryoshka" principle should be true. That is, the best method for total population can be selected from original training samples, and the best method for original training samples can be estimated from training samples generated by resampling plans. Therefore, in terms of Section 2 and 3, a domain of both R_1 and R_2 is the best select method ($R_1(F_0, F_1) \simeq R_2(F_1, F_2) = (\textit{the best method.})$)

3 An Algorithm for MULT-RECITE-R

An algorithm for MULT-RECITE-R can be described by embedding a rule induction method into the following algorithm based on a resampling scheme.

INPUTS: S_0 : Training Samples
 α : Precision for statistical test
 α_{in} : Precision for the Inner Loop
 α_{out} : Precision for the Outer Loop
 L_r : a List and Subprocedures of Rule Induction Methods
 L_m : a List of Metrics
 R : Resampling Scheme

OUTPUTS: BI : the Best Induction method selected by success rate
 M_1 : a List of Induction Methods ordered by success rates
 SR : Overall Success Rate
 BI_p : the Best Induction method selected by adjusted- p Value
 M_{1p} : a List of Induction Methods ordered by adjusted- p Values
 SR_p : Overall (Adjusted-) p Value

- 1) Set Counter to 0 ($i := 0$, $succ := 0$, $p_calc := 0$). And set B_{in} and B_{out} to $[10^{-\alpha_{in}}]$ and $[10^{-\alpha_{out}}]$, respectively².
- 2) Randomly split training samples(S_0) into two parts, one for new training samples(S_1) and the other for new test samples(T_1) using a given resampling plan(R).
- 3) Randomly split training samples(S_1) into two parts, one for new training samples(S_2) and the other for new test samples(T_2) using the same resampling plan(R). Then perform the following subprocedures.
 - 3-a) Induce rules from S_2 for each member of L .
 - 3-b) Test induced results using T_2 and Calculate given metrics (S_2 metrics).

² $\lceil x \rceil$ denotes a maximum integer which do not exceed x . For example, $\lceil 4.6 \rceil$ is equal to 4.

- 3-c) Repeat 3-b) and 3-c) for B_{in} times.
- 3-d) Calculate statistics of S_2 metrics.
- 4) Apply all the rule induction methods to S_1 . Then execute the following procedures.
 - 4-a) Test induced results by using T_1 and Calculate given metrics(S_1 metrics).
 - 4-b) Compare S_1 metrics with S_2 metrics. If the best induction method j for S_1 metrics is the same as that of S_2 metrics, then Count this trial as a success on evaluation ($succ_j := succ_j + 1$). Otherwise, then Count it as a failure.
 - 4-c) Test statistical significance between the best statistics of S_2 metrics and S_1 metrics using student t -test. If not significant, goto 5). Otherwise, Count this trial as a failure ($p_calc_j := p_calc_j + 1$).
- 5) Increment the counter ($i := i + 1$). If the counter is less than the upper bound($i < B_{out}$), goto 2). If not, goto 6).
- 6) Calculate the overall success rate ($SR := \sum succ_j / B_{out}$). And calculate an ordered list of evaluation M_1 with the success rate $succ_j / B_{out}$ of each member in L .
- 7) Calculate the overall adjusted p -value ($p := \sum p_calc_j / B_{out}$). And calculate an ordered list of evaluation M_1 with the success rate p_calc_j / B_{out} of each member in L .
- 8) Interpret the above results by the overall success rates. If a success rate is high, then this estimation method is expected to well-performed, and output the induction method j which gives the best metric is selected as the most suitable induction method ($BI := j$) and an ordered list M_1 . If a success rate is low, then this estimation is expected to be not a good evaluation method. Thus, only a list of machine learning methods ordered by S_1 metrics is returned as an output ($BI := nil$).
- 9) Interpret the above results by the overall adjusted- p values. If $p < \alpha$, then this estimation method is expected to well-performed, and output the induction method j which gives the best metric is selected as the most suitable induction method ($BI_p := j$) and an ordered list M_{1p} . If $p \geq \alpha$, then this estimation is expected to be not a good evaluation method. Thus, only a list of machine learning methods ordered by S_1 metrics is returned as an output ($BI_p := nil$).

4 Experimental Results

We applied this MULT-RECITE-R method to monk datasets in UCI repository [7]. In these experiments, we set L_r , L_m , α , α_{in} and α_{out} be equal to the same values as the above Monk's problems and set R to {2-fold cross-validation, the Bootstrap method}.

Unfortunately, in these databases, test samples are not given independently. So we first have to generate test samples from the original training samples. to evaluate our MULT-RECITE-R methods in the same way as evaluation shown in Section 3. First, given samples are randomly split into training samples(S_0) and test samples(T_0). This T_0 correspond to test samples of Monk's problems, and

Table 1. Results of S_2 and S_1 Metrics(Accuracy)

Domain Samples		S_2 Metric		
		C4.5	AQR	CN2
Monk-1	62	84.3±1.5	90.2±0.9	92.0±1.8
Monk-2	86	62.6±2.4	74.8±1.9	59.1±1.7
Monk-3	62	87.7±1.4	82.5±1.3	84.8±0.9
Domain Samples		S_1 Metric		
		C4.5	AQR	CN2
Monk-1	124	85.3±0.9	91.2±0.5	93.0±0.2
Monk-2	169	66.7±1.3	75.8±0.7	60.1±0.8
Monk-3	122	89.7±0.2	83.5±0.4	83.8±0.5

Table 2. Success Rate (100 Trials)

Domain	Overall			
	Success Rate	Success Rate		
		C4.5	AQR	CN2
Monk-1	94	9	12	73
Monk-2	74	19	31	24
Monk-3	90	79	6	5

Table 3. Adjusted- p Value (100 Trials)

Domain	Overall			
	p -Value	Adjusted- p Value		
		C4.5	AQR	CN2
Monk-1	0.02	0.01	0.01	0.00
Monk-2	0.10	0.04	0.02	0.04
Monk-3	0.05	0.01	0.02	0.02

S_0 correspond to training samples of Monks problems. Then MULT-RECITE-R method is applied to new training samples. This splitting procedure is repeated for 100 times in order for the effect of random sampling to be small.

The above experimental results give us three interesting results, although all of the applied databases are of small size.

First, 2-fold repeated cross validation performs slightly better than the Bootstrap method, which corresponds to the characteristics derived by [2,3]. Therefore, for predictive use, evaluation by cross-validation would be better, although the variance of estimation will be larger.

Second, the best selected method does not always perform better than other two methods. That is, in some generated samples, other methods will perform better. Finally, in the cases when MULT-RECITE-R does not go well, the differences of three rule induction methods in accuracy are not so significant. That is, we can select any of three methods, although the accuracy of each method is not so high.

5 Conclusion

One of the most important problems in rule induction methods is how to estimate which method is the best to be used in an applied domain. For this purpose, we introduce multiple testing based on recursive iteration of resampling methods for rule-induction (MULT-RECITE-R). We apply this MULT-RECITE-R method to three original medical databases and seven UCI databases. The results show that this method gives the best selection of estimation methods in almost all cases.

References

1. Clark, P., Niblett, T.: The CN2 Induction Algorithm. *Machine Learning* 3, 261–283 (1989)
2. Efron, B.: How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* 82, 171–200 (1986)
3. Efron, B., Tibshirani, R.: *An Introduction to the Bootstrap*. Chapman and Hall, London (1994)
4. McLachlan, G.J.: *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, New York (1992)
5. Michalski, R.S.: A Theory and Methodology of Machine Learning. In: Michalski, R.S., Carbonell, J.G., Mitchell, T.M. (eds.) *Machine Learning - An Artificial Intelligence Approach*. Morgan Kaufmann, Palo Alto (1983)
6. Michalski, R.S., et al.: The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. In: *Proceedings of AAAI 1986*, pp. 1041–1045. AAAI Press, Palo Alto (1986)
7. Murphy, P.M., Aha, D.W.: *UCI Repository of machine learning databases [Machine-readable data repository]*. University of California, Department of Information and Computer Science, Irvine
8. Quinlan, J.R.: *C4.5 - Programs for Machine Learning*. Morgan Kaufmann, CA (1993)
9. Schaffer, C.: Overfitting Avoidance as Bias. *Machine Learning* 10, 153–178 (1993)
10. Schaffer, C.: Selecting a Classification Method by Cross-Validation. *Machine Learning* 13, 135–143 (1993)
11. Thrun, S.B., et al.: *The Monk's Problems- A performance Comparison of Different Learning algorithms*. Technical Report CS-CMU-91-197, Carnegie Mellon University (1991)

Adaptive Web-Based Instruction for Enhancing Learning Ability

Wawta Techataweewan

Srinakharinwirot University
114 Sukhumvit 23 Road, Wattana
Bangkok 10110, Thailand
walta@swu.ac.th

Abstract. Web technology in an instructional environment is primarily dedicated to distributing course materials to supplement traditional classroom learning. It also uses designed intelligence to adapt to learners' specific needs. The main purposes of this study were to construct and determine the efficiency of adaptive web-based instruction for LIS students. The web-based content was designed to adapt to three levels of learning ability: high, moderate and low. The system automatically collects data concerning each learner's behavior, determines the level of learning ability and provides suitable content for each learner. In addition, this web-based instruction evaluated using 80/80 standard efficiency criteria and compared the learning achievement of the students who learned through the adaptive WBI and those who learned through traditional methods. Finally, a study was conducted to evaluate students' level of satisfaction with adaptive web-based instruction. The sample consisted of 60 undergraduate students from Srinakharinwirot University, majoring in Library and Information Science. Research instruments included adaptive web-based instruction, achievement test, and satisfaction evaluation form. The findings revealed that the adaptive web-based instruction met the efficiency criteria at 80.78/81.17. The learning achievement of students who learned through adaptive web-based instruction was significantly higher than those who learned through traditional methods at 0.01 level, and they were satisfied with the web-based instruction at a good level.

Keywords: Adaptive learning, Web-based instruction, Learning ability.

1 Introduction

The National Education Act of B.E. 2542 (1999) of Thailand in Section 22 states that, "Education shall be based on the principle that all learners are capable of learning and self-development, and are regarded as being most important; The teaching-learning process shall aim at enabling learners to develop themselves at their own pace and to the best of their potential." This policy introduced the student-centered approach into all education levels (Office of the National Education Commission, 2008). In a systematic analysis of classroom practices that spanned over a century, classroom instruction as a model ranged widely from a teacher-centered to a student-centered curriculum (Khan, 1997: 40). In the teacher-centered approach, knowledge is transmitted from

teacher to students, the latter passively receiving information from the teacher. On the other hand, student-centered learning focuses on student's needs, abilities, interests and learning styles. The teacher facilitates learning that requires students to be active, responsible participants in their own learning and lifelong learning process. Students are responsible for setting personal, realistic goals and determining the steps toward achieving their learning objectives (Koocharoenpaisal, 2005: 20-21).

The web facilitates easy distribution of knowledge and instructional resources regionally, globally or within an enterprise. The current use of the web in the instructional environment is primarily dedicated to distributing course materials to supplement traditional classrooms such as syllabus, slide presentations, assignments, calendars, notes and course readings. Additionally, teachers incorporate electronic mail and news groups to promote learning and interaction between students (Sanrach, 2002: 24). In accordance with the provision of interaction and flexibility, the web facilitates the student-centered approach, creating a motivating and active learning environment (Zangyuan, 2010: 1). For distance education, the web is a unique tool that allows a true shift from a teacher-centered learning paradigm to an authentic student-centered learning paradigm. For students to access a variety of resources, a more flexible course design process is required that considers learners' needs and shapes instructional elements accordingly.

2 Adaptive Web-Based Instruction

Web-based instruction (WBI) can be viewed as an innovative approach for delivering instruction to a remote audience. Instruction involves the delivery of information and activities that facilitate learners' attainment of specific learning goals (Smith *et al.*, 1993: 2). Early WBI offered the same course content using the same didactic presentation method to all students. With growing data, number and diversity of users, the complexity of human-computer interaction continually increases. Increasingly, modern software adapts to users, for example, programs that automatically adapt themselves to unacquainted or inexperienced users (Chen *et al.*, 2005: 104). Recent adaptive learning systems take into account user preferences, skills and interests to enhance knowledge acquisition in a personalized computer-supported teaching environment. To behave adaptively, WBI must be able to recognize a learner's needs or preferences and generate a concept of each learner. Adaptive web-based instruction builds an internal model of an individual learner and modifies instruction accordingly. Adaptive WBI has been defined in many different ways, which are closely related to two aspects of its adaptive systems as follows:

2.1 Adaptive presentation means to adapt the content of a page accessed by learners to current knowledge depending on their learning characteristics. This is similar to a teacher who typically adapts teaching content for each individual student based on their unique needs. Existing adaptive presentation technologies deal with text adaptation that adjusts for different learners at different times who may receive different content even when on the same page. Adaptive presentation also provides the same page to meet the needs of every student.

2.2 Adaptive navigation means to support students with hyperspace orientation and navigation by changing the appearance of visible links. Adaptive navigation can be

considered a generalization of curriculum sequencing technology in the hypermedia context with the same goal of helping students find the optimal path to learning materials. Curriculum sequencing means to organize instructional units or curriculum elements into a hierarchy of related courses, models, lessons and presentations. Each instructional unit typically has an objective (Chen *et al*, 2006: 379). Three techniques of adaptive navigation include adaptive guiding, adaptive link annotation and adaptive link hiding (Sanrach, 2002: 25-26).

The most well-known adaptive learning systems are AHA, InterBook, and APeLS (Sluijs, et al. 2009: 46). AHA is an open source adaptive hypermedia system mainly used in the education domain. It supports several adaptation techniques such as adaptive guiding, link annotation and adaptive link hiding (De Bra *et al*, 2006: 133). InterBook affords an environment for authoring and providing adaptive online textbooks. It supports adaptive navigation that guides users' hyperspace exploration using annotation (Brusilovsky, 1998: 293). APeLS or Adaptive Personalized eLearning Service is a multi-model, meta-data driven adaptive hypermedia system that separates the narrative, content and learner into different models. The adaptive engine in APeLS is a rule-based engine that produces a model for personalized courses based on a narrative and the learner model (Conlan, 2002: 100). Currently, most adaptive learning systems consider learners' preferences and interests for developing individualized service. However, some systems neglect the importance of learner ability when implementing personalized mechanisms. Likewise, some researchers emphasize that personalization should consider levels of learner knowledge, especially in relation to learning (Papanikolaou *et al*, 2002: 338-339). In summary, effective consideration of learner ability can improve individual learning performance and achievement of learning goals.

3 System Design and Components

The system was designed for the adaptation of course content to recognize and appreciate students' learning ability. Students' learning ability level is approached through the use of various learning "agents." The agents are employed to adapt the presentation of course content for each learning unit according to the ability of the learner, as determined by each formative assessment test. The system components include learning interface agent, feedback agent, courseware provision agent and courseware management agent as shown in figure 1.

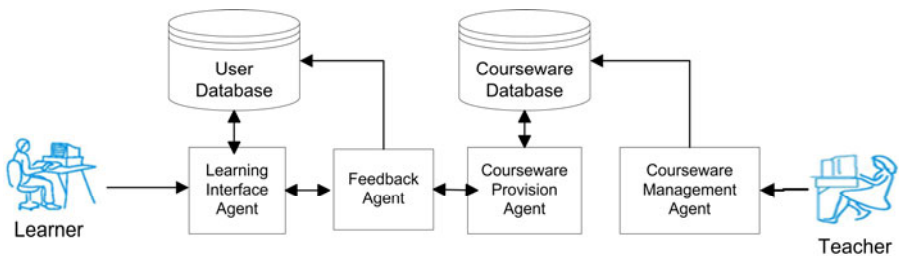


Fig. 1. Adaptive WBI architecture and components

The learning interface agent provides a friendly user interface to interact with learners by checking their accounts in the user database. If the learner has already registered, the system retrieves their individual learning profile. The agent also conveys the learner’s feedback to the feedback agent, and receives results from the system adaptive engines to display appropriate course content to the learner.

The feedback agent obtains learner information from the learning interface agent of the user database and passes the formative assessment scores of each learning unit to the courseware provision agent for evaluating that learner’s ability and providing appropriate course content. Meanwhile, the feedback agent sends information concerning a user’s learning status to the user database for updating the learner’s profile.

The courseware provision agent collects a learner’s scores from the feedback agent to evaluate their ability with criteria such as, 8-10 = high level, 5-7 = moderate level and 3-4 = low level. If the scores lower than 3, the learner has to repeat the past learning unit. The courseware provision agent then retrieves the course content and matches it to the learner’s ability from the courseware database.

The course management agent helps teachers manage the course on the web. They access the system to upload, delete or revise the courseware database, which is maintained by the courseware management agent. A teacher can design course content and configure specific criteria based on students’ various levels of learning ability.

Using the adaptive presentation approach in an adaptive learning system, the course content in this research was designed for high, moderate and low levels of learning ability. Specifically, the course content was divided into three different amounts of lessons per unit for each of the three levels.

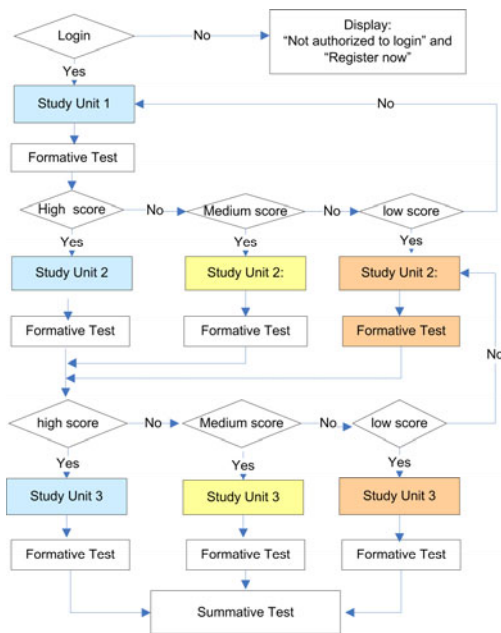


Fig. 2. Adaptive WBI work flow

While studying, students in the experimental group could shift to a higher or lower level depending on their formative assessment scores until they finished the course, as diagrammed in figure 2.

Actually, most LIS students felt that the Dewey Decimal Classification System (DDC) course was difficult for them because there was a lot of content and memorization. Therefore, the web-based screen design was concerned with motivation through a balance of graphics and text, plus animation games for practice, as shown in figures 3.



Fig. 3. Learning webpage design

4 Research Questions and Objectives

The purpose of this research was to develop adaptive web-based instruction on the Dewey Decimal Classification System, which is a compulsory course for Library and Information Science students. Course content was designed to adapt to high, moderate and low levels of learning ability. The system automatically collected data on each student's behavior, determined their level of learning ability, and provided suitable lessons for them. This researched addressed the following questions.

1. How can adaptive WBI be designed to appreciate student's learning ability?
2. How does the outcome from adaptive WBI compare to traditional classroom learning?
3. To what degree are students satisfied with adaptive WBI?

The objectives of this study were:

1. To develop adaptive WBI for students according to their individual learning ability.
2. To elucidate differences between learning achievement in traditional classrooms and adaptive WBI.
3. To analyze students' level of satisfaction with adaptive WBI.

5 Research Methodology

The study employed an experimental research methodology using pre-test, post-test control group design. The research instruments included adaptive WBI, the learning achievement test and the system satisfaction evaluation form. An experiment was conducted to investigate the system's educational effectiveness. The study included 60 Library and Information Science students at Srinakharinwirot University who had not previously taken the Dewey decimal classification course. The students were purposely distributed into two equal groups: the control group consisted of third-year students who were enrolled in a conventional learning environment while the experimental group consisted of second-year students who were placed in an adaptive web-based learning environment. Both groups took the pre-test before the course and post-test after completing it. The efficiency criteria of the system was $E1/E2 = 80/80$, that compared the percentage of formative assessment scores with the summative assessment scores.

6 Research Results

The means and standard deviations of data for the pre-test, post-test and formative tests were analyzed using a t-test, yielding quantitative results. This study aimed to address the research problems as follows:

6.1 Adaptive WBI was designed to appreciate students' learning ability through learning content design and the adaptive engine of the system. The learning ability levels of the learners were evaluated at high, moderate and low levels depending on the formative test scores from each learning unit. The system then determined the course content to match their level. Therefore, the students study the same course content but different amount of content in each unit depending on their learning ability. The outcome for adaptive WBI compared to the traditional classroom indicated that adaptive WBI met the efficiency criteria at 80.78/81.17, as shown in table 1. There were 30 questions on the three formative tests and 20 questions on one summary test with adaptive WBI.

Table 1. Efficiency of adaptive WBI as criteria of E1/E2

Assessment	Scores	\bar{X}	Percentage
Formative tests (E1)	727	24.23	80.78
Summary test (E2)	487	16.23	81.17

6.2 Differences between learning achievement in the traditional classroom and adaptive WBI were examined by comparing the pre-test and post-test scores. Different scores between the experimental and control groups were used as an index of learning achievement. The first set of data in table 2 shows a significant difference at .01 between the achievements of the two groups, indicating that learners benefited more from the adaptive learning environment.

Table 2. Comparison of learning outcome between experimental and control groups

Students	N	\bar{X}	S.D.	t-test
Experimental group	30	16.23	1.25	3.248**
Control group	30	14.73	2.10	

** Significant at $P < .01$.

6.3 All 30 students in the experimental group who used adaptive WBI responded to the satisfaction questionnaire after finishing the course. The satisfaction evaluation form was analyzed using the Likert method, with a scale consisting of the elements: 5 = Very Good, 4 = Good, 3 = Fair, 2 = Less and 1 = Unsatisfied, respectively. Results from the evaluation showed that students were overall satisfied with studying with adaptive WBI at a good level. Specific evaluation topics such as web design, learning facilities and course content also yielded good levels of satisfaction (as shown in table 3). Many students' opinions indicated that they preferred learning on the web, such as, "It isn't difficult to build the DDC number." "It's cool;" "I love the games. It makes me easily memorize the DDC scheme;" "Not boring to study on the web." and so on.

Table 3. Satisfaction evaluation of adaptive WBI

Evaluation lists	\bar{X}	S.D.	Meaning
1. Webpage design			
1.1 Menu sequencing	4.37		Good
1.2 Menu positioning	4.57		Excellent
1.3 Attractive design	4.57		Excellent
1.4 Font formatting	4.53		Excellent
1.5 Font sizes	4.50		Excellent
1.6 Font colors	4.53		Excellent
1.7 Graphics	4.13		Good
1.8 Overall web design	4.43		Good
Total	4.45	0.15	Good
2. Learning management system			
2.1 User friendly	4.40		Good
2.2 Interactive styles	4.10		Good
2.3 Guidance and feedback	4.30		Good
2.4 Full required functions	4.13		Good
2.5 Overall LMS	4.20		Good
Total	4.23	0.12	Good
3.Course content			
3.1 Amount of content	4.27		Good
3.2 Content sequencing	4.30		Good
3.3 Understandable content	4.27		Good
3.4 Suitable content	4.33		Good
3.5 Interesting content	4.10		Good
3.6 Presentation	4.43		Good
3.7 Examples	4.33		Good
Total	4.29	0.10	Good
Total	4.32	0.12	Good

7 Discussion

Adaptive WBI is a more powerful teaching method to support the student-centered learning approach. The program was designed to assign a suitable amount of course content for each student's learning ability regarding their concentration and memorization. Additionally, the DDC scheme also offered practice with animation games, a chat room, bulletin board and course documents that motivated students and helped them remember the material. For examples, games may promote further the students' weakening of attention and concentration is an internal power of their mind, or controlled by external elements such as computer which totally depends on the person's own. The experimental students using WBI improved their learning retention and earned higher scores than the control students. Furthermore, students participating in the DDC course completed the lessons more quickly than those in the conventional classroom. The main reasons for this remarkable difference were the availability of computers and the freedom of learning. Consequently, a lot of students finished the course within a very brief period of time.

8 Conclusions and Future Works

Adaptive WBI aims to formulate an environment that supports students to optimize their individual learning ability. The proposed adaptive WBI for enhancing learning ability provides learning content that can be adapted to the various abilities of students. Meanwhile, this approach can be integrated for all courseware while developing individual learning ability. Experimental results indicated that WBI can successfully recommend appropriate course materials to students based on individual ability and help them learn more effectively in a web-based learning environment. In general, web-based instruction meets the needs of students. In this study, the system supports only the adaption of the learning content to each student. Adaptive WBI proposes an adaptive presentation approach to reveal students' learning needs. Further study should aim to more identify the personal needs and learning styles. Therefore, teachers can set several requirements about a course to response the students' needs and styles. Likewise, the adaptive WBI should compare not only classroom, but also the simple web-based learning environment.

References

- Brusilovsky, P., Eklund, J., Schwarz, E.W.: Web-based education for all: A tool for development adaptive courseware. *Computer Networks and ISDN Systems* 3(1-77), 291–300 (1998)
- Brusilovsky, P., Schwarz, E., Weber, G.: ELM-ART: An intelligent tutoring system on World Wide Web. In: Lesgold, A.M., Frasson, C., Gauthier, G. (eds.) *ITS 1996. LNCS*, vol. 1086, pp. 261–269. Springer, Heidelberg (1996)
- Chen, C., Liu, C., Chang, M.: Personalized curriculum sequencing utilizing modified item response theory for web-based instruction. *Expert Systems with Applications* 30, 378–396 (2006)

- Chen, S.Y., Magoulas, G.D.: Adaptive and adaptive hypermedia systems. In: Universal Access in the Information Society. IRM, Hershey (2005)
- Conlan, O., Wade, V.P., Bruen, C., Gargan, M.: Multi-model, metadata driven approach to adaptive hypermedia services for personalized elearning. In: De Bra, P., Brusilovsky, P., Conejo, R. (eds.) AH 2002. LNCS, vol. 2347, pp. 100–111. Springer, Heidelberg (2002)
- De Bra, P., Smiths, D., Höver, K.M.: The design of AHA! In: Proceedings of the 17th ACM Conference on Hypertext and Hypermedia: Hypertext 2006, Odense, Denmark, pp. 133–134 (December 2006)
- Khan, B.H. (ed.): Web-based instruction. Educational Technology, Englewood Chiffs (1997)
- Koocharoenpibal, N.: A development of learner-centered Science curriculum on “Chemicals in daily Life” for lower secondary students, Dissertation, Ed.D. (Science Education), Bangkok: Srinakharinwirot University (2005)
- Papanikolaou, K.A., Grigoriadou, M., Magoulas, G.D., Kornilakis, H.: Towards new forms of knowledge communication: the adaptive dimension of a web-based learning environment. *Computers & Education* 39, 333–360 (2002)
- Office of the National Education Commission, International Relations and Cooperation Center for Educational Reform, National Education Act of B.E. 2542 (1999), <http://www.moe.go.th/English/edu-act.htm> (retrieved April 17, 2008)
- Sanrach, C.: Adaptive and intelligent web-based learning environment. *Journal of Academic Computer Education* 1(1), 24–45 (2002)
- Sluijs, K., Höver, K.M.: Integrating adaptive functionality in LMS (2009) <http://online-journals.org/i-jet/article/view/958> (retrieved December 10, 2009)
- Smith, P.L., Ragan, T.J.: Instructional design. Macmillan, New York (1993)
- Zangyuan, O.: The Application of an adaptive, web-based learning environment on Oxidation-reduction Reactions. *International Journal of Science and Mathematics* 8(1), 1–23 (2010)

Extracting Comparative Commonsense from the Web

Yanan Cao^{1,2}, Cungen Cao¹, Liangjun Zang^{1,2},
Shi Wang¹, and Dongsheng Wang^{1,2}

¹ Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
No. 6 Kexueyuan South Road Zhongguancun,
Beijing 100190, China

² Graduate University of Chinese Academy of Sciences
No. 19 Yu Quan Road, Shi Jing Shan Distinct,
Beijing 100049, China

Abstract. Commonsense acquisition is one of the most important and challenging topics in Artificial Intelligence. Comparative commonsense, such as "In general, a man is stronger than a woman", denotes that one entity has a property or quality greater or less in extent than that of another. This paper presents an automatic method for acquiring comparative commonsense from the World Wide Web. We firstly extract potential comparative statements from related texts based on multiple lexico-syntactic patterns. Then, we assess the candidates using Web-scale statistical features. To evaluate this approach, we use three measures: coverage of the web corpora, precision and recall which achieved 79.2%, 76.4% and 83.3%, respectively in our experiments. And the experimental results show that this approach profits significantly when the semantic similarity relationships are involved in the commonsense assessment.

1 Introduction

Commonsense knowledge plays an important role in various areas such as natural language understanding (NLU), information retrieval (IR), question answering (QA), etc. For decades, there has been a thirst in artificial intelligence research community for large-scale commonsense knowledge bases. Since the hand-coded knowledge base suffers from semantic gaps and noises, and the human effort should still be involved in the maintenance [1], much recent work focuses on automatic strategies to acquire commonsense knowledge.

As information people use every day, commonsense knowledge encompass many aspects of life. It relates to an individual object, phenomena and activity, or relations between different concepts and a sequence of events. This paper concentrates on a commonsensical concept-level relation, which is defined as follows.

Comparative Relation. We say that concept c_1 and c_2 are comparative if and only if the natural language description "In general, c_1 is relatively * than c_2 " is reasonable and acceptable, where "*" is an adjective expressing the comparative

degree. This statement denotes that the concept c_1 has a property or quality greater or less in extent than that of c_2 . We use the triple (c_1, c_2, p) to express it formally. For a specific adjective in this statement, we call it a comparative property of the given pair of concepts. For example, "In general, a man is relatively stronger than a woman" describes a comparative relation between the concepts man and woman, of which *stronger* is the comparative property. The formalization of this statement is $(man, woman, stronger)$.

In this paper, we aim to acquire comparative commonsense about given pairs of concepts from the Web, which is a vast source of information. Firstly, we submit instantiated lexico-syntactic patterns as query terms to the search engine, such as "men are * than women", and extract potential comparative statements from the retrieved texts. Then, we assess the candidates based on multiple Web-scale statistical features. The semantic similarity relationships, including synonym and antonym, are involved in the computing of these statistics. This improves the effectiveness of our approach. In the experiments, we selected twenty pairs of concepts as the input data and the result set of comparative statements had a precision of 76.4% and a recall of 83.3%. Besides, we show the high coverage of the comparative information in a test web corpus, which achieved 79.2%.

The remainder of this paper is organized as follows. In Section 2, we review some related work on commonsense knowledge acquisition from texts. Section 3 and Section 4 describe the two-stage method for extracting and assessing comparative commonsense in detail. In Section 5, we evaluate our approaches and discuss the experimental results. Finally, we conclude in Section 6.

2 Related Work

To acquire commonsense knowledge, a number of efforts tap on textual data sources. The pioneering work mentioned in [2] recognized text as a source of both explicit and implicit commonsense knowledge. Capitalizing on interpretive rules, they obtained a broad range of general propositions from noun phrases and clauses. And they tried to derive stronger generalizations based on the statistical distribution of the obtained claims. However, to design abundant rules is indeed time-cost and the extraction process was not implemented automatically.

[3] extracted commonsensical inference rules from coordinated verb phrases in Japanese text. This work was based on the assumption that if two events shared a common participant, then the two events would probably be a probably-follow relation. They used an unsupervised approach to select the inference rules based on lexical co-occurrence information. The results were evaluated by multi-subjects, and just 45% of the selected rules were accepted by 80% of the human subjects. The low precision showed that co-occurrence is important but not adequate to assess the specific causal relation between two events, because there are other event-level relations such as *IsPreferableTo* in [4].

[4] used the Web data to clean up and augment existent commonsense knowledge bases. For specific binary event-level relations such as causality, they used

complete or partial event descriptions and lexico-syntactic patterns to mine potential instances of the relations, and assessed these candidates based on Web-scale statistics. For higher-level predicates, the relational definitions were involved in the assessment, which extended the work in [3]. The advantage of this work is that some implicit instances of relations can be assessed at the level of the Web rather than found in one particular sentence.

In our recent work [5], we acquired commonsense knowledge about properties of concepts by analyzing how adjectives are used with nouns in everyday language. We firstly mined a large scale corpus for potential concept-property pairs using lexico-syntactic patterns and then filtered noises based on heuristic rules and statistical approaches. For each concept, we automatically selected the commonsensical properties and evaluated their applicability. The precision of the result-set achieved 67.7%. In the following, we continue our research on the relations between concepts and properties, from the view of comparing.

3 Extracting Candidate Comparative Statements

This section describes the first phase of acquiring comparative commonsense: we extract candidate commonsense from the Web using linguistically-motivated patterns [6]. In the text, comparative relations are expressed in various forms. Table1 lists several patterns followed by corresponding instances, in which *NP* denotes *Noun Phrase*.

Table 1. Examples of patterns designed to extract comparative commonsense

Pattern	Instance
NP_1 be * than NP_2	men are stronger than women
NP_1 be more * than NP_2	men are more powerful than women
NP_2 be not * than NP_1	women are not stronger than men
NP_2 be less * than NP_1	women are less powerful than men
to compare NP_1 and NP_2 , NP_1 be * compared with NP_2 , NP_1 be *	to compare men and women, men are stronger compared with women, men are stronger

When we perform the extraction process, these patterns are automatically instantiated with given pairs of concepts to generate query terms. For example, to acquire comparative statements about *man* and *woman*, we instantiate the first pattern as "men are * than women" and the last one as "compared with women, men are *". Issuing the query terms, we take advantage of the search engine to retrieve relevant texts from the Web and obtain the hit count of each query string. And subsequently, we extract a comparative property from each snippet matched with the "*" wildcard, which constitutes a candidate triple with its corresponding concept-pair.

There are two explicit constraints on the extraction. The first is that the matching snippet should be an adjective phrase, and we just extract the head word no matter if it has a modifier or a complement. For example, from the text "men are much stronger than women", we extracted the comparative property "stronger", while "much" is discarded. We implement this constraint by analyzing the combination of POS in adjective phrases. Second, we extracted the comparative statements in a context-independent manner. That means we don't extract the specific context even if a statement is not acceptable anymore without it. For example, in the sentence "From Shanghai to Lhasa, a train is faster than an airplane", the extracted statement (*train, airplane, faster*) is dependent on the context "from Shanghai to Lhasa", or else it violates the general knowledge. To avoid obvious errors induced by this strategy, we restrict the matching snippet to be a single sentence or to have a common modifier such as "generally speaking", "as we all know", etc.

4 Assessing Candidate Comparative Statements

During the extraction process, we acquire all potential comparative statements. However, the Web contains massive unreliable or biased information [7], and some knowledge we obtained is inconsistent with the commonsensible facts. So, we verify each candidate statement based on multiple Web-scale statistics.

4.1 Statistical Features for Commonsense Assessment

Feature 1 Occurrence Frequency. Because a frequently mentioned piece of information is typically more likely to be correct than an item mentioned once or twice, high frequency is an important feature for truth. Given a candidate statement (c_1, c_2, p) , we could easily obtain its occurrence frequency $Freq(c_1, c_2, p)$ in the web corpora using hit counts of instantiated queries, which are returned by the search engine:

$$Freq(c_1, c_2, p) = \sum_{pt \in PT} hits(pt(c_1, c_2, p)), \quad (1)$$

where PT is the set of predefined patterns (referred in Table 1) and $hits(pt(c_1, c_2, p))$ represents the hit count for a query which is an instance of the pattern pt .

Feature 2 Confidence. Although frequency is an important feature, it's not a guarantee of truth. Given a pair of concepts c_1 and c_2 , we use confidence to indicate the probability of a property p to be their comparative property. More specifically:

$$Conf(c_1, c_2, p) = Freq(c_1, c_2, p) / \sum Freq(c_1, c_2, p) \quad (2)$$

In this formula, the denominator is the number of potential comparative relations which c_1 and c_2 participant in, and "*" is any potential comparative property of c_1 and c_2 .

Feature 3 the Number of the Matched Patterns. The last feature is the number of different expression forms of the same comparative statement. It's motivated by the assumption that if a potential statement is extracted by multi-patterns, it seems more reliable.

4.2 Semantic Similarity Relationships in the Candidate Set

We note that, the candidate statements are not independent, and there are semantic similarity relationships among them. For example, $(woman, man, weaker)$ is consistent with the statement $(man, woman, stronger)$, while $(man, woman, weaker)$ is opposed to it. We make use of these semantic similarity relationships in the computing of the statistics. To see this, we first begin by introducing several fundamental notations.

Given a potential comparative statement (c_1, c_2, p) , we use $SYN_p = \{x|x \text{ is a synonym of } p\} \cup \{p\}$ to denote the set of all synonyms of the property p , and another set $ANT_p = \{y|y \text{ is a synonym of } p\}$, which contains all antonyms of p . Based on these two sets, we divide other statements into the following categories according to the relationship between it and (c_1, c_2, p) .

- A **supporting example**, which has the same meaning of the statement (c_1, c_2, p) . For instance, $(man, woman, more powerful)$ and $(woman, man, weaker)$ are both supporting examples of $(man, woman, stronger)$. It belongs to the set

$$SupSet(c_1, c_2, p) = \{(c_1, c_2, p') | p' \in SYN_p\} \cup \{(c_2, c_1, p'') | p'' \in ANT_p\}$$

- A **counterexample**, which is opposed to the statement (c_1, c_2, p) in semantic. Using above instance $(man, woman, stronger)$, $(man, woman, weaker)$ and $(woman, man, stronger)$ are both its counterexamples. The set of the counterexamples is

$$CntSet(c_1, c_2, p) = \{(c_1, c_2, p') | p' \in ANT_p\} \cup \{(c_2, c_1, p'') | p'' \in SYN_p\}$$

- An **irrelative example** is the statement that is neither a supporting example nor a counterexample. For example, $(man, woman, braver)$ and $(man, woman, stronger)$ are irrelative to each other.

Then, we use $Support(c_1, c_2, p)$ to denote the frequency of supporting examples of the given triple (c_1, c_2, p) . More specifically:

$$Support(c_1, c_2, p) = \sum_{(c_1, c_2, p') \in SupSet(c_1, c_2, p)} Freq(c_1, c_2, p') \quad (3)$$

And we use $Counter(c_1, c_2, p)$ to denote the frequency of counterexamples of the given triple (c_1, c_2, p) . It is computed as follows:

$$Counter(c_1, c_2, p) = \sum_{(c_1, c_2, p') \in CntSet(c_1, c_2, p)} Freq(c_1, c_2, p') \quad (4)$$

Intuitively, if supporting examples of the statement (c_1, c_2, p) are much more than its counterexamples, this statement is more likely to be true. So, we use $ExtFreq(c_1, c_2, p)$ instead of $Freq(c_1, c_2, p)$ to assess potential comparative statements:

$$ExtFreq(c_1, c_2, p) = Support(c_1, c_2, p) - Counter(c_1, c_2, p) \quad (5)$$

And confidence is accordingly computed as follows:

$$ExtConf(c_1, c_2, p) = ExtFreq(c_1, c_2, p) / \sum Freq(c_1, c_2, *) \quad (6)$$

Although the features frequency and confidence are not novel, we combine Web-scale statistics and semantic relationships, and this improves the performance of our approach.

5 Experiments

In this section, we report some experimental results on comparative commonsense acquisition from Chinese web corpora. We computed the estimated coverage to prove the Web a proper resource for comparative commonsense acquisition. And we show that using similarity relationships in the assessment improves the precision and recall of our approach.

5.1 Experimental Settings

We selected twenty pairs of our familiar nouns as the test concept-set, which were averagely divided into four classes: human, plant, animal (except for human) and artificiality. These concepts were all basic level categories [8] (such as "plane"), which are neither too abstract (such as "vehicle") nor too concrete (such as "jet plane"). This is because the concepts on basic level are more likely to be proper for commonsense acquisition [9]. Intuitively, not arbitrary pairs of concepts are comparable. They always have properties in common, but one's property is greater or less in extent than that of the other. To satisfy this assumption, we picked out a concept-pair from the same class, such as "man" and "woman", "watermelon" and "apple", rather than "man" and "apple".

Next, we issue each concept-pair to the Google Search Engine and downloaded relevant texts the search engine returned from the Web. We call it the *Test Corpus* in the following experiments.

5.2 Results: Coverage of the Web Resource

Because an average person is assumed to possess commonsense knowledge, it is not communicated most of the time. That means commonsense knowledge is usually used implicit in the texts. However, it seems infeasible to extract knowledge implicit in texts just using the pattern-matching method. To investigate what

amounts of comparative knowledge we might extract from explicit expressions on the Web, we propose the coverage of comparative information on the web:

$$Coverage = \frac{\text{the number of comparative statements explicit on the Web}}{\text{the number of comparative statements}}$$

In fact, we don't know the accurate quantity of commonsense possessed by people or that of information distributed on the whole web. So we estimated these values based on several human subjects and the *Test Corpus*. More specifically:

$$EstCoverage = \frac{\text{the number of comparative statements in the Test Corpus}}{\text{the number of comparative statements from human subjects}}$$

We asked five human subjects to contribute their comparative commonsense on the test concept-set. The subjects' knowledge was gathered in an iterative way. Firstly, they were limited to derive this knowledge just from their brains. After the subjects submitted their outcome for the first time, we presented them the *Test Corpus*. In the second phase, the subjects manually extracted comparative commonsense from the corpus, which may provide them new clues. Meanwhile, the subjects were allowed to modify their primary submission.

Table 2. Quantity of comparative commonsense obtained from 5 human subjects and that in the test web corpus

Noun Class	Human Subjects		Test Corpus	Coverage
	primary	modified		
Human	99	128	102	79.7%
Plant	27	33	30	90.9%
Animal	57	57	31	54.4%
artificiality	50	61	58	95.1%
total	233	279	221	79.2%

Table 2 shows the quantity of the twice submissions of the subjects and knowledge automatically mined from the web corpus, respectively. We note that, the subjects supplemented 46 records according to the web resource, which were not referred in the first submission. And the average coverage achieved 79.2%, which demonstrated the Web a good resource containing adequate explicit comparative commonsense.

5.3 Results: Effectiveness of Our Approach

In this experiment, we used the feature *Freq* and *Conf* to assess potential statements as the baseline method, and used *ExtFreq* and *ExtConf* as the extended method. Precision and recall are used to evaluate the performance of these approaches.

Before our extraction, we pre-processed the related texts in the test corpus including HTML label analyzing, word segmentation and POS tagging [10]. We extracted 429 potential comparative statements using the heuristic rules (referred in Section 3.1). Then, we expressed these acquired triples in the natural language "In general, c_1 is relatively * than c_2 ", and asked the five subjects to judge whether each description was acceptable. If a potential statement was accepted by 60% of the human subjects, we consider it a piece of comparative commonsense. Finally, we assessed each potential statement using the statistical features in both baseline method and extended method, and the results were compared with that from the human judgments.

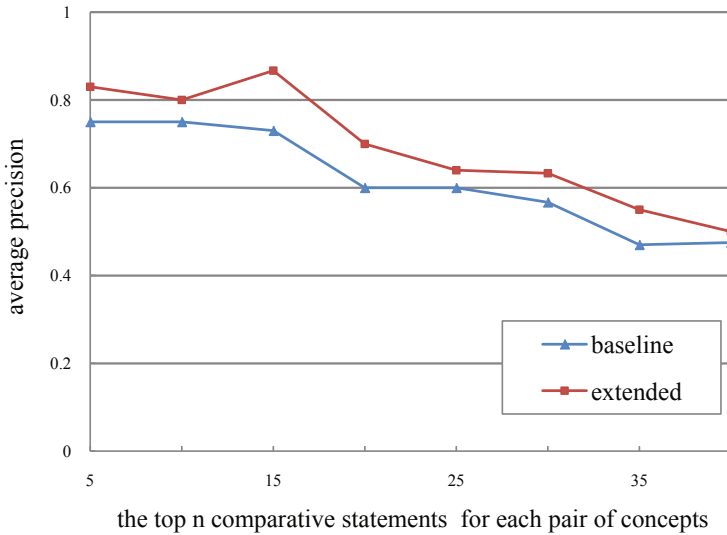


Fig. 1. Comparison of two methods

Fig.1. shows the performance of both the baseline method and extended method. Given the minimum threshold of confidence and the number of matched patterns which are 0.057 and 3 respectively, we ranked the comparative statements according to their frequencies. This graph plots the average precision and the top n acceptable statements ranked for a pair of concepts. It's obvious that the extended method outperformed the baseline scores. The average precision and recall achieved 76.4% and 83.3%, respectively.

6 Conclusion and Future Work

Comparative commonsense describes the concept-level relations from the view of comparing, which denotes that one entity has a property or quality greater or less in extent than that of another. In this paper, we propose an automatic

method for acquiring comparative commonsense from the World Wide Web. We use multiple lexico-syntactic patterns to extract potential comparative statement from related texts retrieved by the search engine. And then, we assess the candidates by combining Web-scale statistics and their semantic similarity relationships including synonym and antonym.

In the experiments, we evaluated the coverage of the explicit comparative information in the web corpus, which demonstrated the Web a good resource for our extraction. The experimental results also showed that the use of semantic similarity relationships in assessment significantly improves precision and recall of our approach when the statements have close dependency.

This work is based on given comparative pairs of concepts. In the future work, we will research on the characteristics of the comparative concepts, and automatically select them as the input data.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant No. 60773059.

References

1. Singh, P.: The Public Acquisition of Commonsense Knowledge. In: Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access (2002)
2. Schubert, L.: Can We Derive General World Knowledge from Texts. In: Proceedings of the Second International Conference on Human Language Technology (2002)
3. Torisawa, K.: An Unsupervised Learning Method for Commonsensical Inference Rules on Events. In: Proceedings of the Second CoLogNet-EISNET Symposium (2003)
4. Popescu, A.M.: Information Extraction from Unstructured Web Text. Ph.D. thesis, University of Washington (2007)
5. Cao, Y.N., Cao, C.G., Zang, L.J., Zhu, Y., Wang, S., Wang, D.S.: Acquiring Commonsense Knowledge about Properties of Concepts from Text. In: Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 4, pp. 155–159 (2008)
6. Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora. In: Proceedings of the 14th International Conference on Computational Linguistics, pp. 539–545 (1992)
7. Kosala, R., Blockeel, H.: Web mining research: a survey. SIGKDD Explorations 2, 1–15 (2000)
8. Zhang, M.: Cognitive Linguistics and Chinese Noun Phrases. China Social Science Press, Beijing (1998)
9. Zhu, Y., Zang, L.J., Cao, Y.N., Wang, D.S., Cao, C.G.: A Manual Experiment on Commonsense Knowledge Acquisition from Web Corpora. In: Proceedings of the 7th International Conference on Machine Learning and Cybernetics, Kunming (2008)
10. Zhang, H.P., Yu, H.K., Xiong, D.Y., Liu, Q.: HMM-based Chinese Lexical Analyzer ICTCLAS. In: Proceedings of the Second SIGHAN Workshop Affiliated with 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan (2003)

Detecting Temporal Pattern and Cluster Changes in Social Networks: A Study Focusing UK Cattle Movement Database

Puteri N.E. Nohuddin¹, Frans Coenen¹, Rob Christley², and Christian Setzkorn²

¹ Department of Computer Science,
University of Liverpool,
L69 3BX Liverpool
+44 (0)151 795 4275
puteri@liverpool.ac.uk,
frans@liverpool.ac.uk

² School of Veterinary Science,
University of Liverpool and National Center for Zoonosis Research,
Leahurst, Neston
+44 (0)151 794 6003
robcb@liverpool.ac.uk,
christian@setzkorn.eu

Abstract. Temporal Data Mining is directed at the identification of knowledge that has some temporal dimension. This paper reports on work conducted to identify temporal frequent patterns in social network data. The focus for the work is the cattle movement database in operation in Great Britain, which can be interpreted as a social network with additional spatial and temporal information. The paper firstly proposes a trend mining framework for identifying frequent pattern trends. Experiments using this framework demonstrate that in many cases a large number of patterns may be produced, and consequently the analysis of the end result is inhibited. To assist in the analysis of the identified trends this paper secondly proposes a trend clustering approach, founded on the concept of Self Organizing Maps (SOMs), to group similar trends and to compare such groups. A distance function is used to compare and analyze the changes in clusters with respect to time.

Keywords: Temporal Data Mining, Social Networks, Trends, Temporal Patterns and Clusters.

1 Introduction

Many data mining techniques have been introduced to identify frequent patterns in large databases. More recently, the prevalence of large time stamped databases, facilitated by advances in technology, has increased. As such, time series analysis techniques are of increasing significance. A time series, at its simplest, consists of a sequence of values associated with an attribute. The work described in this paper

considers time series to comprise several sub-series. As such, the sub-series may be compared to identify changes. For example, we can imagine a time series covering N years, where each twelve month period represents a sub-series; as such the sub-series can be compared to identify (say) seasonal changes or anomalies.

The work described in this paper is specifically directed at the comparison of sequences of time series that exist in social network data. In this respect, the time series are defined in terms of the changing frequency of occurrence of combinations of attributes that exist across particular social networks. We refer to such time series as *trends*.

Social networks are collections of interacting entities typically operating in some social setting (but not necessarily so). The nodes in a social network represent the entities and the arcs the interactions. The focus for the work described in this paper is the cattle movement database in operation within Great Britain (GB). The identification of trends in cattle movements, and changes in trends, is of particular interest to decision makers who are concerned with the potential spread of cattle disease, however the trend analysis techniques described have more general applicability.

A particular issue in trend analysis in social networks (and more generally) is the large number of trends that may be discovered. One solution, and that propose in this paper, is to cluster similar trends using Self Organizing Map (SOM) technology. The use of SOMs provides a visualization technique. Moreover, since we are interested in identifying anomalies in trends, we wish to compare individual SOMs (describing sub-series) so as to be able to observe the “dynamics” of the clustered trends.

2 Background

This section provides some brief background regarding the work described. The section is divided into three sub-sections: temporal frequent pattern mining, social networks and trend clustering and comparison.

2.1 Temporal Frequent Pattern Mining

Temporal data mining is directed at data that comprises sequence of events (Antunes *et al.* 2001). The introduction of advanced computer technologies and data storage mechanisms has afforded many organizations the opportunity to store significant amounts of temporal (and spatio-temporal) data. Consequently, there is a corresponding requirement for the application of temporal data mining techniques. The main aim of temporal data mining is to discover the relationship between non-trivial patterns or events in the temporal database (Roddick *et al.* 2002). This then allows the identification of trends or change points within the data. Many approaches have been explored in the context of temporal data mining. Two common methods are time series analysis (Brockwell *et al.* 2001) (Keogh *et al.* 2003) and sequence analysis (Zaki 2001).

In this work, trends are defined in terms of the changing frequency of frequent patterns with time. A frequent pattern, as first defined by Agrawal *et al.* (1993), is a subset of attributes that frequently co-occur in the input data according to some user specified support threshold. Since then, the frequent pattern idea has been extended in many directions. A number of authors have considered the nature of frequent patterns

with respect to the temporal dimension, for example sequential patterns (Agrawal *et al.* 1995), frequent episodes (Mannila *et al.* 1997) and emerging patterns (Dong *et al.* 1999). Many alternative frequent pattern mining algorithms, that seek to improve on Agrawal's original Apriori algorithm, have also been proposed. One example is the TFP (Total From Partial) algorithm (Coenen *et al.* 2001). The authors have adapted TFP to identify trends as defined above.

2.2 Social Network

A Social Network (SN) describes a social structure of individuals, who are connected directly or indirectly based on a common subject of interest, conflict, financial exchange or activities. A SN depicts the structure of social entities, *actors*, who are connected through ties, links or pairs (Wasserman 2006). Social Network Mining (SNM) has become a significant research area within the domain of data mining. Many SN analysis techniques have been proposed which map and measure the relationships and flows between people, organizations, groups, computers and web sites. SNM can be applied in a static context, which ignores the temporal aspects of the network; or in a dynamic context, which takes temporal aspects into consideration. In the static context, we typically wish to find patterns that exist across the network, or cluster sub-sets of the networks, or build classifiers to categorize nodes and links. In the dynamic context, we wish to identify trends or change points within networks.

2.3 Trend Clustering and Comparison

Trend mining techniques typically identify large numbers of trends. To aide the analysis of the identified trends, the technique proposed in this paper suggests the clustering of trends so that similar trends may be grouped. When this process is repeated for a sequence of sub-trends, the clusters can be compared so as to identify changes or anomalies. Other than simply comparing pairs of data sets (Denny *et al.* 2008), there are several methods that have been proposed to detect cluster changes and cluster membership migration. For example, Lingras *et al.* (2004) proposed the use of Temporal Cluster Migration Matrices (TCMM) for visualizing cluster changes in e-commerce site usage. Hido *et al.* (2008) suggested a technique to identify changes in clusters using a decision tree method. This paper proposes the use of Self Organizing Maps (SOMs).

3 Problem Definition

The work described in this paper assumes a time stamped data set D such that $D = \{D_1, D_2, \dots, D_n\}$, where n is the number of time stamps. Each data set in D comprises a set of records such that each record is a subset of some global attribute set I . A frequent pattern occurring in data set D_k is then some subset of I that occurs in given percentage of the records in D_k , this percentage is termed the support threshold (α). The number of occurrences of a given frequent pattern in D_k is termed its support (s). The sequence of support values for a given pattern can be conceptualized as time series T comprising a number of *episodes* such that $T = \{E_1, E_2, \dots, E_n\}$. The sequence of support values represented in a particular episode describes a trend comprising m time

stamps, thus $E_i = \{t_1, t_2, \dots, t_m\}$ ($0 < i \leq m$). The problem addressed in this paper is firstly the effective identification of these trends, and secondly the comparison of these trends so as to identify interesting information.

4 Frequent Trend Mining and Analysis

An overview of the proposed trend mining process is given in Figure 1. The process comprises two stages: (i) trend mining and (ii) visualization. Separate software units have been developed for each stage, in the figure; these are identified as the *trend mining unit* and the *visualization unit*. The process commences (top left) with a time stamped data set covering a sequence of N episodes. Sequences of N trends can be then identified from within the data using appropriate trend mining software (see below). The trends are stored in a compressed form in a “reverse” set enumeration tree structures (top right of Figure 1). The tree structures allow fast “look up” to extract the actual trends. Some examples are given in the figure. Thus (from the figure), the pattern {a,b,c,d} has a sequence of support values of {0,0,2500,3311,2718,0,0,0,2779} describing a nine time-stamp trend associated with a single episode, similar sequences may be extracted for all N episodes associated with the pattern {a,b,c,d}. Note that a 0 support value indicates a support value below the support threshold. The trend values are the input for visualization unit which produces several maps that cluster yearly trends which are orderly processed by the unit (bottom left of Figure 1).

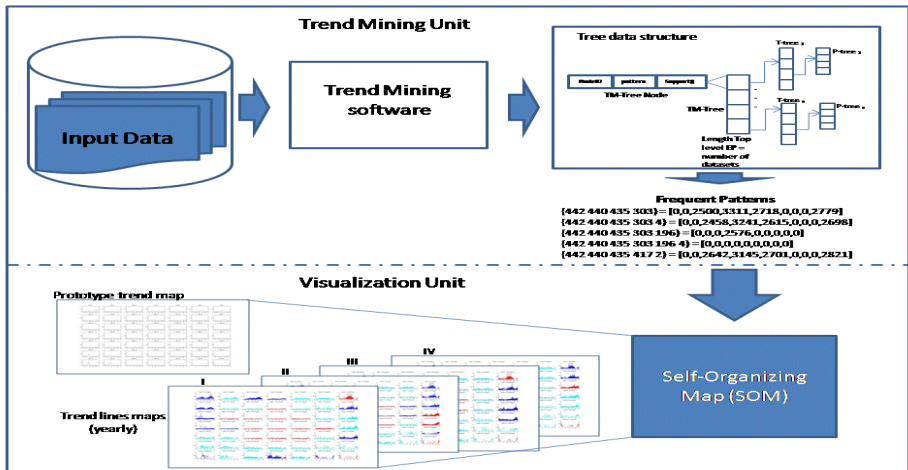


Fig. 1. Trend Mining Analysis

The trend mining unit software identifies and extracts the desired trends. The software was founded on the Total From Partial (TFP) association rule mining algorithm extended to give Trend Mining TFP (TM-TFP) so that sequences of support values could be identified.

The visualization unit is responsible for analyzing the output from TM-TFP and presenting the results. The objective is, other than clustering of the trends, to identify cluster changes. The process commenced with the clustering of the trends in each episode (so that N sets of clusters are produced). The clustering was undertaken using a Self Organizing Map (SOM) (Kohonen 1998). The process commences with the generation of a prototype map using some proportion (or all) of the trends associated with one of the episodes. The SOM map was initialized with $p \times q = N$ nodes such that each node represented a "type" (category) of trend line. Once the prototype map had been generated, the trends associated with each of the episodes were fitted to this map so as to give a sequence of "trend line" maps. In the figure, four episodes are assumed (labeled I, II, III and IV). Once the trend line maps have been derived the change in trends was determined by considering how individual trends, associated with particular frequent patterns, moved (or did not move) across the sequence of maps. A simple Euclidean distance measure was used for this purpose. The maximum change is the diagonal distance across the map.

5 Experimental Evaluation

This section presents and discusses sample results obtained using the proposed trend mining and analysis process. For the experiments, the Cattle Tracing System (CTS) database in operation in Great Britain (GB) was used. The CTS database records, for monitoring purposes, all the movements of cattle registered within or imported into GB. The database is maintained by the Department for Environment, Food and Rural Affairs (DEFRA). Cattle movements can be one off movements to final destinations, or movements between intermediate locations. Movement types include: (i) cattle imports, (ii) movements between locations, (iii) movements in terms of births and (iv) movements in terms of deaths. CTS was introduced in September 1998, and updated in 2001 to support disease control activities. Currently the CTS database holds some 155 Gb of data. The CTS database can be interpreted as a social network where the nodes represent cattle holding areas and the arcs between nodes cattle movements.

The CTS database comprises a number of tables, the most significant of which are the animal, location and movement tables. For the experiments reported here the data from 2003 to 2006 was extracted to form 4 episodes each comprising 12 (one month time stamps). The data was stored in a single data warehouse such that each record represented a single cattle movement instance associated with a particular year (episode) and month (time stamp). The number of CTS records represented in each data episode was about 400,000. Each record in the warehouse comprised: (i) a time stamp (month and year), (ii) the number of cattle moved, (iii) the breed, (iv) the sender's location in terms of easting and northing grid values, (v) the "type" of the sender's location, (vi) the receiver's location in terms of easting and northing grid values, and (vii) the "type" of the receiver's location. If two different breeds of cattle were moved at the same time from the same sender location to the same receiver location this

would generate two records in the warehouse. The maximum number of cattle moved between any pair of locations for a single time stamp was approximately 40 animals.

5.1 Frequent Patterns and Trends

Table 1 presents some statistics indicating the number of trends discovered in the CTS data warehouse using TM-TFP. Recall that TM-TFP was used to identify frequent patterns and their associated support values over a sequences of time stamps. The results presented in Table 1 were generated using support thresholds of 0.5%, 0.8% and 1% respectively. Each row in Table 1 represents the number of trends identified for each of the 4 episodes (12 time stamps per episode). The lower the support threshold the greater the number of discovered frequent patterns and hence the greater the number of trends. However, use of a low support threshold ensures that no potentially interesting trends are omitted. The results presented in Table 1 indicate that a large number of trends can be identified, in a realistically sized dataset, when low support thresholds are used. It is also interesting to note that the variation between years is relatively small.

Table 1. Number of trend lines identified using TM-TFP

Year	Support Threshold		
	0.5%	0.8%	1%
2003	63,117	34,858	25,738
2004	66,870	36,489	27,055
2005	65,154	35,626	25,954
2006	62,713	33,795	24,740

5.2 Temporal Clustering

Figure 2 depicts prototype trend map trained using the 2003 data. With reference to the figure, node 1 (top-left) represents trend lines that have for patterns with high support in spring (March to May) and autumn (September to November) while node 43 (bottom-left) indicates trend lines with high support in spring only (March to April). Note that the distance between nodes indicates the dissimilarity between nodes; the greatest dissimilarity is thus between nodes at opposite ends of the diagonals. Once the initial proto-type map has been generated a sequence of trend line maps can be produced, one for each episode. An example is given in Figure 3 for the 2003 cattle movement data of the trend line maps which have been produced. These trend line maps are referring to 2003 prototype map for the same cluster structure. Each node has been annotated with the number of trends in the “clusters”. Thus, from Figure 3, there are 1970 trend lines in node 1. The shading used in Figure 3 indicates the number of trend lines in each node, the darker the shading, the greater the number of trends.

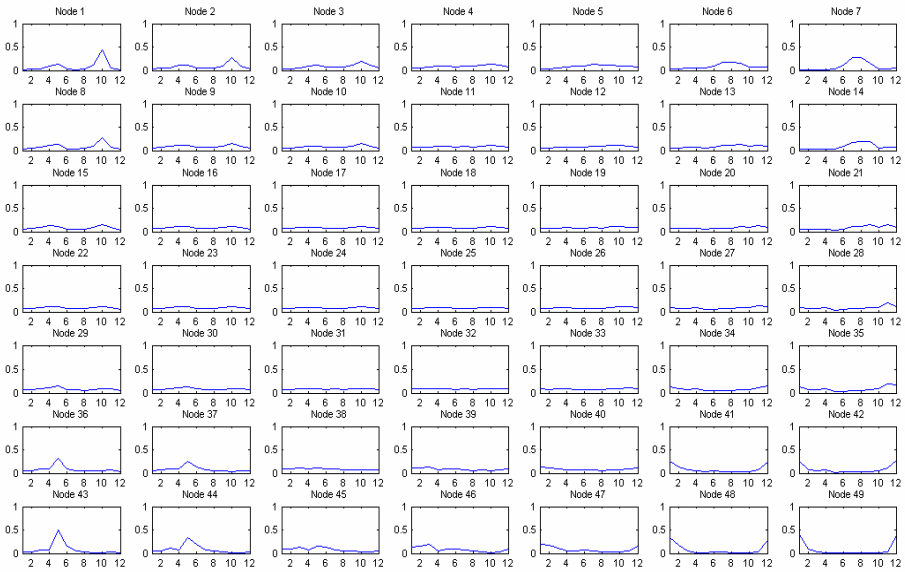


Fig. 2. Prototype map

5.3 Temporal Cluster Changes

Figure 4 indicates the number of trends in each node (cluster) for the 4 years (episodes) included in the described study. From Figure 4, the greatest differences are observed for nodes 23 and 31. Whatever the case, from the figure, it is clear that the number of trends per node is not static. Given a sequence of trend-line maps comparisons can be made to see how trends associated with individual frequent patterns change by analyzing the nodes in which they appear. Some trends may remain within the same node for the entire sequence of episodes. Some other trends may oscillate between nodes, while some further trends may slowly migrate across the map. By translating the trend line maps into a rectangular (D-plane) set of coordinates a Euclidean distance function was applied to observe the similarities and differences of trends within each node across the episodes. By comparing the values produced by the distance function, the degree of movement could be determined. This could be interpreted in a number of ways, for example the greater the distance moved the more interesting the change may be deemed to be.

Table 2 shows examples of how some trends (representing frequent patterns) migrate from one cluster to another. For example, the trend line representing the pattern {441 436 329 301 213 4 3} which translates as: {numberAnimalsMoved <=5, Receiver PTI = NULL, Receiver Location Type = Calf Collection Centre, Sender Location Type = Agricultural Holding, Sender area = 14, Animal age <= 1 year old, Gender = female} was in node 49 (bottom right in Figure 2) in 2003 and 2004, but then migrated to node 48 in 2005 and disappeared in 2006. Table 3 gives some further

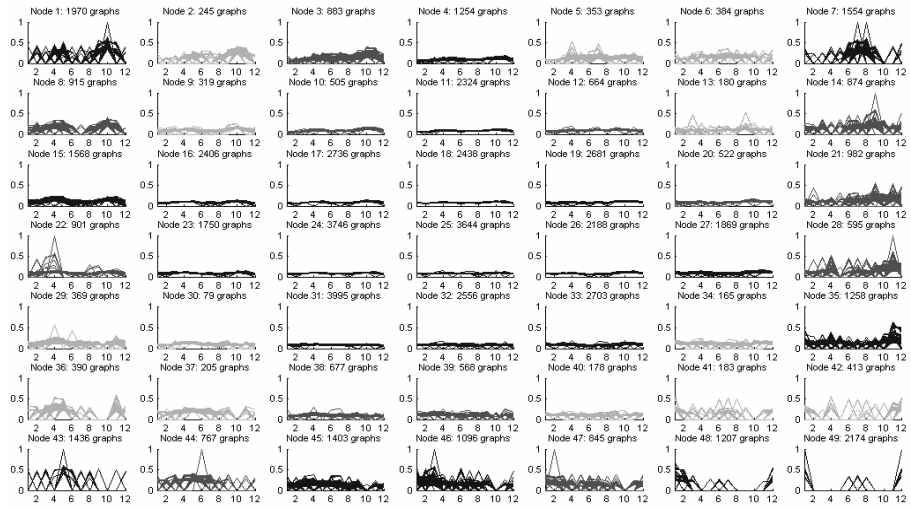


Fig. 3. Trend line map for 2003

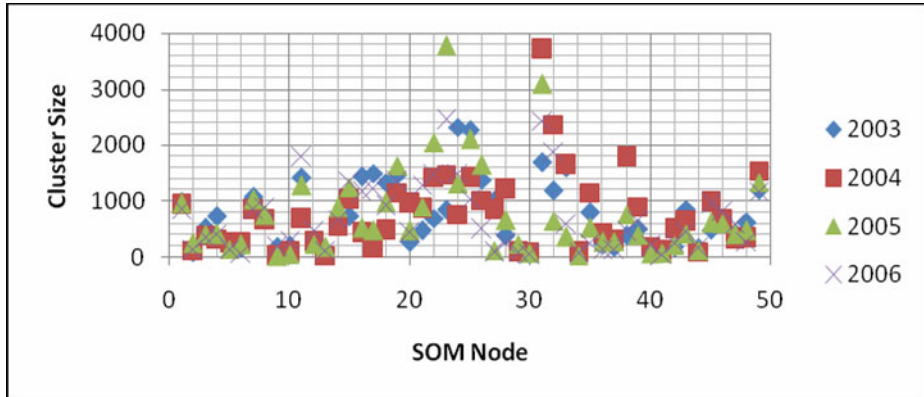


Fig. 4. Comparison cluster size between 2003 and 2006

statistics regarding the movement of trends in the context of the CTS database. There are 79894 distinct frequent patterns generated between 2003 and 2006 data episodes. But only 4193 patterns that remain in the same nodes across the years whereas the rest of the patterns moved to different cluster nodes. Table 3 also shows statistics of spatio-temporal patterns between 2003 and 2006.

Table 2. Example of frequent patterns that migrated to other clusters

Frequent Patterns	Node 2003	Dist	Node 2004	Dist	Node 2005	Dist	Node 2006
{441 436 329 301 213 4 3}	49	0	49	1	48	0	0
{441 436 329 301 213 196}	48	1	49	4.1	38	3.2	48
{378 301 263}	39	0	39	3.2	49	3.2	39
{378 301 263 4}	46	0	46	0	0	0	46
{378 301 263 196}	47	1	46	0	0	0	49
{441 318 301 212 4}	14	0	14	5	16	5.4	7
{441 329 214}	47	2	49	0	0	0	49

Table 3. Clusters memberships

Number of Patterns	Quantity
Distinct frequent patterns described by trend lines between 2003 and 2006	79894
Trends stayed in the same node (unchanged) between 2003 and 2006	4193
Trends migrated to other clusters between 2003 and 2006	75701
Spatio-temporal trends stayed in the same cluster between 2003 and 2006	637
Spatio-temporal trends migrated to other clusters with greater distance values (distance>4) between 2003 and 2006	2061

6 Conclusions

This paper has described a trend mining framework, TM-TFP that successfully identifies trends in large social networks. The framework is supported by a SOM technique that provides a powerful mechanism for grouping similar trends, and a trend migration identification mechanism to show changes in the nature of individual trends associated with frequent patterns. The mechanism has been tested and evaluated using data from GB’s cattle tracking database. The research team is currently looking at other ways in which change detection can be made more effective in the context of decision makers and stakeholders.

References

Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of ACM SIGMOD Conference (1993)

Agrawal, R., Srikant, R.: Mining sequential patterns. In: 11th International Conference on Data Engineering (1995)

Antunes, C.M., Oliveira, A.L.: Temporal Data Mining: An Overview. In: Proc. ACM SIGKDD Workshop Data Mining, August 2001, pp. 1–13 (August 2001)

Brockwell, P., Davis, R.: Time Series: Theory and Methods. Springer, Heidelberg (2001)

Coenen, F.P., Goulbourne, G., Leng, P.: Computing Association Rules Using Partial Totals. In: Siebes, A., De Raedt, L. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 54–66. Springer, Heidelberg (2001)

Denny, Williams, G.J., Christen, P.: reDSOM: relative density visualization of temporal changes in cluster structures using self-organizing maps. In: IEEE International Conference on Data Mining (ICDM), pp. 173–182. IEEE Computer Society, Los Alamitos (2008)

- Dong, G., Li, J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: Proceeding of Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (1999)
- Hido, S., Idé, T., Kashima, H., Kubo, H., Matsuzawa, H.: Unsupervised changes analysis using supervised learning. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 148–159. Springer, Heidelberg (2008)
- Keogh, E., Kasetty, S.: On the need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration. *Data Mining and Knowledge Discovery* 7(4), 349–371 (2003)
- Kohonen, T.: The Self Organizing Maps. *Neurocomputing* 21, 1–6 (1998)
- Lingras, P., Hogo, M., Snorek, M.: Temporal Cluster Migration Matrices for Web Usage Mining. In: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (2004)
- Mannila, H., Toivonen, H., Verkamo, A.: Discovery of Frequent Episodes in Event Sequences. *Data Mining and Knowledge Discovery* 1, 259–289 (1997)
- Roddick, J., Spiliopoulou, M.: A Survey of Temporal Knowledge Discovery Paradigms and Methods. *IEEE Trans. Knowledge and Data Eng.* 14(4), 750–767 (2002)
- Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, New York (2006)
- Zaki, M.: SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning* 42(1-2), 31–60 (2001)

Unstructured P2P-Enabled Service Discovery in the Cloud Environment

Jing Zhou^{1,2} and Zhongzhi Shi²

¹ Communication University of China, Beijing, 100025, China

² The Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing,
100190, China
zhoujing@cuc.edu.cn

Abstract. As the Cloud computing appears to be part of the mainstream computing in a few years, the number of the services it provides, its users, and the requests of these services will be on the rise accordingly. To deliver satisfactory experience to Cloud users, it is essential that highly scalable techniques for service discovery should be available. We embarked on a preliminary study, in Cloud environments, on service discovery by adopting an unstructured P2P technique. In the context of Cloud computing, we start by examining proposed and deployed solutions to service discovery, discuss the methodology of developing efficient mechanisms for service description, description indexing, and query routing, and finally identify open issues.

1 Introduction

Armbrust *et al.* defined in [1] that Cloud computing comprises the applications delivered as services over the Internet and the hardware and system software in the datacenters that offer those services. Typically, Cloud computing comes in three kinds: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [2]. The classification emphasizes the core concept of “X as a Service”, that is, software, platform, and infrastructure all can be provided to end users on demand and on a subscription basis.

Cloud services are currently available from providers such as Amazon, Microsoft, Google App Engine, Eucalyptus, and GoGrid. Users can access Cloud services by visiting service provider’s website, establishing a runtime environment in response to instructions, creating a user account, and configuring related tools. There is no complicated mechanism for service discovery involved. As the Cloud computing will soon become part of the mainstream computing in a few years, the number of the services it provides, its users, and the requests of these services will increase in orders of magnitude. To deliver satisfactory experience to Cloud users, highly scalable techniques for service discovery should be available.

We observed in distributed systems that service discovery mechanisms are primarily centralized in which a central index server (or several such servers) stores and maintains all information about services being offered. Centralized

service discovery is prone to the following issues: 1) single point of failure; 2) lack of satisfactory scalability; 3) requirements for powerful computing capabilities to serve large amounts of service discovery and update queries on the central server; and 4) performance bottlenecks and network congestion. To overcome these obstacles so as to deliver a Cloud service discovery mechanism of high performance, we will rely on peer-to-peer (or P2P for short) computing [3].

The popularity of P2P computing stems from its self-organization, fault tolerance, and scalability, which make P2P computing an obvious candidate for tackling large scale service discovery in the Cloud environment. Studies [4] showed that satisfactory scalability could be achieved if P2P techniques based on DHTs (distributed hash tables) were applied to service discovery in Cloud computing. We, however, argue that such structured P2P techniques place severe constraints on the network topology and the placement of services (or their indices), which makes structured P2P unable to better model the real world.

We aim to address the issue of service discovery in the Cloud computing by exploring the Semantic Web technology to describe Cloud services and to facilitate service discovery and service matching, and investigating unstructured P2P-based techniques that support highly efficient and scalable routing mechanisms for service requests.

The balance of the paper is structured as follows. Related work is reviewed in Sect. 2. We present and discuss the primary design issues in Sect. 3 in greater details. Finally, we conclude the paper by identifying open issues in Sect. 4.

2 Related Work

2.1 Unstructured P2P and Structured P2P

Unstructured P2P systems including Gnutella, Freenet, and FastTrack carry out object lookup and downloading operations in the absence of a central index server. Each peer maintains indices for the resources it currently holds. A lookup operation in such systems will not necessarily be successful and there is no upper and lower bounds on the successful operation, that is, non-deterministic query performance is offered. In addition, unstructured P2P systems support one-dimensional and multi-dimensional point queries and range queries.

Similarly, there is no central index server in structured P2P built upon DHTs (such as CAN, Chord, Pastry, and Tapestry). Each peer maintains the indices of data items on another $O(\log n)$ peers where n is the number of peers in the system. The keys of the data items and nodes are mapped onto the overlay network in which each node is responsible for managing a small number of data items. Whenever a peer receives a lookup request, it is able to locate the data item of interest within $O(\log n)$ hops. Therefore, the query performance delivered by structured P2P is deterministic. One-dimensional point queries can be well supported but, in order to implement one-dimensional range queries and multi-dimensional point queries and range queries, the system needs to employ spatial indices [5] such as the Space Filling Curve, kd-trees and MX-CIF quadtree, and carry out sufficient extension to the basic DHT mechanism [6].

Note that each peer in structured P2P maintains the indices of data items of others (but not the data items of its own!) and reliance among peers exists. However, each Cloud service provider maintains its own services, describing the services, creating indices according to service description, and publishing service description to facilitate service discovery. Furthermore, service discovery in Cloud computing will regularly be performed based on *multiple* (instead of single) attributes of the service in most cases. We believe that the search mechanism enabled by unstructured P2P techniques can better resolve multi-dimensional queries by incorporating support for semantics, whereas the basic DHT technique delivers satisfactory performance in keyword search.

2.2 P2P-Based Service Discovery

A robust, automatic, and reliable mechanism for service discovery is indispensable to distributed systems such as Grid computing, Web services, and pervasive computing because it helps users to locate services or resources required in such systems [7,8,9,10]. To overcome the drawbacks and inefficiencies of centralized service discovery in terms of scalability, fault-tolerance, and network congestion, a number of fully decentralized solutions were proposed. Among others, the approach that adopts the P2P paradigm draws a great deal of attention (see [11] for details on P2P-based resource¹ discovery in computational grids).

The first step towards implementing a P2P-based mechanism for service discovery is to design a decentralized index system in which the appropriate data structure is selected for building indices upon the service description. Since a service is typically characterized by both static and dynamic attributes, users can specify the requirement for multiple attributes in a query, that is, a multi-dimensional query is allowed. In unstructured P2P, each peer maintains its own services and service descriptions, hence simple indices, such as the two dimensional table, can be employed to facilitate resolving multi-dimensional queries.

The routing mechanism for multi-dimensional queries in the P2P paradigm should also be developed, which helps forward service requests to their destination peers efficiently. Flooding and its variants [12,13] are the widely used approaches to message routing in unstructured P2P. The main drawback of such techniques includes that excessive unnecessary query messages and traffic are generated. Various heuristics [14,15] were therefore developed to guide message routing as well as to increase system scalability.

When a query is eventually delivered to a potential destination peer, a matching scheme is needed to confirm a perfect match between the service description and the user requirement expressed in the query message. Currently, keyword-based methods widely used in structured P2P, due to their unsatisfactory search performance, have been gradually replaced by semantics-based solutions. This topic is however not the focus of our paper and we leave it to the future work.

¹ The “resource” here may refer to clusters, supercomputers, and desktop computers and this concept has an overlap with that of the “service” in Cloud computing.

2.3 Service Description and Discovery Using Semantics

When describing a service, elements such as the service properties, capabilities, and constraints should be taken into account. To service the request, the Semantic Web offers a number of powerful tools. The integration of the Semantic Web technologies is beneficial to the realization of efficient service discovery due to the following reasons: 1) Ontology can be used to describe concepts and the relationships among the concepts within a specific domain in a disambiguous way and hence, people often employ ontology to describe services and encode the description semantically; 2) Languages such as OWL can be used to describe ontologies in order to support semantic inference for relationships among various concepts; 3) A number of tools in the Semantic Web community have been developed for service (Web Services in particular) description purposes, including OWL-S [16], WSMO (Web Services Modeling Ontology) [17] and Web Service Semantics - WSDL-S [18]; and 4) The semantic service description with powerful expresiveness is necessary to service matching based on semantics.

Semantics-based service matching comes in two forms: signature matching [19] and specification matching [20]. In signature matching, the subsumption relationship between concepts that are defined in ontologies and are used to describe the input capabilities (that is, the required services) and output capabilities (that is, the available service descriptions), is identified. Specification matching is, however, performed between the pre-conditions and post-conditions of the functional semantics of software components using automated theorem proving or query containment. Semantic matching has been applied to a number of distributed systems, including Grid computing [21,22], P2P computing [23,24], Web Services [25], and pervasive computing [10].

To address the issue of service matching in the Cloud environment in particular, Zeng *et al.* proposed a matching algorithm which extends the keywords that describe the input and output capabilities of services by using WordNet [26]. A function is employed to evaluate the semantic similarity between the concept sets (consisting of concepts extended from the keywords) of any two services. The primary drawback of the approach is lack of support for range queries.

3 System Design

3.1 System Overview

Our proposed system consists of Cloud service providers (peers in P2P terminology) that describe their own services, create indices according to service descriptions, and publishing service descriptions for sharing purposes. Each new peer or node, upon arrival, will exchange its service descriptions with others (and maintain the information in its local index) that are already in the network, that is, 1-step replication [27] is utilized, thus forming a neighboring relationship with those nodes. When a peer voluntarily leaves the system, or is lost due to topology re-organization, the index information for that neighbor gets flushed.

According to our previous work on unstructured P2P, flooding and random walk search techniques can hardly offer scalability comparable to that of DHT-based approaches. One efficient solution to the problem is to introduce a certain degree of centralized control in the P2P network, that is, supernodes [14] [27]. We therefore allow a few number of peers with high capacity to act as supernodes as needed. For instance, when a peer finds its knowledge about services provided by other peers is rich enough², it can elect itself to be a supernode by informing all neighbors of the decision. If the latter agree to accept the peer as a supernode, they will update their indices to reflect the change. The primary motive underlying the introduction of supernodes is to bias query messages towards nodes that provide a shortcut to the destination peers with a high probability.

Moreover, to avoid overburdening nodes with query messages from neighbors, we allow nodes to add links to more useful neighbors and drops links to useless ones, thus leading to reorganization of the P2P overlay topology. Topology reorganization is often used by P2P systems to facilitate resource/service discovery by shortening the distance (in the form of hops) between the potential destination peers and the resource/service requester based on predictions of future requests.

We elaborate on issues comprising service description, description indexing, and routing of service requests in the following sections.

3.2 Service Description

To describe Cloud services, we start by using WSDL-S—currently a W3C member submission that offers a lightweight solution to creation of semantic service description. In WSDL-S, the expressivity of WSDL was augmented with semantics by adopting concepts similar to those in OWL-S. We may gradually extend and enhance WSDL-S by means of the extensibility provided by WSDL itself according to the specific requirements for Cloud service descriptions.

In a nutshell, WSDL-S provides a few extensibility elements to realize the URI reference mechanism as follows [18].

wssem:modelReference specifies the association between a WSDL entity and a concept in a semantic model.

wssem:schemaMapping handles the structural difference between the schema elements of a Web service and their corresponding concepts in a semantic model.

wssem:precondition and **wssem:effect** are specified as child elements of the element *operation* and describe the semantics of the operation.

wssem:serviceCategorization comprises service categorization information that could be used when publishing a service in a Web Services registry.

A Cloud service can be semantically annotated by borrowing all these constructs from WSDL-S. Figure 1 presents an example Cloud service description. Note

² In related work such as [14], supernodes refer to peers with higher bandwidth connectivity, whereas in our system in which service description is encoded with semantics (see Sect. 3.2) and we argue that nodes with richer knowledge should be able to better serve service requests and thus be elected as supernodes.

that a precondition is a set of statements that should be true before a service can be successfully invoked. Hence, we can describe part of a user’s requirement for a service by means of preconditions. For instance, if a user is looking for a disk storage service, she can send out a “DiskStorage” request and specify the capacity should be 200GB and the transfer rate of the disk should be 50Mb/s. However, in proposal [18] at most one precondition (as well as one effect) is allowed so multiple preconditions should be captured into one high level precondition. We can combine those two statements via “AND” into one precondition since **WSSemantics.xsd** defines an attribute “expression” of type “string” for **wssem:precondition**.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<definitions name="DiskStorage"
  targetNamespace="http://www.example.com/wsd1-s/examples/diskstorage.wsd1"
  xmlns="http://www.w3.org/2004/08/wsd1"
  .....
```

```
  xmlns:wssem="http://www.example.com/wsd1-s/examples/diskStorage.wsd1"
  xmlns:DSOntology="http://www.example.com/wsd1-s/ontologies/DiskStorage.owl"
  <types>
    .....
```

```
    <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
      targetNamespace="http://www.example.com/wsd1-s/examples/diskStorage.wsd1"
      xmlns="http://www.example.com/wsd1-s/examples/diskStorage.wsd1" >
      <xs:complexType name="processDiskStorageRequest" >
        <xs:all>
          <xs:element name="billingInfo" type="xsd1:DSBilling" />
          <xs:element name="storageItem" type="xsd1:DSItem" />
        </xs:all>
      </xs:complexType>
      <!-- Semantic annotation is added directly to leaf element -->
      <xs:element name="processDiskStorageResponse" type="xs:string"
        wssem:modelReference="DSOntology#StorageConfirmation" />
      </xs:schema>
    </types>
    <interface name="DiskStorage" >
      <operation name="processDiskStorage" pattern="wsdl:in-out"
        wssem:modelReference="DSOntology:RequestDiskStorage" >
        <input messageLabel="processDiskStorageRequest"
          element="tns:processDiskStorageRequest" />
        <output messageLabel="processDiskStorageResponse"
          element="processDiskStorageResponse" />
        <wssem:precondition name="AvailableStoragePrecond"
          wssem:modelReference="DSOntology#StorageAvailable"
          expression="DSOntology#Capacity=200GB&&
            DSOntology#Transferrate=50Mb/s" />
        <wssem:effect name="StorageOccupiedEffect"
          wssem:modelReference="DSOntology#StorageOccupied" />
      </operation>
    </interface>
  </definitions>
```

Fig. 1. An example Cloud service description

3.3 Description Indexing

Among all the constructs used to describe a Cloud service, the precondition is the most important resource upon which we can build indices for all the services in a P2P network. Suppose a user issues a Cloud service request and we can then compile part of the request (which is specified in **wssem:precondition**) as $(attri_1, value_1)$, $(attri_2, value_2)$, \dots , $(attri_i, value_i)$.

As such, a user request can be converted to a multi-dimensional query by which the user specifies multiple conditions should be met on several attributes. We adapted one of the database approaches—the kd-tree—to establish a data space for all service descriptions. Once the data space is constructed, we set out to develop the routing strategy (in the following section) by which a multi-dimensional query can be efficiently forwarded to the relevant peers in such a space.

The kd-tree is a data structure that decomposes a multi-dimensional data space into hyperrectangles and each node corresponds to a hyperrectangle. The fields of a kd-tree node include the splitting dimension number, splitting value, left kd-tree, and right kd-tree. According to the first two fields, each node splits the space into two subspaces. Searching for a point in the dataset represented in a kd-tree can be carried out in a traversal of the tree from root to leaf ($O(\log n)$ if there are n data points). In the following, we demonstrate a kd-tree representation of 4 points (200GB, 50Mb/s), (300GB, 80Mb/s), (600GB, 30Mb/s), and (800GB, 90Mb/s). For simplicity, a 2-d space is used for illustration.

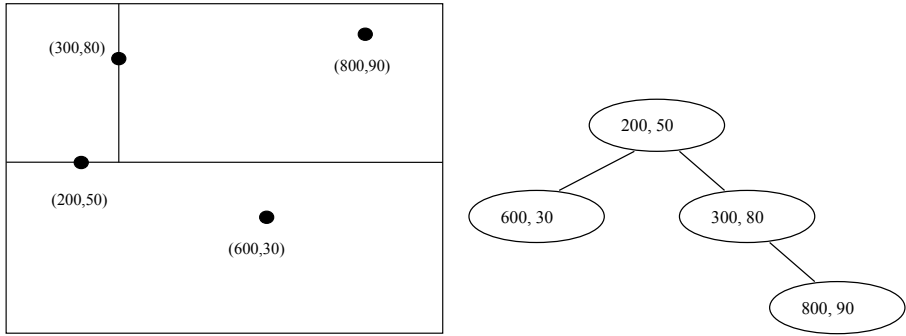


Fig. 2. A k-d tree example

In Fig. 2, the root node (200GB, 50Mb/s) splits the space in the y -axis into two subspaces. The point (600GB, 30Mb/s) lies in the lower space, that is, $\{(x, y) | y < 50\}$ and hence is in the subtree of the root node. By traversing such a kd-tree, the hyperrectangle leaf that potentially contains the target point can be located.

In our system, each new node locally creates an index in the form of a kd-tree for all the service descriptions it is aware of. The index will gradually grow and evolve as nodes join and depart from the system. The description indices are therefore distributed among individual nodes and will be used for routing service requests.

3.4 Routing of Service Requests

Part of a user's service request (encoded in **wssm:precondition**) is used to formulate a multi-dimensional query. To execute the query over a P2P network, it must be routed to the set of nodes that contain data relevant to the query. We describe routing as follows.

1. According to the content of the query, the node issuing the query will first traverse its kd-tree to locate the set of potential target nodes and subsequently forward the query to those nodes. If no such nodes are found, the

node simply sends the query to its associated supernodes, if any. Otherwise, it will flood the query message to all neighbors.

2. Upon receipt of an incoming query, the receiving node checks its local index to find the target service that might satisfy the user query³ and forwards the query to the potential targets. If no such nodes are found, the node routes the query to its associated supernodes, if any. Otherwise, it will flood the query message to all neighbors if the message was not flooded in its last hop. A flooded message, in this case, will be sent to a randomly chosen neighbor.

This process is repeated until: 1) there are no other potential targets to which a query can be further routed, 2) the TTL (Time-To-Live) in the query message is decreased to zero, or 3) the service requester explicitly claims that enough results have been collected.

3.5 Discussions

- **Decentralized indices:** By locally creating a kd-tree to index all service descriptions that a peer node is aware of, we obtain a fully distributed index (consists of all the local indices) for service discovery in the proposed P2P system. Neighbors simply delete index information about services of a leaving node from their kd-tree, whereas in a DHT-based structured P2P system, node departure incurs much more operations.
- **Support for point and range queries:** Thanks to the use of the kd-tree, both a point query and a range query can be similarly handled in our system. This is in contrast to other approaches in which it is easy to resolve the point query but sufficient extensions should be made to the basic mechanism (DHTs for instance) if range queries are also supported.
- **Localized scheme:** We proposed a localized scheme when constructing the kd-tree index for each peer node, that is, no global knowledge is required. However, in [28] when “skip pointers” (shortcuts used for optimized routing based on kd-trees) are built, the set of all nodes should be known in advance.

4 Future Work

In our preliminary work on unstructured P2P-enabled service discovery in Cloud computing, much still is left to be investigated. For instance, we have yet to experimentally demonstrate the efficiency and the scalability of the routing algorithm. Moreover, developing matching and sorting algorithms for semantics-based service matching is also indispensable for realizing effective and efficient service discovery in the context of Cloud computing. We are currently exploring these issues and will report our ongoing work in a forthcoming paper soon.

³ A perfect match is only confirmed after the semantics-based service matching has been carried out.

Acknowledgment

This work is funded by 382 Research Foundation for Talented Scholars (No. G08382320), the Engineering Disciplines Planning Project (No. XNG0921), and the Leading Academic Discipline Program (3rd phase of 211 Project for the Communication University of China). We also acknowledge the input of the National Natural Science Foundation of China (No. 60775035, No.60933004, No.60970088), the National High-Tech Research and Development Plan of China (No. 2007AA01Z132), National Basic Research Priorities Programme (No. 2007CB311004) and National Science and Technology Support Plan (No.2006BAC08B06), Dean Foundation of Graduate University of Chinese Academy of Sciences(O85101JM03).

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M.: Above the clouds: A berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, Department of Electrical Engineering and Computer Sciences, University of California at Berkeley (2009)
2. Buyya, R., Pandey, S., Vecchiola, C.: Cloudbus toolkit for market-oriented cloud computing. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) *CloudCom 2009*. LNCS, vol. 5931, pp. 24–44. Springer, Heidelberg (2009)
3. Clark, D.: Face-to-face with peer-to-peer networking. *Computer Journal* 34(1), 18–21 (2001)
4. Ranjan, R., Zhao, L., Wu, X., Liu, A.: Peer-to-peer cloud provisioning: Service discovery and load-balancing. The Computing Research Repository abs/0912.1905 (2009)
5. Samet, H.: *The Design and Analysis of Spatial Data Structure*. Addison-Wesley Publishing Company, Reading (1990)
6. Cai, M., Frank, M., Chen, J., Szekely, P.: Maan: A multi-attribute addressable network for grid information services. *Journal of Grid Computing* 2, 3–14 (2004)
7. Toma, I., Iqbal, K., Roman, D., Strang, T., Fensel, D., Sapkota, B., Moran, M., Gomez, J.M.: Discovery in grid and web services environments: A survey and evaluation. *Multiagent and Grid Systems* 3(3), 341–352 (2007)
8. Bachlechner, D., Siorpaes, K., Lausen, H., Fensel, D.: Web service discovery - a reality check. In: *Demos and Posters Session of the 3rd European Semantic Web Conference*, Budva, Montenegro (June 2006)
9. Zhu, F., Mutka, M.W., Ni, L.M.: Service discovery in pervasive computing environments. *IEEE Pervasive Computing* 4(4), 81–90
10. Mokhtar, S.B., Preuveneers, D., Georgantas, N., Issarny, V., Berbers, Y.: Easy: Efficient semantic service discovery in pervasive computing environments with qos and context support. *Journal of Systems and Software* 81(5), 785–808
11. Ranjan, R., Harwood, A., Buyya, R.: Peer-to-peer-based resource discovery in global grids: A tutorial. *IEEE Communications Surveys and Tutorials* 10(1-4), 6–33 (2008)
12. Lv, Q., Cao, P., Cohen, E., Li, K., Shenker, S.: Search and replication in unstructured peer-to-peer systems. In: *Proceedings of the 16th International Conference on Supercomputing*, New York, USA, pp. 84–95 (2002)

13. Iamnitchi, A., Foster, I., Nurmi, D.C.: A peer-to-peer approach to resource location in grid environments. In: Proceedings of the 11th IEEE International Symposium on High Performance Distributed Computing, p. 419 (2002)
14. Chawathe, Y., Ratnasamy, S., Breslau, L., Lanham, N., Shenker, S.: Making gnutella-like p2p systems scalable. In: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Karlsruhe, Germany, pp. 407–418 (2003)
15. Zhou, J., Hall, W., Roure, D.C.D., Dialani, V.K.: Supporting ad-hoc resource sharing on the web: A peer-to-peer approach to hypermedia link services. *ACM Transactions on Internet Technology* 7(2) (May 2007)
16. Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., Narayanan, S., Paolucci, M., Parsia, B., Payne, T., Sirin, E., Srinivasan, N., Sycara, K.: Owl-s: Semantic markup for web services (2010), <http://www.daml.org/services/owl-s/1.2/>
17. Roman, D., Keller, U., Lausen, H., de Bruijn, J., Lara, R., Stollberg, M., Polleres, A., Feier, C., Bussler, C., Fensel, D.: Web service modeling ontology. *Applied Ontology* 1(1), 77–106 (2005)
18. Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M.T., Sheth, A., Verma, K.: Web service semantics - wsdl-s. W3C Member Submission (2005), <http://www.w3.org/Submission/WSDL-S/>
19. Zaremski, A.M., Wing, J.M.: Signature matching: a tool for using software libraries. *ACM Transactions on Software Engineering and Methodology* 4(2), 146–170 (1995)
20. Zaremski, A.M., Wing, J.M.: Specification matching of software components. *ACM Transactions on Software Engineering and Methodology* 6(4), 333–369 (1997)
21. Harth, A., Decker, S., He, Y., Tangmunarunkit, H., Kesselman, C.: A semantic matchmaker service on the grid. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, New York, NY, USA, pp. 326–327
22. Ludwig, S.A., Reyhani, S.M.S.: Introduction of semantic matchmaking to grid computing. *Journal of Parallel and Distributed Computing* 65(12), 1533–1541
23. Zhou, G., Yu, J., Chen, R., Zhang, H.: Scalable web service discovery on p2p overlay network. In: Proceedings of the IEEE International Conference on Services Computing (SCC 2007), Salt Lake City, Utah, USA, pp. 122–129 (2007)
24. Li, Y., Zou, F., Wu, Z., Ma, F.: Pwsd: A scalable web service discovery architecture based on peer-to-peer overlay network. In: Proceedings of the 6th Asia-Pacific Web Conference, pp. 291–300 (2004)
25. Paolucci, M., Kawamura, T., Payne, T.R., Sycara, K.P.: Semantic matching of web services capabilities. In: Horrocks, I., Hendler, J. (eds.) *ISWC 2002*. LNCS, vol. 2342, pp. 333–347. Springer, Heidelberg (2002)
26. Zeng, C., Guo, X., Ou, W., Han, D.: Cloud computing service composition and search based on semantic. In: Proceedings of the 1st International Conference on Cloud Computing, pp. 290–300 (2009)
27. Gkantsidis, C., Mihail, M., Saberi, A.: Hybrid search schemes for unstructured peer-to-peer networks. In: Proceedings of IEEE INFOCOM, pp. 1526–1537 (2005)
28. Ganesan, P., Yang, B., Garcia-Molina, H.: One torus to rule them all: multi-dimensional queries in p2p systems. In: Proceedings of the 7th International Workshop on the Web and Databases, Paris, France, pp. 19–24 (2004)

Using Global Statistics to Rank Retrieval Systems without Relevance Judgments

Zhiwei Shi¹, Bin Wang¹, Peng Li¹, and Zhongzhi Shi²

¹ Information Retrieval Group, Center for Advanced Computing Research,
Institute of Computing Technology, CAS, Beijing, 100190, China
{shizhiwei, wangbin, lipeng01}@ict.ac.cn

² Key Lab of Intelligent Information Processing,
Institute of Computing Technology, CAS, Beijing, 100190, China
shizz@ics.ict.ac.cn

Abstract. How to reduce the amount of relevance judgments is an important issue in retrieval evaluation. In this paper, we propose a novel method using global statistics to rank retrieval systems without relevance judgments. In our method, a series of global statistics of a system, which indicate the percentage of its documents found by k out of all the N systems ($k = 1, 2, \dots, N$), are selected, then a linear combination of the series of global statistics is utilized to fit the mean average precision (MAP) of the retrieval system. Optimal coefficients are obtained by linear regression. No human relevance judgments are required in the entire process. Compared with existing evaluation methods without relevance judgments, our method has two advantages. Firstly, it outperforms all early attempts. Secondly, it is adjustable for different effectiveness measurements, e.g. MAP, precision at n , and so forth.

Keywords: Information retrieval, evaluation, without relevance judgments, regression.

1 Introduction

Generally, to compare the effectiveness of information retrieval systems, we need to prepare a test collection composed of a set of documents, a set of query topics, and a set of relevance judgments indicating which documents are relevant to which topics. Among these requirements, relevance judgment is the most human resource exhausting and time consuming part. It even becomes incapable when the test collection is extremely large. To address this problem, the TREC conferences used a pooling technology [10], where the top n (e.g., $n=100$) documents retrieved by each participating system are collected into a pool and then only the documents in the pool are judged for system comparison. Zobel [12] has shown that this pooling method leads to reliable results in term of determining the effectiveness of retrieval systems and their relative rankings. Yet, the relevance determination process is still very resource intensive especially when the test collection reaches or exceeds terabyte, or much more queries are included. More seriously, when we change to a new document collection, we have to redo the entire evaluation process.

There are two possible solutions to the problem above, evaluation with incomplete relevance judgments and evaluation without relevance judgments. The former is well studied. Many well designed ranking methods with incomplete judgments were carried out. Two of them, Minimal Test Collection (MTC) method [4] and Statistical evaluation (statMAP) method [2], even got practical application in the Million Query (1MQ) track in TREC 2007 [1], and achieved satisfactory evaluation performance. The latter is comparatively less studied. Only a few papers concentrate on the issue of evaluating retrieval systems without relevance judgments. In Section 2 of this paper, we will briefly review some representative methods. We will see what they are and how they work.

In this paper, we focus our effort on the retrieval evaluation without relevance judgments. Although ‘blind’ evaluation is really a hard problem and its evaluation performance is far less than that of methods with incomplete judgments, it is undeniable that non-judgment evaluation has its own advantages. In some cases, relevance judgments are non-attainable. For example, when researchers compare their novel retrieval algorithms to existing methods, or search for optimal parameters of their algorithms, or conduct data fusion in a dynamic environment, relevance judgment usually seems impossible. Besides, to construct a good evaluation method without relevance judgments, researchers need to mine the retrieval results thoroughly, and try to find laws that indicate the correlation between the effectiveness of a system and features of its retrieval result. These laws are not only useful for ‘blind’ evaluation methods but also valuable for evaluation methods with incomplete judgments.

The main contribution of this paper is that we propose a non-judgment retrieval evaluation method using global statistics of retrieval results, where a linear combination of a series of global statistics of a retrieval system is utilized as an indicator of its retrieval performance. Details of this method will be presented in Section 3. Experimental results, which are reported in Section 4, demonstrate that the proposed method outperforms all the existing methods without relevance judgments. Finally, we conclude our work in Section 5.

2 Related Work

In 2001, Soboroff et al. [6] firstly proposed the concept of evaluating retrieval systems in the absence of relevance judgments. They generated a set of pseudo-relevance judgments by randomly selecting and declaring some documents from the pool of top 100 documents as relevant. This set of pseudo-relevance judgments (instead of a set of human relevance judgments) was then used to determine the effectiveness of the retrieval systems. Four versions of this random pseudo-relevance method were designed and tested on data from the ad hoc track in TREC 3, 5, 6, 7 and 8. They were simple random pseudo-relevance method, the variant with duplicate documents, the variant with Shallow pools and the variant with Exact-fraction sampling. All their resulting system assessments and rankings were well correlated with actual TREC rankings, and the variant with duplicate documents in pools got the best performance, with an average Kendall’s tau value 0.50 over the data of TREC 3, 5, 6, 7 and 8.

Soboroff et al.’s idea came from two results in retrieval evaluation. One is that incomplete judgments do not harm evaluation results greatly. Zobel’s research [12] had

showed that the results obtained using pooling technology were quite reliable given a pool depth of 100. He also found that even though the pool depth was limited to 10, the relative performance among systems changed little, although actual precision scores did change for some systems. The other is that partially incorrect relevance judgments do not harm evaluation results greatly. Voorhees [9] ascertained that despite a low average overlap between assessment sets, and wide variation in overlap among particular topics, the relative rankings of systems remained largely unchanged across the different sets of relevance judgments. These two points are bases of Soboroff et al.'s random pseudo-relevance method, and give explanation to the result that their rankings were positively related to that of the actual TRECs. As a matter of fact, the two points are bases of all the retrieval evaluation methods without or with incomplete relevance judgments.

Aslam and Savell [3] devised a method to measure the relative retrieval effectiveness of systems through system similarity computation. In their work, the similarity between two retrieval systems was the ratio of the number of documents in their intersection and union. Each system was scored by the average similarity between it and all other systems. This measurement produced results that were highly correlated with the random pseudo-relevance method. Aslam and Savell hypothesized that this was caused by 'tyranny of the masses' effect, and these two related methods were assessing the systems based on 'popularity' instead of 'performance'. The analysis by Sporerri [7] suggested that the 'popularity' effect was caused by considering all the runs submitted by a retrieval system, instead of only selecting one run per system. Our later experimental results will show that this point of view is partially correct. The 'popularity' effect could not be avoided completely by only selecting one run per system. This is indeed a hard problem for all the evaluation methods without relevance judgments.

Wu and Crestani [11] developed multiple 'reference count' based methods to rank retrieval systems. They made the distinction between an 'original' document and its duplicates in all other lists, called the 'reference' documents, when computing a document's score. A system's score is the (weighted) sum of the scores of its 'original' documents. Several versions of reference count method were carried out and tested. The basic method (Basic) scored each 'original' document by the number of its 'reference' documents. The first variant (V1) assigned different weights to 'reference' documents based on their ranking positions. The second variant (V2) assigned different weights to the 'original' document based on its ranking position. The third variant (V3) assigned different weights to both the 'original' documents and the 'reference' documents based on their ranking positions. The fourth variant (V4) was similar to V3, except that it normalized the weights to 'reference' documents. Wu and Crestani's method output similar evaluation performance to that of the random pseudo-relevance method. Their work also showed that the similarity between the multiple runs submitted by the same retrieval system affected the ranking process. If only one run was selected for any of the participant system for any query, for 3-9 systems, V3 outperformed random pseudo-relevance method by 45.6%; for 10-15 systems, random pseudo-relevance method outperformed V3 by 6.5%.

Nuray and Can [5] introduced a method to rank retrieval systems automatically using data fusion. Their method consists of two parts. One is selecting systems for data fusion, and the other is selecting documents as pseudo relevant documents as the fusion result.

In the former part, they hypothesized that systems returning documents different from the majority could provide better discrimination among the documents and systems. In return, this could lead to a more accurate pseudo relevant documents and more accurate rankings. To find proper systems, they introduced the ‘bias’ concept for system selection. In their work, bias was 1 minus the similarity between a system and the majority, where the similarity is a normalized dot product of two vectors. In the latter part, Nuray and Can tested three criteria, namely Rank position, Borda count and Condorcet. Experimental results on data from TREC 3, 5, 6 and 7 showed that bias plus Condorcet got the best evaluation results and it outperformed the reference count method and random pseudo relevance method greatly.

More recently, Spoerri proposed a method using the structure of overlap between search results to rank retrieval systems. This method provides us a new view on how to rank retrieval systems without relevance judgments. He used local statistics of retrieval results as indicators of relative effectiveness of retrieval systems. Concretely, if there are N systems to be ranked, N groups are constructed randomly with the constraint that each group contains five systems and each system will appear in five groups; then the percentages of a system’s documents not found by other systems (Single%) as well as the difference between the percentages of documents found by a single system and all five systems (Single%-AllFive%) are calculated as indicators of relative effectiveness respectively. Spoerri found that these two local statistics were highly and negatively correlated with the mean average precision and precision at 1000 scores of the systems. By utilizing the two statistics to rank systems from subsets of TREC 3, 6, 7 and 8, Spoerri obtained appealing evaluation results. The overlap structure of the top 50 documents were sufficient to rank retrieval systems and produced the best results, which outperformed previous attempts to rank retrieval systems without relevance judgments significantly.

So far, we have reviewed 5 representatives of non-judgment evaluation methods. Among these methods, Single% method proposed by Spoerri [8] is the most appealing one. Its average Spearman’s rank correlation coefficient achieves 0.80 over data of TREC 3, 6, 7 and 8. More meaningfully, Spoerri’s method provides us a new view of what information in retrieval results is more valuable for system ranking. Only the random grouping is a little bit confusing. Following study will show that more explicit information can be used in non-judgment retrieval evaluation.

3 Methodology

In this section, we will introduce our method for ranking retrieval systems using global statistics. Basically, our idea comes from the careful study of Spoerri’s work in 2007 [8]. We find that the expectation of local statistics utilized in Spoerri’s research, e.g. Single%, is actually a linear combination of a series of global statistics. So, why don’t we seek for a series of optimal coefficients to make the combination better fit the measurement of systems’ retrieval effectiveness, e.g. MAP or some measurement else? Here comes our idea of ranking retrieval system with global statistics. Before we go into more details of our method, let us check the statistics in Spoerri’s work first.

We have just described Spoerri's method in the previous section. The statistics 'Single%' is the percentage of documents found by a single system and not by other four in a random group. Apparently, 'Single%' is a local statistics, for it involves five systems in a random group. In Spoerri's work, the value of this local statistics is obtained experimentally. More concretely, for a given system, 'Single%' is calculated on each of the five random groups containing this system and each of the 50 topics, then these 'Single%' values are averaged. Obviously, if we replace the average value of 'Single%' with its expectation, the result will remain the same, or become statistically more accurate. Now we check the expectation of 'Single%'.

Suppose that we have N systems, each of which is a document list. Consider a given system and a random group containing it. This means that we have a certain system and four other random systems in the group. For any document that is found by the given system, if it has ever appeared in k out of N systems ($k = 1, 2, \dots, N$), the probability that it appears in the group as 'single' is:

$$p_k^{(1)} = \frac{C_{k-1}^0 \cdot C_{N-k}^4}{C_{N-1}^4} \quad (1)$$

This can be interpreted as the probability that we pick 0 out of $k-1$ systems that contain the document, 4 out of $N-k$ systems that do not contain the document and the given system to form the group.

Thus, by the law of total probability, we have the expectation of 'Single%' as follows:

$$E(\text{Single}\%) = \sum_{k=1}^N p_k^{(1)} (N_k \%) \quad (2)$$

where $p_k^{(1)}$ is described in formula (1), and $N_k \%$ is the percentage of the given system's documents found by k (including the given system) out of all the N systems ($k = 1, 2, \dots, N$). Notice that $N_k \%$ is a global statistics opposite to local statistics.

Similarly, we can write the expectations of 'AllFive%' and other local statistics in the form like formula (2):

$$E(\text{AllFive}\%) = \sum_{k=1}^N p_k^{(5)} (N_k \%) \quad (3)$$

and

$$E(\text{Single}\% - \text{AllFive}\%) = \sum_{k=1}^N (p_k^{(1)} - p_k^{(5)}) (N_k \%) \quad (4)$$

where

$$p_k^{(5)} = \frac{C_{k-1}^4 \cdot C_{N-k}^0}{C_{N-1}^4} \quad (5)$$

Now we get that the expectation of local statistics used in Spoerri's method is actually a linear combination of a series of global statistics. With formula (2) and (4), we do not need to consider the random grouping any more. If we have these global statistics from the retrieval document lists, we can obtain the expectations of statistics used in Spoerri's method.

Here comes the question. A linear combination of these global statistics with fixed coefficients can be a good indicator of system's relative effectiveness, what if we replace the fixed coefficients with the optimal coefficients? It will definitely produce better system rankings. Besides, the optimal coefficients can be tunable. Different coefficients could be optimized corresponding to different effectiveness measurements, e.g. MAP, precision at n , or any sound measurements. This is our idea.

To make our method experimentally comparable to early methods without relevance judgments, we will use the MAP measurement as the target of our optimization in this work. That is, we are seeking for a series of coefficients a_1, a_2, \dots, a_M , so that we can minimize the sum of squares of errors with the true MAP:

$$\sum_{i=1}^N (y_i - MAP_i)^2 \quad (6)$$

where MAP_i is the MAP value of the i th system and y_i is defined as:

$$y_i = \sum_{k=1}^M a_k (N_k^{(i)} \%) \quad (7)$$

where a_k is the coefficient to be optimized, and $N_k^{(i)}\%$ is the percentage of the i th system's documents found by k out of all the N systems ($k = 1, 2, \dots, M, M \leq N$). By using linear regression, we can easily get these optimal coefficients. In turn, we calculate y_i for the i th system and obtain their rankings eventually.

Typically, when devising methods for retrieval evaluation without relevance judgments, researchers often seek for some law(s) inside a small part of data and apply the law(s) on the entire data set to see whether it works well. Accordingly, we will generate 5 series of coefficients optimized based on the data from TREC 3, 5, 6, 7 and 8 respectively, and examine their ranking performances on all the 5 data sets. Each of the 5 series of coefficients is in fact an implementation of our evaluation method. R3, R5, R6, R7 and R8 are short for these 5 series of coefficients as well as their corresponding ranking methods, where Rx means the method comes from TREC x ($x=3, 5, 6, 7, 8$).

4 Experimental Results

4.1 Some Clarification

Before we come to the experimental results, we would like to make some details clear first.

Firstly, in our experiments, the value of M in formula (7) is set to 30. The number of systems (runs), N , varies in different TREC data (see Table 1 for details). To make

our method optimized based on one TREC data capable for being applied to others, we need a fixed number of M , which fits for all N of TREC 3, 5, 6, 7, 8. We also noticed that as parameter k goes from 1 to N , the statistics $N_k^{(i)\%}$ decreases rapidly to zero. A fixed number of M , if not too small, will not make the model lose too much information. 30 is a good but not only choice for parameter M . It fits all TREC data, and is not too small.

Table 1. Number of TREC runs

TREC	Number of Runs
3	40
5	61
6	74
7	103
8	129

Secondly, different from Spoerri's work, we plan to rank all the systems for each TREC opposite to a subset of them. Without any limitation, we will definitely encounter the problem of 'popularity' effect mentioned previously in Section 2. To avoid this situation, when we calculate statistics $N_k^{(i)\%}$, different runs from same system will be counted only once.

Besides, to make a fair comparison, we need to repeat Spoerri's method over all systems for each TREC. The repetition is not exactly the same as Spoerri's original one. Based on the analysis in previous subsection, we replace the average of 'Single%' with the expectation of 'Single%', so that we can eliminate random turbulence in the original method. We process the statistics 'Single%-AllFive%' in the same way.

At last, the correlation between the rankings from our proposed methods, as well as other methods to be compared with, and the TREC official rankings (based on MAP) is measured using the Spearman's rank correlation coefficient. One reason is that it suits better for evaluating correlation between ratio sequences, e.g. MAP, than Kendall's tau. The other reason is that we can directly compare our results with those of previous attempts reviewed in Section 2, since most of them provided Spearman's rank correlation coefficient results.

4.2 Model Selection

Spoerri had stated that the overlap structure of the top 50 documents were sufficient to rank retrieval systems and produced the best results [8]. We just test our methods and Spoerri's methods with three pool sizes, namely 20, 30 and 50.

Among all our methods, R5 with pool size 20 produces the best result. It also works very stable. So we take R5 as the representative of our method. For Spoerri's methods, Single% with pool size 20 is selected as a representative, for it works slightly better than Single%-AllFive% on all systems.

4.3 Comparison with All Previous Attempts

We make a comparison between our method and all previous attempts. The comparison result is given in Table 2.

Table 2. Spearman’s correlation coefficients for best results from different methods

	RC	RS	BC	SS	Single%	R5
Trec3	0.587	0.627	0.867	0.751	0.824	0.716
Trec5	0.421	0.429	0.657*	0.488	0.563	0.912
Trec6	0.384	0.436	0.717	0.609	0.618	0.601
Trec7	0.382	0.411	0.453	0.551	0.550	0.603
Avg.3-7	0.444	0.476	<u>0.674</u>	0.600	0.639	0.708
Std.3-7	0.097	0.101	0.210	0.112	0.127	0.146
Trec8	-	-	-	0.613	0.569	0.514
Avg.3-8	-	-	-	0.602	<u>0.625</u>	0.669

In Table 2, RC is the best result produced by reference count method; RS represents the result of random pseudo relevance method, where relevance ratio is set to 10% rather than the actual ratio in its original version; BC accounts for the result of Bias plus Condorcet method, a data fusion based method. Results of these three methods are cited from Nuray and Can’s paper [5]. They did not provide results on TREC 8, so we just have their results on TREC 3, 5, 6 and 7. For the number with a ‘*’ (BC on TREC 5), in their original paper, same result in different tables conflict, and we pick the bigger number presenting in Table 5. SS is short for method based on system similarity. Since there is no Spearman’s rank correlation coefficient result available in Aslam and Savell’s work [3], we make an implementation of this method. In the implementation, we have tested several pool depths, where pool depth 100 produces the best result, thus is presented in Table 2. Single% is the representative of Spoerri’s overlap structure based method, and R5 is the representative of our method.

Each line of Table 2 presents results from different methods on same TREC data. The bold number indicates the best result on a TREC data. We can see that over all five TREC data, BC method achieves best twice, SS method wins best once, and R5 method gets best twice. Especially, R5 method gets the best result on the two average evaluation performances. When averaging on TREC 3-7, R5 method outperforms the second best result (from BC with underline in Table 5) 5%. When averaging on TREC 3-8, R5 method outperforms the second best result (from Single% with underline in Table 5) 7%. In a word, regarding the average Spearman’s correlation coefficients on TREC data 3, 5, 6, 7 and 8, R5 method outperforms all the existing retrieval evaluation methods that do not use human relevance judgments.

Besides, we find that all methods work quite unstably. The stand deviation of Spearman’s correlation coefficients for all methods on TREC data 3, 5, 6 and 7 runs from 0.097 to 0.210. The better average evaluation result a method gets, the more instable it is. Our R5 method is an exception. It gets the best average result but the second large deviation.

5 Conclusion and Future Work

We end this paper with a conclusion reemphasizing the main points of our work.

In this work, we propose a retrieval evaluation method using global statistics of retrieval systems, where a linear combination of a series of global statistics of a retrieval system is utilized as an indicator of its retrieval performance. Compared with existing evaluation methods without relevance judgments, our method has two advantages. Firstly, it outperforms all early attempts regarding data on TREC 3, 5, 6, 7 and 8. Secondly, the method is adjustable for different effectiveness measurements, e.g. MAP, precision at n , and so forth. In contrast, some early attempts, e.g. reference account method, system similarity method and Single% method, can not change their scoring strategy to fit different effectiveness measurements.

The proposed method has its weakness as well. It works unstably on different data set, and mixes best systems with ordinary ones. This is also the common problem for all non-judgment evaluation methods. With meticulous analysis, we have found the fundamental factor that depresses the performance of non-judgment evaluation. How to tackle this problem is our future work.

Acknowledgments. This work is supported by the National Science Foundation of China under Grant No. 60776797, the Major State Basic Research Project of China (973 Program) under Grant No. 2007CB311103 and the National High Technology Research and Development Program of China (863 Program) under Grant No. 2006AA010105.

References

1. Allan, J., Carterette, B., Aslam, J.A., Pavlu, V., Dachev, B., Kanoulas, E.: Overview of the TREC 2007 Million Query Track. In: Proceedings of TREC (2007)
2. Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2006)
3. Aslam, J.A., Savell, R.: On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2003)
4. Carterette, B., Allan, J., Sitaraman, R.: Minimal test collections for retrieval evaluation. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA (2006)
5. Nuray, R., Can, F.: Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management: an International Journal* 42(3), 595–614 (2006)
6. Soboroff, I., Nicholas, C., Cahan, P.: Ranking retrieval systems without relevance judgments. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, Louisiana, United States, pp. 66–73 (2001)
7. Spoerri, A.: How the overlap between search results correlates with relevance. In: Proceedings of the 68th Annual Meeting of the American Society for Information Science and Technology (2005)

8. Spoerri, A.: Using the structure of overlap between search results to rank retrieval systems without relevance judgments. *Information Processing and Management: an International Journal* 43(4), 1059–1070 (2007)
9. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 315–323 (1998)
10. Voorhees, E.M., Harman, D.: Overview of the eighth text retrieval conference (TREC-8). In: *The 8th Text Retrieval Conference (TREC-8)*, Gaithersburg, MD, USA (1999)
11. Wu, S., Crestani, F.: Methods for ranking information retrieval systems without relevance judgments. In: *Proceedings of the 2003 ACM Symposium on Applied Computing*, Melbourne, Florida (2003)
12. Zobel, J.: How reliable are the results of large-scale information retrieval experiments? In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 307–314 (1998)

Rule Learning with Negation: Issues Regarding Effectiveness

Stephanie Chua, Frans Coenen, and Grant Malcolm

University of Liverpool
Department of Computer Science,
Ashton Building, Ashton Street, L69 3BX Liverpool,
United Kingdom
s.chua@liverpool.ac.uk,
coenen@liverpool.ac.uk,
grant@liverpool.ac.uk

Abstract. An investigation of rule learning processes that allow the inclusion of negated features is described. The objective is to establish whether the use of negation in inductive rule learning systems is effective with respect to classification. This paper seeks to answer this question by considering two issues relevant to such systems; feature identification and rule refinement. Both synthetic and real datasets are used to illustrate solutions to the identified issues and to demonstrate that the use of negative features in inductive rule learning systems is indeed beneficial.

Keywords: Inductive rule learning, Negation, Classification.

1 Introduction

Inductive Rule Learning (IRL) is a generic term used to describe machine learning techniques for the derivation of rules from data. IRL has many applications; this paper is concerned with IRL techniques to build rule-based classifiers. The advantage offered by IRL, over many other forms of machine learning techniques (such as support vector machines, neural networks and self organising maps) is that the disjunctive normal form (DNF) rules produced are expressive while at the same time being easily interpretable by humans.

In the context of classification, the derived rules are typically of the form *condition* \rightarrow *conclusion*; where the *condition* (antecedent) consists of a conjunction of features, while the *conclusion* (consequent) is the resulting class label associated with the condition. For example, the rule $a \wedge b \wedge c \rightarrow x$ (where a , b and c are features that appear in a dataset, and x is a class label) is interpreted as, if a and b and c occur together in a document, then classify the document as class x . With respect to most IRL systems, rules do not normally include the negation of features. For example, $a \wedge b \wedge \neg c \rightarrow x$, which would be interpreted as, if a and b occur together in a document and c does not occur, then classify the document as class x . Intuitively, rules that include negation seem to provide a powerful mechanism for distinguishing examples for classification; the inclusion of negation

should serve to improve classification accuracy. This paper seeks to establish whether the use of negation in IRL is indeed beneficial with respect to classification. When considering the effectiveness of IRL with negation, there are two significant issues that need to be considered:

- a. Feature identification: The identification of appropriate features to be negated.
- b. Rule refinement strategies: The strategies for learning rule with negation.

The rest of this paper is organized as follows. A brief review of relevant previous work is presented in Section 2. In Section 3, a scenario illustrating the need for rules with negation is presented. Section 4 will discuss the issues highlighted. Section 5 describes the experiments carried out to determine the effectiveness of rules with negation, as well as the results and analysis. Section 6 concludes.

2 Previous Work

Existing work on IRL for classification tends to adopt a two-stage process: rule learning, followed by rule pruning. Examples of such systems include: (i) Reduced Error Pruning (REP) (Brunk et al., 1991), which incorporates an adaptation of decision tree pruning; (ii) Incremental Reduced Error Pruning (IREP) (Fürnkranz et al., 1994), an enhancement over REP, (iii) Repeated Incremental Pruning to Produce Error Reduction (RIPPER) (Cohen, 1995), a further enhancement over IREP, and (iv) Swap-1 (Weiss et al., 1993). All these systems use the covering algorithm for rule learning, shown in Figure 1, whereby rules are “learned” sequentially based on training examples. The examples “covered” by a learnt rule are then removed and the process is repeated until some terminating condition is met.

Algorithm: Sequential covering. Learn a set of rules for classification.

Input:

- D , a data set class-labelled tuples;
- Att_vals , the set of all attributes and their possible values;

Output: A set of IF-THEN rules.

Method:

```

Rule_set = { }; //initial set of rules learned is empty
for each class  $c$  do
    repeat
        Rule = Learn_One_Rule( $D$ ,  $Att\_vals$ ,  $c$ );
        remove tuples covered by  $Rule$  from  $D$ ;
    until terminating condition;
    Rule_set = Rule_set + Rule; //add new rule to rule set
endfor
return Rule_set;

```

Fig. 1. Basic sequential covering algorithm (Han et al., 2006)

None of the above exemplar systems include an option to build negation into the generated rules. Examples of IRL approaches that generate rules with negation are much rarer. Wu et al. (Wu et al., 2002) and Antonie et al. (Antonie et al., 2004) considered both positive and negative Association Rules (ARs) in their work on AR mining (a classification rule of the form described in Section 1 may be considered to be a special type of AR). Negative features are also used by Zheng et al. (Zheng et al., 2003). However, their work does not involve the direct generation of rules with negation. They combined positive and negative features in their feature selection method for text classification using the Naïve Bayes classifier. Galavotti et al. (Galavotti et al., 2000) use negative evidence in a novel variant of k-NN. None of these systems can be truly described as being classification rule learning systems.

More recently, Rullo et al. (Rullo et al., 2007) have proposed a system called Olex that used both positive and negative features for rule learning. The system was directed at text classification and comprised a single stage rule learning process with no post-learning optimization (i.e. pruning). Rullo et al. proposed a paradigm of “one positive term, more negative terms”, where the positive term allows the identification of the right documents, thus, giving high recall values; while the negative terms help reduce the number of wrong classifications, thus, improving precision. The core of their method was in the selection of discriminating terms, which were selected from a reduced vocabulary to maximize the F1-measure value when using that set of terms to generate rules for classification. Each rule generated consisted of conjunctions of a single positive feature with none or more negative features. While the notion of using both positive and negative features seemed very promising, Rullo et al. also highlighted that their approach was not able to express co-occurrence based on feature dependencies (by allowing exactly one positive feature in a rule antecedent) and that this could affect the effectiveness of the text classifier. Thus, Olex is unable to generate rules of the form $a \wedge b \wedge c \rightarrow x$.

It is of course possible to define features that describe the negation of features; given a feature “blue”, we can define two binary-valued features: *blue* and \neg *blue*, which can then be considered by a “standard” IRL system. However, in the opinion of the authors, this is not a true IRL with negation approach. To the best knowledge of the authors, there are no reported IRL systems that incorporate the concept of negation as defined here.

3 Motivation

As noted in Section 1, rules of the form of *condition* \rightarrow *conclusion* are the standard output from IRL algorithms; the *condition* part is usually a conjunction of positive features. Rules of this form are often sufficient for the classification of new and unseen data. However, there are cases where rules with negation produce a more effective rule set. This section seeks to establish that IRL with negation is necessary with respect to some data scenarios.

Assume a set of features $A = \{a, b, c, d\}$ and a set of class labels $C = \{x, y, z\}$ that can occur in a data set. Thus, we might have a data set of the form given in Table 1.

Table 1. Example data set 1

{a, b, x}
{a, c, x}
{a, d, y}
{a, d, y}

Table 2. Example data set 2

{a, b, x}
{a, b, c, y}
{a, c, z}

To apply IRL, the features must first be ordered according to which are the best discriminators, thus {d, b, c, a} (b, c and d are all excellent discriminators but d covers more records so is listed first). The strategies described in this paper (see Section 4.2) use chi square ordering. Processing this data set in the standard IRL manner (without negative features) produces these rules: $b \rightarrow x$, $c \rightarrow x$ and $d \rightarrow y$, respectively. By introducing negation, we can get a more succinct set of rules: $a \wedge d \rightarrow x$ and $d \rightarrow y$. Thus, in this case the use of negation has produced what may be argued to be a better (smaller and therefore more effective) rule set.

Considering the data set given in Table 2, it is more difficult to order the features. However, features b and c can be argued to be better discriminators than a because at least, they distinguish between one and the remaining classes, thus {b, c, a}. Starting with the first record, the rule $b \rightarrow x$ will be produced, which would have to be refined to $b \wedge c \rightarrow x$ to give the correct result. Moving on to the next record will give $b \rightarrow y$, and then $c \rightarrow z$. Rearranging the ordering of the data set does not avoid the need for a negated rule. This example clearly illustrates the need for IRL with negation.

4 Inductive Rule Learning with Negation

The illustration in Section 3 provides a clear motivation for IRL with negation. However, this leads to the question of which feature to add to a rule when refining a rule. If a rule with negation is to be generated, which feature should be negated? If both positive and negative features are available, is the rule better refined with a positive feature or a negative feature? This section discusses these two issues.

4.1 Feature Identification

Using our proposed approach, rules are initiated by selecting a feature associated with a class from a chi-square ordered list of features. Thus, all rules start with a single positive feature. If a rule covers both positive and negative examples, then the rule has to be further refined in order to learn a rule that can separate the examples. Positive examples are those training set records that are classified correctly given a current rule; negative examples are those that are classified incorrectly. Using our approach, the search space can be conceptualised as containing features that belong to positive and negative examples. This paper proposes that the search space be divided into three sub-spaces that contain different kinds of feature: (i) unique positive (UP) features which are found only in positive examples, (ii) unique negative (UN) features found only in negative examples, and (iii) overlap (Ov) features that are found in both positive and negative examples. This division allows efficient and effective identification of features that can be negated. It should be noted that the UP, UN and Ov feature

categories may be empty as the existence of these features is dependent upon the examples covered by a rule. Where categories contain more than one feature, the features are ordered according to the frequency with which each feature occurs in the collection of examples covered by the current rule (one count per example).

4.2 Rule Refinement Strategies

If a rule is refined with a UP or an Ov feature, then a rule with no negation is generated. If a rule is refined with a UN feature, then a rule with negation is generated. When refining a rule with a UP or UN feature, the feature with the highest document frequency (appears in the most covered examples) is selected. When refining a rule with an Ov feature, the feature with the highest frequency difference (i.e. positive frequency minus negative frequency) is selected.

Table 3. Example of rule refinement with UP, UN and Ov features

<p>Feature set for class $x = \{bike, ride, harley, seat, motorcycles, honda\}$ Initial rule learnt = $bike \rightarrow x$</p> <p>The rule covers three examples (two +ve examples and one -ve example):</p> <p>$\{bike, ride, motorcycles, x\}$ $\{seat, harley, bike, ride, x\}$ $\{bike, ride, honda, y\}$</p> <p>Identify UP, UN and Ov features UP features = $\{motorcycles, seat, harley\}$ UN features = $\{honda\}$ Ov features = $\{ride\}$</p> <p>Strategies for rule refinement Refine with UP feature = $bike \square motorcycles \rightarrow x$ Refine with UN feature = $bike \square \neg honda \rightarrow x$ Refine with Ov feature = $bike \square ride \rightarrow x$</p>

Table 3 shows an example of refining a rule with UP, UN and Ov features. The refinement process will be repeated until the stopping condition is met; either: (i) when the rule no longer covers negative examples, (ii) the rule antecedent size reaches a pre-defined threshold or (iii) there are no more features that can be added to the rule. At every round of refinement, the examples covered will change and therefore, the search space will also change.

Given the UP, UN and Ov feature categories, a number of strategies can be identified whereby these categories can be utilized. These strategies may be defined according to the order in which they are considered. The Ov category, which comprises features that occurs in both positive and negative examples, is the least likely to result in successful refinement. Thus, it is argued that this should be considered last. Thus, we

have two possible strategies involving all three categories in sequence: UP-UN-Ov (UP first if it is not empty, then UN, then Ov) and UN-UP-Ov. Alternatively, we can refine rules using only the UP or UN collection. This gives rise to two more strategies: UP and UN. Note that the UP strategy, which does not entail negation, is the bench-mark strategy (use of negation must improve on this). Note also that the UN strategy produces rules that are identical to the rule structure that Olex (Rullo et al., 2007) generates as described in Section 2.

When refining rules using UP or UN, only one type of feature is used for the refinement. In contrast, the sequence combinations of UP-UN-Ov and UN-UP-Ov allow the use of UP, UN and Ov features when the preceding feature category in the sequence does not exist. A more flexible proposed strategy is UP-or-UN. The mechanism for this is to refine a rule by generating two versions and selecting the better version; one version is refined by UP and another version is refined by UN. The rule with the higher Laplace estimation accuracy is selected as the better rule.

5 Experimental Evaluation

This section describes the experimental setup used to investigate the proposed use of feature sub-spaces (UP, UN and Ov) and the five different rule refinement strategies devised. The results and analysis of each experiment are also discussed. Three different categories of data set were used for the experimental evaluation: (i) a collection of synthetic data sets covering all possible combination of a given set of features and classes, (ii) text mining data sets extracted from the well known 20 Newsgroups collection, and (iii) a selection of data sets taken from the UCI repository (Blake et al. 1998). In all cases, single-labelled (as opposed to multi-labelled) classification was conducted.

5.1 Synthetic Datasets

The synthetic data sets were constructed by considering every combination of a set of features $A = \{a, b, c\}$ and a set of class labels $C = \{x, y, z\}$. Given that $|A| = 3$, there are $2^3 - 1 = 7$ possible feature combinations. It was assumed that each record could contain only a single class label. Thus, there were $7 * 3 = 21$ variations per record. Each data set was assumed to comprise 3 records, thus overall $21^3 = 9261$ data sets were generated covering all possible record permutations (including data sets containing contradictions). The five strategies described in Section 4.2 were applied to the data sets. The results are shown in Table 4. The rows in Table 4 indicate the number of synthetic data sets where the generated classifier accurately covered all 3 records (100% accuracy), only 2 records (67% accuracy) and only 1 record (33% accuracy) respectively. Comparing the results using the UP and UN strategies in Table 4 provides further evidence for the need for IRL with negation. Using the UN strategy, many more 100% accurate classifiers are generated than using the UP strategy. Using the UP-UN-Ov and UN-UP-Ov strategies allows the inclusion of all feature types which enhances the result even further. Inspection of the 2,436 cases where 100%

accuracy was not obtained using both the UP-UN-Ov and UN-UP-Ov strategies, indicates that these mostly include contradictions which can never be entirely satisfactorily resolved. Use of the UP-or-UN strategy produces identical results to when the UN strategy is used; suggesting that at every round of refinement in the UP-or-UN strategy, the rule refined by UN is a better rule that is selected. The reason that the results for the UP-UN-Ov and UN-UP-Ov strategies, and for the UN and UP-or-UN strategies are identical is also due to the small size of the individual data sets used in the experiment, where the number of features is small. In general, it can be observed that strategies involving the generation of rules with negation produce better results than strategies without the use of negation.

Table 4. Results for synthetic data sets

Accuracy	Rule Refinement Strategy				
	UP	UN	UP-UN-Ov	UN-UP-Ov	UP-or-UN
100%	4,503	6,717	6,825	6,825	6,717
67%	3,324	2,352	2,316	2,316	2,352
33%	1,434	192	120	120	192
Total	9,261	9,261	9,261	9,261	9,261

5.2 Text Mining Datasets

For the text mining experiment, the 20 Newsgroups data set¹ was used in the context of binary classification. The 20 Newsgroups dataset is a collection of news items comprising 19,997 documents and 20 classes. The dataset was split into two parts: 20 Newsgroups A (20NGA) comprising 10,000 documents and the first 10 classes, and 20 Newsgroups B (20NGB) comprising 9,997 documents and the remaining 10 classes. Stop words removal was applied; followed by feature selection, based on the chi-square metric, where the top 1,000 features in each class was selected to be used in the text representation vector. Chi-square was chosen as the feature selection method due to its reported success in the literature (Yang et al., 1997; Debole et al., 2003; Zheng et al., 2003). The 1,000 features threshold was chosen to ensure a sufficiently large collection of features for each class is obtained. A rule size threshold of five was imposed on rule learning to generate rules that were not overly specific. Post-processing of the generated rule set was conducted by removing rules with coverage lower than a pre-defined threshold of 1.5% of the documents in the class (i.e. 15 documents with respect to the 20 Newsgroups), and a Laplace estimation rule accuracy value lower than 60%. Average ten-fold cross validation accuracy and F1-measure results across all classes in each fold using the different refinement strategies are presented in Table 5 (best results are highlighted in **bold font**).

¹ <http://people.csail.mit.edu/jrennie/20Newsgroups/>

From Table 5, it is noted that the UN strategy has the best results for accuracy in both 20NGA and 20NGB. In terms of the F1-measure, the UN strategy has the highest value in 20NGB while the UP-or-UN strategy did best in 20NGA. The UP and UP-UN-Ov strategies recorded the same results, suggesting that at every round of rule refinement, UP features exist and therefore, only rules without negation are generated. The UN-UP-Ov strategy did not improve on the UN strategy. This hinted that using the UN strategy may be sufficient in learning an effective rule set. The UP-or-UN strategy obtained a slightly higher F1-measure than the UN strategy although its accuracy is slightly lower. Overall, the results indicate sound support for the use of negation in IRL.

Table 5. Results for 20 Newsgroups datasets

Datasets	Rule refinement with									
	UP		UN		UP-UN-Ov		UN-UP-Ov		UP-or-UN	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
20NGA	92.6	62.7	93.1	63.0	92.6	62.7	92.6	61.3	92.4	63.7
20NGB	93.4	66.6	94.0	70.8	93.4	66.6	93.7	67.3	93.2	68.0

5.3 UCI Datasets

Further binary classification experiments were conducted using data sets selected from the UCI repository (Blake et al. 1998), namely: Anneal, Breast Cancer, Iris, Pima Indians and Wine. The datasets were first normalised and discretized using the LUCS-KDD normalisation software². Again, a rule size threshold of five was imposed on rule learning. Post-processing of the generated classification rules was conducted by removing rules with a Laplace estimation rule accuracy value lower than 60%. Average accuracy and F1-measure value using ten-fold cross validation across all classes in each fold with the different rule refinement strategies are presented in Table 6 (again, best results are highlighted in **bold** font).

From Table 6, it can be observed that results are mixed. The first observation that can be made is that there are notable differences in the results obtained for UP-UN-Ov and UN-UP-Ov with that of UP and UN respectively, indicating that with respect to some of the generated rules there are no UPs and/or UNs. The best overall accuracy recorded for the Anneal data set was using the UP-UN-Ov strategy, while the highest overall F1-measure was obtained using the UN strategy. In the Breast Cancer data set, the UP-UN-Ov and UN-UP-Ov strategies produce the highest accuracy and F1-measure. It is also worth noting that in this case UP-UN-Ov and UN-UP-Ov significantly out-performed the other strategies. The UP-or-UN strategy produced the best accuracy and F1-measure for the Iris data set; and the UN strategy recorded the best accuracy and F1-measure for the Pima data set. The only data set where the UP strategy recorded the best accuracy and F1-measure was the Wine data set. It can also be

² http://www.csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN_ARM/lucs-kdd_DN.html

observed that using the UP-UN-Ov strategy always improves on the UP strategy except in the Wine data set. Overall, the results indicate that strategies that allow the generation of rules with negation generally perform better than strategies that generate rules without negation.

Table 6. Results for UCI datasets

Datasets	Rule refinement with									
	UP		UN		UP-UN-Ov		UN-UP-Ov		UP-or-UN	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Anneal	96.7	64.8	97.0	66.7	97.6	65.6	96.4	64.8	97.5	64.4
Breast	78.8	83.0	77.2	83.0	92.6	92.3	92.6	92.3	85.5	87.1
Iris	90.2	85.1	96.7	94.8	95.1	91.7	95.3	92.5	96.9	95.0
Pima	51.4	34.6	73.3	66.7	70.7	60.5	72.1	64.3	66.1	52.2
Wine	91.0	86.1	87.6	77.4	89.5	83.6	90.6	84.6	89.7	85.1

6 Conclusion

This paper has sought to establish whether IRL with negation is effective or not with respect to the classification problem. This entails two issues: (i) the mechanism for identifying features to be negated and (ii) the strategies for deciding whether to add a positive or a negative feature. The paper proposes a solution to the first by dividing the search space, with respect to a current rule, into three sub-spaces designated as UP, UN and Ov. Five strategies for refining rules are considered, including a benchmark strategy (UP) that does not generate negated rules. The reported experimental results indicate that the use of negation in IRL is indeed beneficial. For future work, the authors intend to conduct further experiments and investigate alternative strategies. This includes the comparison of different feature selection methods with respect to IRL with negation.

References

- Antonie, M.-L., Zaïane, O.R.: An associative classifier based on positive and negative rules. In: Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp. 64–69 (2004)
- Blake, C.L., Merz, C.J.U.: Repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, CA (1998)
<http://www.ics.uci.edu/~mllearn/MLRepository.html>
- Brunk, C., Pazzani, M.: Noise-tolerant relational concept learning algorithms. In: Proceedings of the 8th International Workshop on Machine Learning, Morgan Kaufmann, New York (1991)

- Cohen, W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning (ICML), pp. 115–123. Morgan Kaufmann, San Francisco (1995)
- Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. In: Proceedings of the 18th ACM Symposium on Applied Computing, pp. 784–788 (2003)
- Fürnkranz, J., Widmer, G.: Incremental reduced error pruning. In: Proceedings of the 11th International Conference on Machine Learning (ICML), Morgan Kaufmann, San Francisco (1994)
- Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the use of feature selection and negative evidence in automated text categorization. In: Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries, pp. 59–68 (2000)
- Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco (2006)
- Rullo, P., Cumbo, C., Policicchio, V.L.: Learning rules with negation for text categorization. In: Proceedings of the 2007 ACM Symposium on Applied Computing, pp. 409–416. ACM, New York (2007)
- Weiss, S.M., Indurkha, N.: Optimized rule induction. *IEEE Expert: Intelligent Systems and Their Applications* 8(6), 61–69 (1993)
- Wu, Z., Zhang, C., Zhang, S.: Mining both positive and negative association rules. In: Proceedings of the 19th International Conference on Machine Learning, pp. 658–665 (2002)
- Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning (ICML), pp. 413–420 (1997)
- Zheng, Z., Srihari, R.: Optimally combining positive and negative features for text categorization. In: Proceedings of the International Conference on Machine Learning (ICML), Workshop on Learning from Imbalanced Datasets II (2003)

Integrating Web Videos for Faceted Search Based on Duplicates, Contexts and Rules

Zhuhua Liao^{1,2,3}, Jing Yang¹, Chuan Fu¹, and Guoqing Zhang¹

¹ Institute of Computing Technology, Chinese Academy of Sciences

² Graduate School of the Chinese Academy of Sciences

³ Key Laboratory of Knowledge Processing and Networked Manufacturing,
College of Hunan, Xiangtan, China

{liaozhuhua, jingyang, chuanfu, gqzhang}@ict.ac.cn

Abstract. We propose a novel video integration architecture, INTERVIDEO, for faceted search on web-scale. First, we demonstrate that the traditional video integration techniques are no longer valid in face of such heterogeneity and scale. Then, we present three new integrating techniques to build a global relation schema for organizing web videos and aiding user to retrieve faceted results. Finally, we conduct an experimental study and demonstrate the ability of our system to automatically integrate videos and build a complete and concise high-level relation schema on large, heterogeneous web sites.

Keywords: Video integration, local relation view, global relation schema, faceted search.

1 Introduction

Since there has been exponential growth with the popularity of social media in Web 2.0, the video collection environments are leading to the need for flexible video retrieval systems which deal with adaptive, multi-faceted search [1]. Faceted search provides flexible access to information by one or more facets which represent dimensions of information (e.g., category, time and location). However, there are many challenges for such faceted search to web videos. First, the semantic knowledge of videos such as annotation is very sparse, where the problem of query answering with incomplete information is intractable. Second, there are lacks of integration approaches on multiple dimensions for relevant content which reside at different video sources to organize web videos and enrich video's knowledge.

Similar aspects of research can be found on faceted search [1,2], data integration [3,4] and video retrieval system [5,6]. However, the work of faceted search only focus on the faceted metadata and category-based interface design, but not the information organization with multi-facets, especially the web videos' organization; the traditional work of data integration were mostly based on deep-web sources and mapping or reformulating of heterogeneous data schemata, such as the Meta-Querier project [7] and the PayGo architecture [8]. Recently, many content sites can share structured data to users and other web sites by initiatives like OpenID and OpenSocial, In [9], the authors propose the SocialScope to integrate data based on OpenID and OpenSocial. But

in the all work, they do not consider video integration on heterogeneous and video collection with the features of sparse annotations and distributing discrete, nonintegrated videos on the Web. And the video retrieval system’s work is only intended for matching by text, image, and concept, etc.

Video integration for efficient video search has two broad goals[4]: increasing the completeness and increasing the conciseness of relation view over video collections that is available to query and index to users and applications. An increase in completeness is achieved by adding more video sources (more videos, more attributes describing video) to the system and integrating sources that supply additional attributes to the relation. An increase in conciseness is achieved by removing redundant videos and links, and aggregating duplicates and merging common attributes into one.

The goal of our video integration system is to combine the annotations, contexts and various relations of relevant videos which residing at different sources, providing the user with a unified relation view, called *global relation schema*. User formulates queries over the global relation schema, and the system suitably queries the sources, providing complete, concise and faceted results to the user.

2 System Overview

This section describes the design and implementation of the INTERVIDEO system. INTERVIDEO is modeled as a client-server system, where the search clients interact with both web video sites and video integration server. The overall system architecture is presented in Figure 1. In the system, we first use information retrieving tools to retrieve video’s annotations and relationships for building local relation view of videos. Then, we integrate various local relation views with new techniques to build global relation schema and refine it.

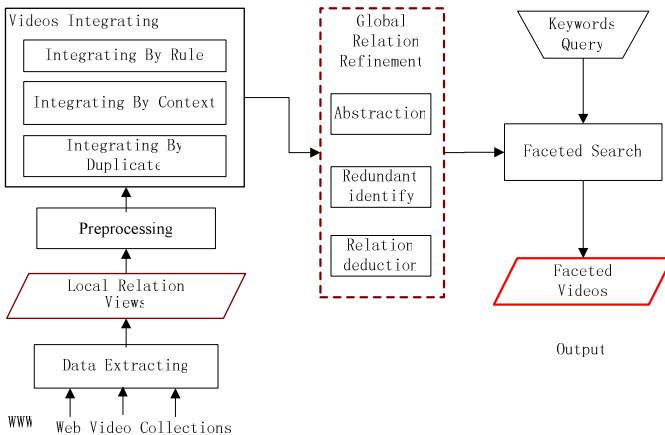


Fig. 1. INTERVIDEO System Architecture

● **Local relation view retrieving.** In general, the intense semantic relationships of videos can be found in the published web pages. At present, many techniques have

been proposed to mine and retrieve the relation links imbedded in web pages. In order to extracting local relation view from HTML codes we use information extracting tool.

● **Global relation schema building.** In order to building the global relation schema, we propose three classes of novel techniques. These are (1) duplicate-based integration technique which takes relationship of immediately duplicate to integrate videos and enrich video's annotations; (2) the context-based integration technique which leverages the contexts such as tagging to identify the relationships between videos; (3) the rule-based integration technique which uses rules that user specified to integrate videos. In the section 4, we will describe three techniques in detail.

3 Local Relation View Retrieving and Duplicate Detecting

Information extracting from HTML [10,11,12,13] is usually performed by software modules called *wrapper*. In most cases, a practicable wrapper should be able to identify the template, and hence extract the data fields from any new pages having the same template. In the system, we use the RoadRunner[10] and specify templates to extract local relation view from web pages on different web sites. The templates defined by HTML codes and compiled to a consistent specification in XHTML, a restrictive variant of HTML. The specification defines a set of interrelated entities: a video element links a set of duplicates videos and annotation; and a set of correlative videos with logic relation (e.g. sequential relation). The data extraction process is introduced in [10] in detail.

Among huge video collections with many near duplicate videos (X. Wu observe that on average there are 27% redundant videos[14]), efficient near duplicate detection is essential for effective search, retrieval and integration. We built a video duplicate detector to detect near duplicate [14] in video collections based on the work originally presented in [16]. This fingerprint-based method relies on robust hash functions, which take an input message (video frames in our case) and generate a compact output hash value, with the condition that similar input messages generate similar output values. All videos involved in the detection are converted into hash values, and detection is performed as a search problem in the hash space. The system uses the robust hash functions and search procedure which described in [16]. The precision-recall was verified approximately 0.8 [15].

4 Global Relation Schema Building

In the paper, we consider the relation view of videos as a relational graph and the integration of relation views is equal to merge two or more graphs.

Definition 4.1 (Graph). A (relational) graph is a tuple $G=(V, E,R,W)$ where V is a set of nodes, E is a set of edges, R is a set of relation of each edge, and W is a weight matrix of each relation. The relation set can be included similarity, time or space proximity, sequence etc.

We define an operator on graphs, Union, as follows:

Definition 4.2. Union(\cup): Let G_i and G_j be two relational graphs that present the relation between videos. The $G_i \cup G_j = \{G \mid V=V_i \triangle V_j, E= E_i \triangle E_j, R= R_i \triangle R_j, W=W(R)\}$, where \triangle is the operation of symmetric difference in logical algebra.

4.1 Duplicate-Based Integration

Generally, the relation view of one duplicate represents a faceted semantic relation of the duplicate in bigger space. So the duplicate-based integration can help to build a global relation schema on video sources. In view of the neighbours of duplicate may be duplicate, we can not simple merge these local relation views. We consider eliminating the common nodes which represent the same video in these views. Algorithm **IntegrateByDuplicate** describes the steps in integrating local relation view G_i and G_j . At a high-level we first detect all duplicates between G_i and G_j and update the names of nodes that represent the duplicates for name consistency. Then we consider the pre-processing of relations and weights for relation consistency. The pre-processing of relations is included the relation transform, such as: the video A was created on “2010.4.9” and B was created on “2010.4.19”, and there are exist the relation r_1 =“is same month” with $w=1-(1/3)$ in G_i . But in G_j there are used the relation r_2 =“is same year”, so for ensuring relation consistency in union view, r_1 can be transformed to r_2 with $w=1-(10/365)$. Note that, we transform relation from old relation in combining views, but do not delete the old relation.

Algorithm 1. IntegrateByDuplicate(G_i, G_j : View of duplicate)

```

1: Dset=DetectDuplicateNodes( $G_i, G_j$ );
2: for each duplicate do
    if there are duplicates between  $G_i$  and  $G_j$  then update the names of nodes that
    represent the duplicates to the same but different with other non-duplicates' nodes;
    end for
3: Preprocessing:
    if  $r_i(R_i) \subset r_j(R_j)$  or  $r_j(R_j) \subset r_i(R_i)$  then do relation transform; End if
    if  $r_i(R_i)$  is the same as  $r_j(R_j)$  and  $w_i(r_i) \neq w_j(r_j)$  then
         $w_{r_i} = w_{r_j} = (w_{r_i} + w_{r_j}) / 2$ ;
    End if
4:  $G = G_i \cup G_j$ ;
5: return  $G$ ;
```

Note that the algorithm **IntegrateByDuplicate** is the main idea that integrating two local relation views by near-duplicate. In the whole video collections, if there are multiple local relation views and duplicates between them, the algorithm **IntegrateBy-Duplicate** will be called repeatedly until there are not duplicates in all local relation views.

4.2 Context-Based Integration

Although no near duplicates in some local relation views, we observed that some of videos in different views will similar in semantics if their annotations such as tagging, description are very similar. In the paper, we take the annotation and comments of a

video as context of the video. On account of integrating these videos and their relevant videos can help to retrieve bigger relation view, the integration based on context is useful technique for our system. Algorithm **IntegrateByContext** summarizes that how we integrate the graph G_i and G_j if we find there are relations with highly weight W_{ij} between nodes $v_i (v_i \in G_i)$ and $v_j (v_j \in G_j)$.

Firstly, in algorithm 2, we integrate the duplicates by using the algorithm **IntegrateByDuplicate** for merging the same videos. Then deducing the relation type of videos (such as similarity, sequence) between G_i and G_j in which these videos have same attributes or keywords in the context. In the step of relation establishment, we establish the directional relation for sequential relation, and the similar relation with the similarity computing technique [16] to compute the similarity of video's annotation as the weight. Note that for determining what similarity of context between videos is considered to be integrated, we use a threshold of the similarity σ which can be set by user or system.

Algorithm 2. IntegrateByContext(G_i, G_j ; Graph)

1: Firstly, integrating by duplicates:

G = IntegrateByDuplicate(G_i, G_j);

2: Finding same attributes or keywords in the context of V_i and V_j

Deducing the relation type between V_i and V_j

If there are existing relation and not edge between V_i and V_j **then**

Generate a edge between V_i and V_j

End if

If there are sequential relation **then** Establish the directional relation r_{ij} ; **end if**

Else if there are similar relation **then**

Computing similarity of the values of attribute both in V_i and V_j on same attribute

If the similarity great than the value of a threshold that user set **then**

Establish the relation r_{ij} and assign the similarity to $w(r_{ij})$

End if

End if

3: **return G;**

4.3 Rules-Based Integration

The techniques introduced above can automatically integrate correlative videos by duplicate or context. There is one type of video integration which can not be integrated with obviously correlative relationship, but can be integrated with logic rules, such as constituent, time or space distance, etc. In general, there are mainly two classes of rules: (1) Numerical Rules that integrate a set of videos by a numerical bound; (2) Set Rules that integrate a set of videos by enumerative tags.

Definition 4.3. Numerical Rule. Let S be a set of videos, A be a set of common attributes of S , i.e. $A = \{a_1, a_2, a_3, \dots\}$, the simple rule $R_s = \{ \forall I S.a_i \otimes E \}$, where \otimes is one of $<$, $=$, $>$, E is a expression which limit the bound, and V is the videos that satisfied the rules.

As an example, Let A be a set of common attributes of a video collection S , such as load time (lt), length, and so on. The function $R_s = \{ \forall I S.lt > DATE \}$ integrates the set of all video that their load time late than the DATE.

Definition 4.4. Set Rule. Let S be a set of videos, A be a set of common attributes of S , i.e. $A=\{a_1, a_2, a_3, \dots\}$, the set rule $R_s=\{\forall I S.a_i \oplus T\}$, where \oplus is similarity operator, and T is a set of enumerative phrases.

It is easy to see that sometime there are not a video's content covered a subject, but a set of videos, so user can specify a set of sub subject names to query.

The algorithm **IntegrateByRules** describes how we integrate videos based on some rules that user given. At first, we use the traditional match algorithms e.g. Vector Space Match [17] to select the videos V that satisfied numerical rules and set rules, then generate edges for V and merge the graphs of these videos.

Algorithm 3. IntegrateByRules(R_s : specific rules; S : video set)

```

1:  $\forall r_i \in R_s ; G = \{\}$ ;
2: if  $r_i$  is Numerical Rule then
3:  $V = S.a_i \otimes E$ ;
4: else if  $r_i$  is Set Rule then
5:  $V = S.a_i \oplus T$ ;
6: end if
7: for  $v_i, v_j \in V$  do
8:  $G_i =$ the graph of  $v_i$ ;  $G_j =$ the graph of  $v_j$ ;
9:  $e_{ij} =$ edge between  $v_i$  and  $v_j$ ;
10:  $G = G \cup G_i \cup G_j$ ;
11:end for
12:return  $G$ ;

```

5 Global Relation Refinement

Using the algorithms of section 4, we can build the global relation schema on video collections, but the global relation schema is complex, and there are redundant and conflicted relations which will impede the faceted search seriously. In the section, we will consider the abstraction of nodes, redundant relation rectification and relation deduction for refining the global relation of video collections.

(1)**Nodes abstraction.** The name of some relations has implicitly declared that the nodes belong to one category or have same feature, such as “is same (time, color, etc)”, “is belong to (common command in computer networks, electronic commerce course, etc)”. So we can build an abstract node to link these nodes and use the same feature and category name as its name which shows in figure 2(a). To some abstract nodes if they be included a wider category or have common features, we can build an abstract nodes on these abstract nodes which show in figure 2(b).

(2)**Redundant identify.** Although in the integrating process, we try to integrate all duplicate videos or videos with same tags. But it is hard to keep the global relation schema with no redundancy. There are may be some relations with different their name but they are the same relation in real, so in the global relation schema we need to identify the redundant relations by the semantic analysis, such that synonyms, alias, etc.

(3)**Relation deduction.** Some relations between a set of videos have the transitive or symmetrical characteristics. For videos a, b, c , that is:

if $a \xrightarrow{r} b$ then $b \xrightarrow{r} a$ (symmetry);

if $a \xrightarrow{r} b$ and $b \xrightarrow{r} c$ then $a \xrightarrow{r} c$ (transitivity).

So we can deduce:

r is symmetrical relation and $a \xrightarrow{r} b \Rightarrow b \xrightarrow{r} a$;

r is transitive relation and $a \xrightarrow{r} b, b \xrightarrow{r} c \Rightarrow a \xrightarrow{r} c$.

By the relation deduction for transitive or symmetrical relations, we can complement the relations that implied in videos.

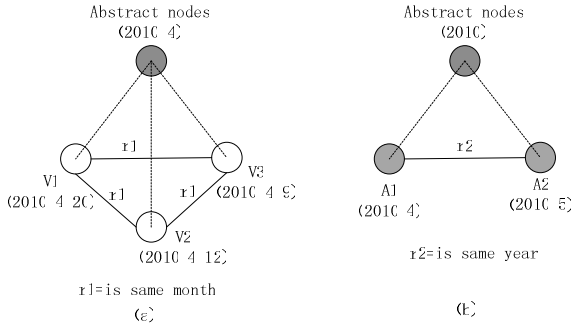


Fig. 2. Nodes abstraction

In short, through the nodes abstraction, redundant rectify and relation deduction, we will get a compact, clear and hierarchical global relation schema which will make the faceted search became effectively, quickly and completely.

6 Experiments

We conducted an experimental study the performance of the system. In the experiment, we mainly consider two integrating techniques: duplicate-based integration and context-based integration. The goal of the study was to understand the effect of our technique to integrate videos on heterogeneous video collections and the contributions of the various constituents in the system.

6.1 Experimental Setup

All the video set in our experiments are crawled by searching on the Google Video and Yahoo!Video. We select 80 popular keywords about “Computer Networks” topic as the queries (in Google Video we using Chinese keywords). For each query, we get 100 top-ranked videos and their corresponding web pages. We refer to the dataset from Google Video and Yahoo!Video as GV and YV respectively. We use the Road-Runner to extract local relation view from web pages. Then we use the two integrating techniques to build the global relation schema.

To measure the effectiveness of our techniques for automatic video integration, we perform video integrating to estimate the conciseness, completeness, integration gain

respectively and compared the video systems of Google, Yahoo in the following experiments by randomly select 10 keywords.

6.2 Effect of Extensional Conciseness

The conciseness measures the uniqueness of videos representations and boosting the video tagging, as well as the capability of eliminating copy, in video collections. Referred to [4], we define the extensional conciseness (EC) is the number of unique videos in a collection in relation to the overall number of video representations in the collection.

$$EC = \frac{\| \text{unique videos in video collection} \|}{\| \text{all videos in video collection} \|} = \frac{a}{a + b} \tag{1}$$

The example in the figure 3 shows the EC on the INTERVIDEO based on the experimental dataset of 10 keywords queries from GV and YV respectively. We observed that we can get EC=83.5% by our system. And further, we use the method of Nodes Abstraction (NA) to integrate all segments of videos, for example, using the “common command in computer networks” to representing the “part 1 of common command in computer networks” and the “part 2 of common command in computer networks”, we can reduce more the EC, which is displayed as NA on GV and YV respectively in figure 3.

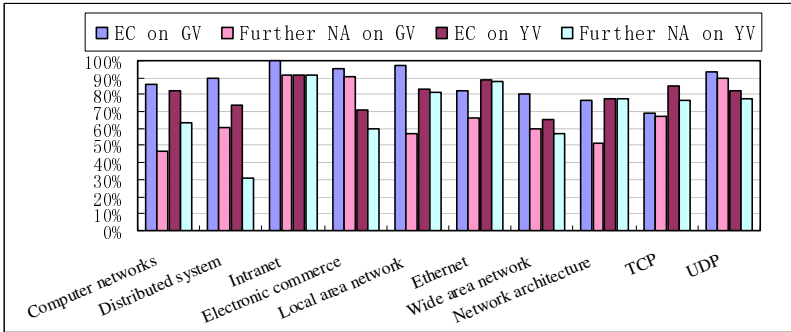


Fig. 3. Measuring the EC on the INTERVIDEO

6.3 Effect of Extensional Completeness

The extensional completeness (EP) is the number of unique video representations in a dataset in relation to the overall number of unique videos in the Web, such as, in all the sources of an integrated system, referred to [4]. It measures the percentage of videos, which in the Web, covered by that dataset. We assume that we are able to identify same videos on the Web, for example, by an identifier created during duplicate detection.

$$EP = \frac{\| \text{unique videos in video collection} \|}{\| \text{all unique videos in the Web} \|} = \frac{a}{a + c} \tag{2}$$

In order to evaluate the EP, we not only use the GV and YV but also retrieving the videos that queried by Google Web. If we consider only the “intense relevant videos”, which is meaning the videos belong to the semantic space of the keywords, we observed that the EP equal to or slightly larger than 1. Because these dataset most from popular video website (e.g. www.youtube.com), in which the relevant videos in same web page is queried by same keywords in most case. But if we take the videos that queried by Google Web as experimental dataset, we can get high EP which in general the value great more than 1, and in most case the value can get to 5~8. We observed that the relevant videos with the videos we queried same in the web page is pre-defined and with same topic in these case. Note that the results that all returned by system together with topics but no discrete and disorder.

6.4 Integration Gains

The integration gain (IG) is measuring average size of connected graph compared before and after video integrating. It evaluates the ability of interlinking with various semantic dimensions to a system.

$$IG = \frac{\text{average size of connected graph before integrating}}{\text{average size of connected graph after integrating}} \quad (3)$$

Generally, the videos queried by search engine are discrete and incomplete, and relevant videos are not linked. In our system, we can integrate the discrete videos with sorted and interlink to groups. The figure 4 shows the IG from our system with GV dataset, which has not been processed by global relation refinement. The results indicate that the results will be more semantic integration ability and comprehensive by our integrating techniques.

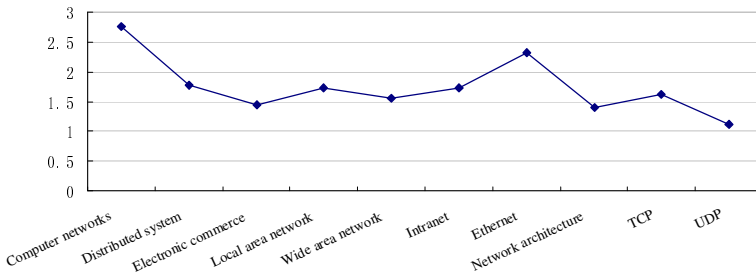


Fig. 4. The Integration Gains (IG) of 10 queries in GV

7 Conclusions

In this paper, we have proposed a novel video integration framework for faceted search on the Web. More specifically, in what is a novel hybrid approach, we have used near duplicates, correlative contexts and specified rules to build global relation schema over heterogeneous video collections. The global relation schema which involves various relations and rich knowledge of videos enables faceted search. Our

experiments show that the relevant videos fusion can largely improve concisely and completely structure and organization of content; our preliminary evaluation indicates an information gain and efficiency for videos searching. In the future, we plan to resolve the integration conflict, which include the schematic conflict, and data conflict, etc. We also plan to automatically generate faceted metadata based on the global relation schema to boost the query refinement or results presentation.

Acknowledgement

We are grateful to the National High-Tech Research and Development Plan of China under Grant No. 2008AA01Z203 for funding our research.

References

1. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: Proc. of the SIGCHI Conference on Human Factors in Computing Systems (2003)
2. Teevan, J., Dumais, S.T., Gutt, Z.: Challenges for Supporting Faceted Search in Large, Heterogeneous Corpora like the Web. In: Proceedings of HCIR (2008)
3. Barish, G., Shin Chen, Y., Dipasquo, D., Knoblock, C.A., et al.: Theaterloc: Using information integration technology to rapidly build virtual applications. In: ICDE (2000)
4. Bleiholder, J., Naumann, F.: Data fusion. *ACM Computing Surveys* 41(1) (December 2008)
5. Cao, J., Zhang, Y.D., et al.: VideoMap: An Interactive Video Retrieval System of MCG-ICT-CAS. In: CIVR 2009 (July 2009)
6. Christel, M.G., Yan, R.: Merging Storyboard Strategies and Automatic Retrieval for Improving Interactive Video Search. In: CIVR 2007 (July 2007)
7. Chang, K., He, B., Zhang, Z.: Toward large scale integration: Building a MetaQuerier over database on the web. In: CIDR (2005)
8. Madhavan, J., Jeffery, S.R., Cohen, S., et al.: Web-scale data integration: you can only afford to pay as you go. In: CIDR (2007)
9. Amer-Yahia, S., Lakshmanan, L., Yu, C.: SocialScope: Enabling Information Discovery on Social Content Sites [C]. In: CIDR (2009)
10. Crescenzi, V., Mecca, G., Merialdo, P.: RoadRunner: Towards automatic data extraction from large web wites. In: VLDB (2001)
11. Arasu, A., Molina, H.G.: Extracting structured data from Web pages. In: SIGMOD (2003)
12. Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: WWW (2005)
13. Hung, M., Zou, Y.: Recovering workflows from multi tiered e-commerce systems. In: 15th IEEE International Conference on Program Comprehension, ICPC 2007 (2007)
14. Wu, X., Hauptmann, A.G., Ngo, C.-W.: Practical elimination of near-duplicates from web video search. In: ACM Multimedia, MM 2007 (2007)
15. Siersdorfer, S., Pedro, J.S., Sanderson, M.: Automatic video tagging using content redundancy. In: SIGIR 2009, July 19-23 (2009)
16. Pedro, J.S., Dominguez, S.: Network-aware identification of video clip fragments. In: CIVR 2007, pp. 317–324. ACM Press, New York (2007)
17. Abbasi, R., Staab, S.: RichVSM: enRiched vector space models for folksonomies. In: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (2009)

An Efficient Data Indexing Approach on Hadoop Using Java Persistence API

Yang Lai^{1,2} and Shi ZhongZhi¹

¹ The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

² Graduate University of Chinese Academy of Sciences, Beijing 100039, China
{yanglai, shizz}@ics.ict.ac.cn

Abstract. Data indexing is common in data mining when working with high-dimensional, large-scale data sets. Hadoop, a cloud computing project using the MapReduce framework in Java, has become of significant interest in distributed data mining. To resolve problems of globalization, random-write and duration in Hadoop, a data indexing approach on Hadoop using the Java Persistence API (JPA) is elaborated in the implementation of a KD-tree algorithm on Hadoop. An improved intersection algorithm for distributed data indexing on Hadoop is proposed, it performs $O(M+\log N)$, and is suitable for occasions of multiple intersections. We compare the data indexing algorithm on open dataset and synthetic dataset in a modest cloud environment. The results show the algorithms are feasible in large-scale data mining.

Keywords: Data Indexing, KD-tree, Data Mining, Distributed applications, JPA, ORM, Distributed file systems, Cloud computing.

1 Introduction

Many approaches have been proposed for handling high-dimensional and large-scale data, in which query processing is the bottleneck [1]. Business intelligence and data warehouses can hold a Terabyte or more of data. Cloud computing has emerged for the subsequently increasing demands of data mining. MapReduce is a programming framework and an associated implementation for large data sets [2].

A concise indexing Hadoop implementation of MapReduce is presented in McCreadie's work [3]. Ralf proposes a basic program skeleton to underlie MapReduce computations [4]. Moretti presents an abstraction for scalable data mining, in which data and computation are distributed in a computing cloud with minimal user efforts [5]. Gillick uses Hadoop to implement query-based learning [6].

Most data mining algorithms are based on object-oriented programming (OOP), which runs in memory. Researchers elaborate many of these methods [7-10].

However, the following features in the MapReduce framework are unsuitable for data mining. First, in globality, map tasks are irrelevant to each other, as are reducing tasks. Data mining requires that all of the training data be converted into a global model, such as a KD-tree or clustering tree. The tasks in the MapReduce framework only handle its partition of the entire data set and output its results into the Hadoop distributed file

system (HDFS). Second, random-write operations are disallowed by the HDFS, thus disabling link-based data models in Hadoop, such as linked-lists, trees, and graphs. Finally, the duration of both map and reduce tasks are based on scanning processing, and will end when the partitioning of the training dataset is finished. Data mining requires a persistent model for following testing processing.

A database is an ideal persistent repository for objects generated by data mining using Hadoop tasks. A data mining framework on Hadoop using the Java Persistence API (JPA) and MySQL Cluster is proposed [11]. To mine high-dimensional and large-scale data on Hadoop, we employ Object-Relation Mapping (ORM), which stores objects whose size may surpass memory limits in a relational database. The Java Persistence API (JPA) provides a persistence model for ORM [12]. A distributed database is a suitable solution to ensure robustness in distributed handling. MySQL Cluster is designed to withstand any single point of failure [13], which is consistent with Hadoop.

We performed the same work that McCreddie's performed [3] and now propose a novel indexing Hadoop implementation for continuous values. Using JPA and MySQL Cluster on Hadoop, we propose an efficient data mining framework, which is elaborated by a KD-tree implementation. We also propose an improved intersection algorithm for the distributed data indexing on Hadoop [11], which can be suitable for many situations.

The rest of the paper is organized as follows. In Section 2 we elucidate the related index structures, flowchart and algorithms. Section 3 proposes the KD-tree indexing on Hadoop and the improved intersection algorithm. Section 4 provides descriptions of our experimental setting and results. In Section 5, we offer conclusions and suggest possible directions for future work.

2 Related Work

2.1 Naïve Data Indexing [11]

Data indexing is necessary for querying data in classification or clustering algorithms. Data indexing can dramatically decrease the complexity of querying.

2.1.1 Definition

A $(n \times m)$ dataset is an $(n \times m)$ matrix containing n rows of data, which contain m columns of float numbers. Let $\min(i)$, $\max(i)$, $\text{sum}(i)$, $\text{avg}(i)$, $\text{std}(i)$, $\text{cnt}(i)$ ($i \in [1..m]$) equal the minimum, maximum, sum, average, standard deviation and count of the i -column, respectively—i.e., the essential statistical measures of the i -column of the dataset. Let $\text{sum2}(i)$ be the dot product of the i -column, which is used for $\text{std}(i)$. From the

$\text{ZSCORE}(x) = \frac{x - \text{avg}()}{\text{std}()} [7]$, the formula $\text{YSCORE}(x) = \text{round}\left(\frac{(x - \text{avg}()) \times \text{iStep}}{\text{std}()}\right)$, is defined, where iStep is an experimental integral parameter.

2.1.2 Simple Design of the Inverted Index

In general, high-dimensional datasets are always stored as tables in a sophisticated DBMS for most data mining tasks. Figure 1 shows the flowchart in which the whole procedure is illustrated for indexing this kind of dataset and similarity querying [11]:

Discretization calculates the essential statistical measures, then performs YSCORE binning on the FloatData.TXT file (CSV format) and outputs the IntData.TXT file in which each line is labeled with a unique number (LineNo) consecutively.

The encoder transforms the text format file IntData.TXT into the Hadoop Sequence File format IntData.SEQ and generates IntData.DAT. To locate a record for a query rapidly, the length of the record has to be equivalent. The IntData.DAT contains an (n×m) matrix, i.e., n rows of records containing m columns of IDs of bins, just as does IntData.SEQ.

With the indexer, the index files Val2NO.DAT and Col2Val.DAT are outputted with MapReduce from the IntData.SEQ file. To obtain the list of LineNo’s for a specific bin, it is needed to store the list in a structural file (Val2NO.DAT); the address and the length of the list should be stored in another structural file (Col2Val.DAT). The two structural files are easy to access by offset.

The searcher, given a query—which should be a vector with m float values— performs YSCORE binning with the same statistical measures and yields a vector with m integers for searching; it then outputs a list of the number of records.

In intersection, the positions of each integer in Val2NO.DAT for a particular query will be extracted from Col2Val.DAT, and the lists of LineNo’s for each integer can be found. The result will be computed by the intersection of the lists. The algorithm for improved intersection is elucidated in (3.1).

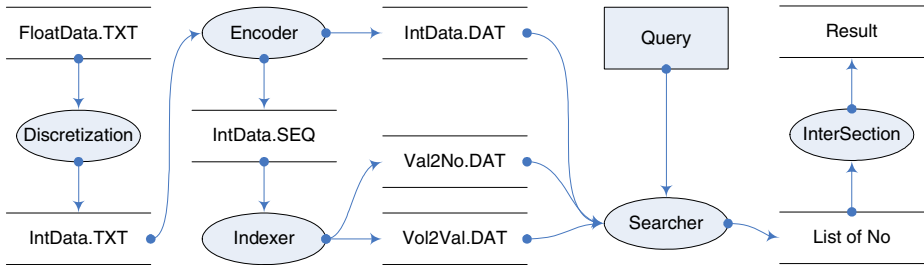


Fig. 1. Flowchart for indexing and searching on Hadoop

2.1.3 Statistics Phase

In Hadoop, data sets are scanned once to calculate the four distributive measures, $min(i)$, $max(i)$, $sum(i)$ and $sum2(i)$ as well as $cnt(i)$ for the i -column; it then computes the algebraic measures of average and deviation. Therefore, a (5×m) array is used for the keys in the MapReduce process: the first row contains the keys of the minimum of a column; the second row contains the keys for the maximum; the third contains the keys for the sum; the fourth contains the keys for the square of the sum; and the fifth contains the keys for the row count.

The algorithm performs $O(N/mapper)$, where N represents the number of records in the data sets, and $mapper$ is the number of the task of map functions. All partitions will have been calculated into the five distributive measures, the reduce function merges them into global measures. Keys for the map function are the positions of the next line, and values are the next line of text.

2.1.4 YSCORE Phase

In Figure 2, the maximum number in the bins reduces to a reasonable level, by increasing the *iStep*. The lowered numbers administer to inverted indexing.

These interval values are put into bins using YSCORE binning [7]. Then, the inverted index will be of the format: the first column holds all fields in the original data-set; the second holds the <bin, RecNoArray> tuples.

The YSCORE binning phase does not need a reduce function, thus eliminating the cost of sorting and transferring; the result files are precisely the partitions on which map functions work. Keys in the map function are the positions of the next line, and values are the next line of text. The input file will be FloatDate.TXT, and the output file will be IntData.TXT, as shown in Figure 2. After the statistics' phase and the YSCORE binning phase, Hadoop can handle discrete data as in the commonly-used MapReduce word count example.

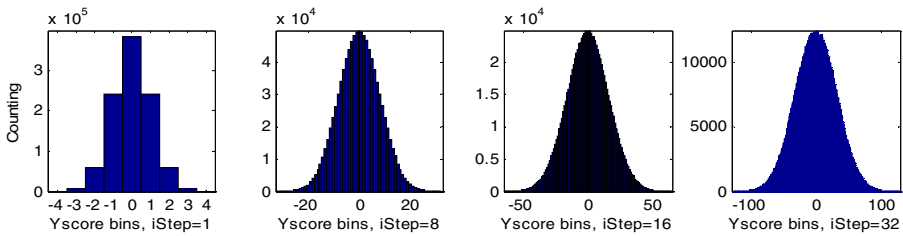


Fig. 2. YSCORE binning of 1,000,000 values from a normal distribution with mean 0 and standard deviation 1. These YSCORE binnings use *iStep* of 1, 8, 16, and 32, respectively.

2.2 Data Mining Framework on Hadoop [11]

To apply common object-oriented data mining algorithms in Hadoop, the three problems concerning globalization, random-write and duration, posed in the introduction, should be handled. For data mining with Hadoop, a database was an ideal persistent repository for handling the three problems mentioned. In the data mining framework, a large-scale dataset is scanned by Hadoop. The necessary information is extracted using classical data mining modules, which will be embedded in the Hadoop map or reduce tasks. The extracted objects, such as lists, trees, and graphs, will be sent automatically into a database—in this case, MySQL Cluster.

3 Method

3.1 Improved Intersection

Searching in the naïve data indexing is based on intersection of multiple index files [11]. In general, the lengths of each index file are not the same order of magnitude. We propose an improved algorithm for this kind of unbalanced intersection.

The length of an intersection is always less than or equal to the two candidates. Figure 3 shows an example of intersection of ten arrays. Let δ be the intersection coefficient; Let x be the length of the first candidate in Figure 3. Therefore the lengths of

intersections of multi-arrays are proximately equal to a geometric progression: $x, x\delta \dots, x\delta^{n-1}$. For best performance, the arrays should be in descending order; the x should be the shortest. If the number of arrays is large enough, the length of the final intersection will be significant small.

The algorithm first checks whether the proportion between the two candidates is beyond a threshold $iTimesLength$. If no then a normal intersection turns out, else the improved intersection will be performed. The proportion determine which array is larger to be applied in binary search. All the elements in the smaller will be scanned, while only a few elements in the larger will be checked. Let m be the length of the smaller, n for the larger. The complexity is $O(m+\log(n))$. If the lengths of the two candidates are not the same order of magnitude, the algorithm will perform better significantly. In fact, high-dimensional datasets always give rise to unbalanced intersection on Hadoop.

Though the algorithm performs better than the naïve data indexing, a flaw is that the procedure of intersection can not be paralleled to MapReduce model effectively.

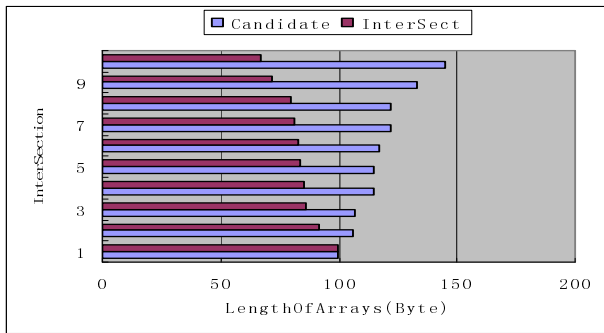


Fig. 3. Abstract of improved intersection between multi-Arrays. The blue lower bars are the original arrays in ascend order for optimal complexity. The red upper bars are the results of intersection. The 10th red bar means the final intersection of the ten arrays.

3.2 KD-Tree on Hadoop

KD-tree is a traditional data indexing algorithm. Bentley [9] discusses it in detail. The naïve KD-tree algorithm divides high-dimensional dataset on each dimension according to the split value in a defined criterion (Figure 4 a).

3.2.1 Hierarchical Bucket

This algorithm handles large-scale datasets on Hadoop, applies the naïve KD-tree algorithm to small datasets under the limit of memory (Figure 4). Let $iMemoryLimit$ be the limit of memory, a KD-tree will be divided into 4 levels, bucket 0, 1, 2, 3. The triangles denote bucket-1, which just below the $iMemoryLimit$. The bucket-0 means the internal nodes beyond the triangles, which denote the sub dataset whose size is too large to fit in memory. The bucket-2 and bucket-3 are not shown in Figure 4, which are exactly the in-memory nodes described by Bentley [14].

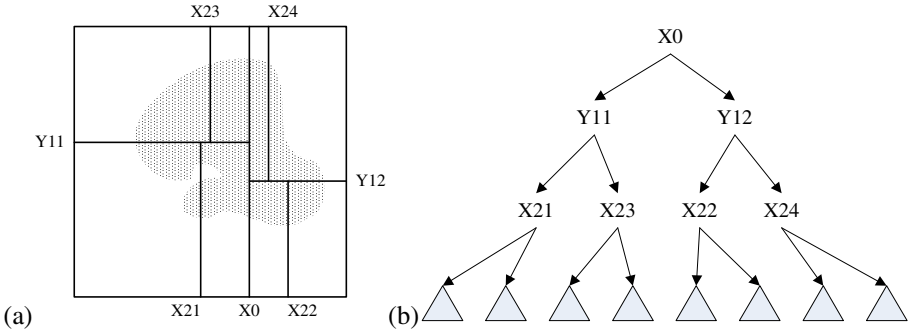


Fig. 4. Dataset in 2-space stored in JPA. The dotted area in (a) denotes a dataset. The labels denote the split points. The tree in (b) denotes the 2-d tree in JPA generated from the dataset in (a). The triangles denote the basic blocks of the dataset, which are the directories in HDFS.

Let R be the ratio of dataset to `iMemoryLimit`, the number of levels for bucket-1 is $\lceil \log_2 R \rceil$, and the number of the internal bucket-0 nodes, i.e. the number of splitting on Hadoop, is $2^{\lceil \log_2 R \rceil}$, approximately R . A node in KD-tree using JPA can be represented by the following MySQL commands:

```
CREATE TABLE `KD-tree` (
  `id` INT(11) PRIMARY KEY, `cnt` INT(11), `mean` INT(11),
  `std` INT(11), `attribute` INT(11), `median` FLOAT,
  `childleft` INT(11), `childright` INT(11),
  `bucket` TINYINT(3), `block` VARCHAR(1000));
```

Where ‘`id`’ is the primary key for each node, i.e. the reference for an object in OOP; ‘`attribute`’ is to tell which column should be concerned; ‘`median`’ is the balanced split-point; ‘`bucket`’ is to show what type of the node; ‘`block`’ is important for Hadoop by providing the HDFS path of the dataset represented by the node.

3.2.2 Median Selecting

A balanced KD-tree needs the median of each subset. As shown in the textbook [7], median is a holistic measure that is notable in complexity. However, there is a method to approximate it in $O(n)$. Let D be a dataset, D is grouped in bins according to an attribute and the frequency of each bin is known. The approximate median of the attribute is given by

$$median = L_m + \frac{N \times 0.5 - \sum f_{less}}{f_m} \times width \tag{1}$$

Where $median$ is the median of a field and L_m is the lower boundary of the bin in which the $median$ is; f_m is the frequency of the bin in which the $median$ is; f_{less} is the frequency of the bin in which all values are below the $median$; N is the size of D ; $width$ is the width of a bin. [7]

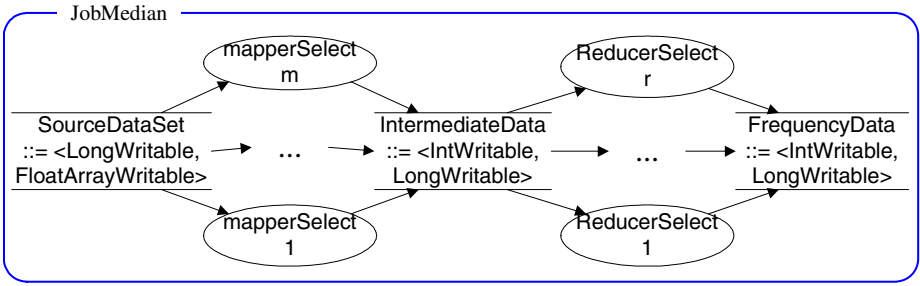


Fig. 5. The Data Flow Diagram of Median Selecting in KD-tree on Hadoop

In Figure 5, JobMedian will scan a dataset once and the YSCORE (2.1.4) can be applied in this situation. The $avg(i)$ and $std(i)$ of each subset will be ready for the best attribute, then the subset will be grouped in bins whose width is $std(i)/\Delta Step$ in $mapperSelect(1..m)$. The frequencies of the bins will be similar to Figure 2. Finally, one $reduceSelect$ collects the numbers from each map task, and outputs the median V_0 of the dataset according to formula (2). The notable thing is the $mapperSelect$ s handle the huge dataset once, after then the $reduceSelect$ calculate a few numbers. The sorting overhead is trivial compared with scanning. In JobMedian, only read operations occur with full speed.

3.2.3 Splitting

JobSplit contains only $mapperSplit$ tasks, which do not yield normal HDFS output using the method $output.collect()$ (Figure 6); instead, it writes directly to two sequence files in the HDFS to eliminate unwanted sorting I/O overheads in reduce task. The SourceDataSet is D ; the ChildDataLeft is D_1 ; the ChildDataRight is D_2 ; and all of these are, in fact directories in HDFS, which contain the outputs from each $mapperSplit$ task.

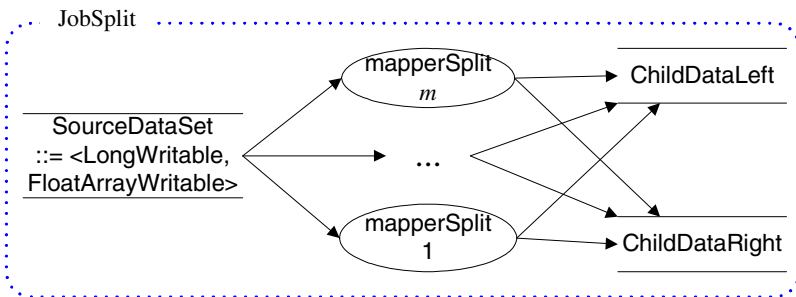


Fig. 6. The Data Flow Diagram of Split of KD-tree on Hadoop

Each node in the tree has to contain a member ATTRIBUTE and a member MEDIAN. Given a test record (an array of continuous values), the former tells which field should be checked, and the latter tells which child node should be selected. If the test record is less than or equal to the member VALUE in the member ATTRIBUTE, then the left child node is selected, otherwise the right child node is selected.

As shown in Figure 4, the internal nodes in KD-tree will be persisted in database use JPA, and external nodes (denoted by triangles) will be directories in HDFS.

The statistics algorithm runs in JobSplit on both sub dataset. The results $sum(i)$, $sum2(i)$ and $cnt(i)$ of each attributes of each sub dataset on each mapper node will be sent by JPA to calculate $avg(i)$, $std(i)$, the attribute with the largest $std(i)$ will be selected as the best.

Before splitting, a node in JPA will be created corresponding to D, with the best attribute A_0 and best splitting value V_0 . While scanning the dataset D, each mapper yields two separate files to HDFS, read and write operations occur.

3.2.4 Merging

By default, each mapper handles a segment of a file no more than 64MB, the number of split files may increase exponentially after JobSplit. Instead of using normal Hadoop Path as input, the list of Paths will be used as input of Mappers. This makes each mapper to scan multiple files, not a single 64MB block. Meanwhile, the average size of files will be checked, if files are all small enough, JobMerge will merge them into one file on each mapper node. The transfer rate of data will be close to the extreme.

3.2.5 Tree Building

The above processing is just for one node. To build a whole KD-tree, the processing needs to be repeated recursively. A complete KD-tree Hadoop algorithm is shown:

FUNCTION BuildTreeDFS (Path pD)

1. $rootD.Path \leftarrow pD$, JPA.persist($rootD$), $Q \leftarrow rootD$;
2. **WHILE** Q is not empty **DO**
 - a. $nodeD \leftarrow Q.removeFirst()$, JobMerge($nodeD$);
 - b. **IF** $nodeD.cnt$ less than $iMemoryLimit$ **THEN**
 - 1) JPA.begin(), $nodeD.bucket \leftarrow 0$; JPA.commit();
 - c. **ELSE**
 - 1) JPA.begin(), $nodeD.bucket \leftarrow 1$, JPA.commit();
 - 2) JPA.begin(), JobMedian($nodeD$), JPA.commit();
 - 3) $nodeLeft, nodeRight \leftarrow JobSplit(nodeD)$;
 - 4) JPA.persist($nodeLeft$), JPA.persist($nodeRight$);
 - 5) $Q.addLast(nodeLeft)$, $Q.addLast(nodeRight)$;
 - d. **ENDIF**
3. **ENDWHILE**

Where JPA.persist() function means a new node is persisted by JPA; the assignments enclosed by JPA.begin() and JPA.commit() mean these operations are monitored by JPA; JobMerge(), JobMedian(), JobSplit() are mapreduce jobs in Hadoop, which will be distributed to many nodes in a cluster.

4 Experimental Setup

4.1 Computing Environment

Experiments were conducted using Linux CentOS 5.0, Hadoop 0.19.1, JDK 1.6, a 1Gbps switch network, and 4 ASUS RS100-X5 servers (CPU: Dual Core 2GHz, Cache: 1024 KB, Memory: 3.4GB, NIC: 1Gbps, Hard Disk: 1TB).

4.2 Performance of Inverted Indexing

We evaluated several synthetic datasets and an open dataset. Two different-sized experiments are shown in Table 1. A $(1,000,000 \times 200)$ synthetic data set was generated at random for test1. A $(102,294 \times 117)$ data set was taken from KDDCUP 2008 [15] for test2. The procedure was followed according to the design (see 2.1.2). Of both tests, the latter takes fewer seconds for searching than the former. The SearcherTime is the average for a single query.

Table 1. Comparison of synthetic and open dataset

Item	Test1	Test2
FloatData(MB)	1,614	204
DiscretizationTime(s)	180	59
EncoderTime(s)	482	42
IndexerTime(s)	8473	397
SearcherTime(s)	1.07	0.68

4.3 Improved Intersection

In Figure 7, the longer candidate has $1E7$ sorted random integers; the x coordinate gives the proportions of the shorter candidate to $1E7$. The red line (normal) shows a constant higher time complexity between unbalanced arrays; the yellow line (improved) shows the time decrease dramatically along with the proportion of the shorter to the longer. The improved algorithm performs $O(M+\log N)$.

We also find the critical proportion is about 200 for the parameter `iTimesLength` in the improved intersection algorithm. Because binary searches arose many cache miss and sequential searches make use of cache hit [16] [17], the algorithm has no advantage in balanced intersections with proportion below 200.

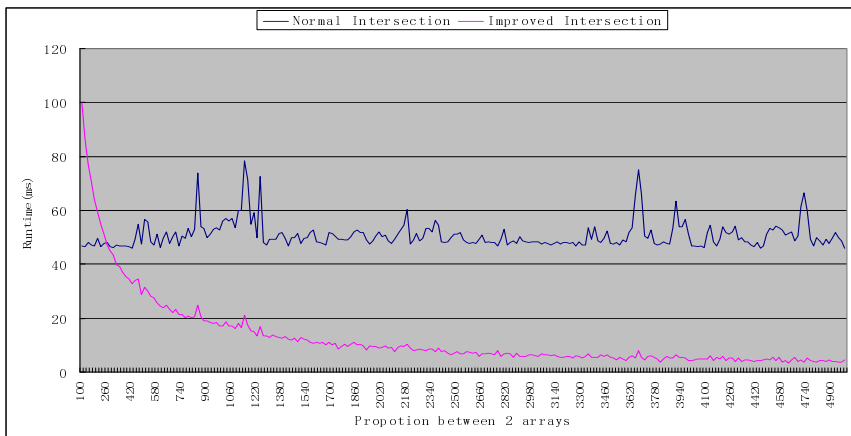


Fig. 7. Runtime comparison between normal and improved Intersections. The X coordinate means the proportion between the two arrays, by which the length of one array will become shorter; while the other remains $1E7$ elements.

4.4 Data Indexing

The essential settings in the algorithm are listed in **Table 2**. The setting `dfs.replication` should not be 1, which will ignore the redundancy and the network transfer rate.

Table 2. Main setting. The item ^{a, b, c} are the setting in Hadoop, the item ^{d, e} is in algorithm

Item	Transfer rate
<code>io.file.buffer.size</code> ^a	65536
<code>dfs.replication</code> ^b	2
<code>mapred.child.java.opts</code> ^c	<code>-Xmx800m</code>
<code>iMemoryLimit</code> ^d	1024*1024*1024
<code>iBucketLimit</code> ^e	1024*64

The data shown in Table 3 are average from multiple results. The data will be different on the situation of failure of a Hadoop cluster. The size of bucket-1 nodes in the KD-tree is just smaller than `iMemoryLimit`, it ensures that the naïve KD-tree algorithm can split the bucket-1 nodes into the bucket-3 nodes (in-memory nodes). The “time cost of bucket-1” denotes the cost of `JobMedian`, `JobSplit`, and `JobMerge` on the files and directories on HDFS. The “time cost of bucket-3” denotes the cost on memory.

Table 3. Comparison between datasets with different size in building KD-tree on Hadoop. All the data is from synthesis algorithm, with 117 fields, except `KDDCUP`.

Item	Time cost of bucket-1	Time cost of bucket-3
Build-KDDCUP	159 seconds	39 second
Build-1GB	313 seconds	51 second
Build-6GB	1963 seconds	243 seconds
Build-32GB	30010 seconds	2703 seconds

The transfer rates (Table 4) describe clearly the performance of data handling in the data mining framework on Hadoop is close to the extreme of the rack of 4 servers.

Table 4. Comparison of transfer rate. The item ^{a, b} has only read operation, the item ^c has R/W operation. The item ^d means file copy between two nodes using the linux command `scp`; the item ^e means the direct NIC transfer rate by UDP packet.

Item	Transfer rate
<code>JobMerge</code> ^a	59.7MB/sec
<code>JobMedian</code> ^b	64.3MB/sec
<code>JobSplit</code> ^c	24.6MB/sec
<code>Net copy</code> ^d	38.0MB/sec
<code>UDP transfer</code> ^e	120.0MB/sec

Finally, the engine type of MySQL is `MYISAM` or `NDBcluster`. The difference between them is trivial in the large-scale dataset, for time costs associate with “time cost of bucket-1” (Table 3). A strange error will occur with engine type of `InnoDB`.

The codes will be open at a cloud-based open source site, JavaForge, the SVN address is <http://svn.javaforge.com/svn/HadoopJPA>.

5 Conclusion

An efficient high-dimension large-scale data mining framework is proposed by the implementation of KD-tree algorithm on Hadoop; it employs the JPA and MySQL Cluster to resolve problems of globalization, random-write and duration in Hadoop. Experimental results show that its performances reach the peak of the transfer rate in our environment and it is technically feasible.

An improved intersection algorithm is proposed to enhance the naïve data indexing approach in paper [11]. Experimental results show that the algorithm performs better. The improved intersection is generic for other situations.

We will consider how to implement more tree-based algorithms with JPA, and improve our inverted indexing in future work in an effort to enhance the performance in larger-scale data mining. A convenient tool for the data mining framework will be a focus of future efforts.

Acknowledgements

This work is supported by the National Basic Research Priorities Programme (No. 2007CB311004) and National Science Foundation of China (No.60775035, 60903141, 60933004,60970088), National Science and Technology Support Plan (No. 2006BAC08B06).

References

1. Bohm, C., et al.: Multidimensional index structures in relational databases. In: Mohania, M., Tjoa, A.M. (eds.) DaWaK 1999. LNCS, vol. 1676, Springer, Heidelberg (1999)
2. Dean, J., Ghemawat, S., Usenix: MapReduce: Simplified data processing on large clusters. In: 6th Symposium on Operating Systems Design and Implementation (OSDI 2004), San Francisco, CA,
3. McCreadie, R.M.C., Macdonald, C., Ounis, I.: On Single-Pass Indexing with MapReduce. In: Sanderson, M., et al. (eds.) Proceedings 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 742–743. Assoc. Computing Machinery, New York (2009)
4. Lammel, R.: Google's MapReduce programming model - Revisited. *Science of Computer Programming* 70(1), 1–30 (2008)
5. Moretti, C., et al.: Scaling Up Classifiers to Cloud Computers. In: IEEE International Conference on Data Mining, Pisa, Italy (2008), <http://icdm08.isti.cnr.it/Paper-Submissions/32/accepted-papers>
6. Gillick, D., Faria, A., DeNero, J.: MapReduce: Distributed Computing for Machine Learning (2006), http://www.icsi.berkeley.edu/~arlo/publications/gillick_cs262a_proj.pdf

7. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. In: Jim Gray, M.R. (ed.) The Morgan Kaufmann Series in Data Management Systems, 2nd edn., Morgan Kaufmann, San Francisco (2006)
8. Berchtold, S., Keim, D.A., Kriegel, H.P.: The X-tree: An Index Structure for High-Dimensional Data. Readings in Multimedia Computing and Networking (2001)
9. Bentley, J.L.: Multidimensional binary search trees used for associative searching. Communications of the ACM 18(9), 509–517 (1975)
10. Arya, S., et al.: Approximate Nearest Neighbor Queries in Fixed Dimensions. In: 4th Annual ACM-SIAM Symp. on Discrete Algorithms, SIAM, Austin (1993)
11. Yang, L., Shi, Z.: An Efficient Data Mining Framework on Hadoop using Java Persistence API. In: The 10th IEEE International Conference on Computer and Information Technology (CIT-2010), Bradford, UK (2010)
12. Biswas, R., Ort, E.: Java Persistence API - A Simpler Programming Model for Entity Persistence (2009),
<http://java.sun.com/developer/technicalArticles/J2EE/jpa/>
13. Hinz, S., et al.: MySQL Cluster (2009),
<http://dev.mysql.com/doc/refman/5.0/en/mysql-cluster-overview.html>
14. Bentley, J.L.: K-d trees for semidynamic point sets. In: Proceedings of the Sixth Annual Symposium on Computational Geometry. ACM, New York (1990)
15. Siemens Medical Solutions, USA, kddcup data (2008),
<http://www.kddcup2008.com/KDDsite/Data.htm>
16. Lam, M.S., Rothberg, E.E., Wolf, M.E.: The Cache Performance and Optimizations of Blocked Algorithms. In: 4th International Conf. on Architectural Support for Programming Languages and Operating Systems. Assoc. Computing Machinery, Santa Clara (1991)
17. Przybylski, S.A.: Cache and Memory Hierarchy Design: A Performance Directed Approach. Morgan Kaufmann, San Francisco (1990)

Knowledge Engineering for Non-engineers

Tatiana Gavrilova

Abstract. This paper presents one approach for the innovative organization training for business analysts in developing enterprise ontologies. The underlying teaching framework is pursuing a methodology that will aid the process of knowledge structuring and practical ontology design, with emphasis on visual techniques. The described approach may be helpful for those companies which are interested in the practical knowledge management and need skillful knowledge workers. The paper proposes some new ideas of practical ontology design and evaluation and may be interesting for the knowledge engineering research and practising community.

Keywords: Knowledge engineering, learning, thinking, analyst training, ontology design.

1 Introduction

During the last decade, knowledge has become a key consideration in our economies and it is heavily associated with learning and innovation. Central problems for supporting all phases of knowledge processing are the productivity of the knowledge workers and the effectiveness of the usage of the special professional techniques. These techniques and models help to elicit, structure and integrate various knowledge patterns within and across enterprises. Knowledge work deals with analyzing and structuring in general. Top managers and IT analysts are continually challenged by the need to analyze massive volumes and varieties of multilingual and multimedia data. This situation is not limited to e-business, but is seen in nearly all companies and institutions. Knowledge base of a company can be operationalized, both in terms of measurement and by providing simulation models (Leydesdorff, 2006). Special interest to knowledge work is paid in the virtual and open organizations.

Company staff and employees require support and guidelines for knowledge sharing about information analysis, theories, methodologies and tools. Knowledge management (KM) is one of the powerful approaches to solve these problems in new information age with huge information overload and sophistication (Firestone and McElroy, 2005). Sophistication needs professionals. Professional knowledge analysts are still very rare on human resources market. Unfortunately, they also differ considerably in both backgrounds and cognitive styles (Wiig and Wiig, 1999).

Knowledge Engineering (KE) traditionally emphasizes and develops a range of techniques and tools including knowledge acquisition, conceptual structuring and representation models (Scott et al, 1994; Firestone, 2003). But for practitioners as enterprise analysts it is still a rather new, eclectic domain that draws upon areas like cognitive science. Accordingly, knowledge engineering has been, and still is, in danger from fragmentation, incoherence and superficiality. Still few universities deliver courses in practical knowledge engineering.

This paper describes recent experience in such training for some Russian subsidiaries of the international companies (British-American Tobacco, Siemens Business Services, etc.). The total number of trainees that received certificates of knowledge analysts is more than 60.

Theoretical part of the Training on Knowledge Engineering (TKE) is based on university courses in intelligent-systems development, cognitive sciences, user modeling and human-computer interaction delivered by author in 1992-2008 at the University of Pittsburgh (USA), University of Milano (Italy), University of Espoo EVTEK (Finland), Tartu University (Estonia), First Independent University of Warsaw (Poland) and Saint-Petersbutg State University (Russia). TKE proposes information structuring multi-disciplinary methodology, including the principles, practices, issues, methods, techniques involved with the knowledge elicitation, structuring and formalizing. Emphasis is put not on the technologies and tools, but in the training of analytical skills. Ontological Engineering is a further development of knowledge engineering towards ontology design and creating.

2 Knowledge Analysts Training Outline and Organization

The discipline of Knowledge Engineering traditionally emphasized and rapidly developed a range of techniques and tools including knowledge acquisition, conceptual structuring and representation models. These developments have underpinned an emerging methodology that can bridge the gap between the ability of the human brain to structure and store knowledge, and the knowledge engineers' ability to model this process. But for practitioners, knowledge engineering is still a rather new, eclectic domain that draws upon a wide range of areas, including cognitive science, etc. Accordingly, knowledge engineering has been, and still is, in danger of fragmentation, incoherence and superficiality.

Since 2000, a major interest of researchers has focused on building customized tools that aid in the process of knowledge capture and structuring. Trainees are introduced to major issues in the field and to the role of the knowledge analyst in strategic information system development. We include a lot of interdisciplinary knowledge elicitation and structuring methods that can help the knowledge work, such as the conducting unstructured interview, mastering the verbal reports, business process modelling techniques, road mapping, brainstorming, etc.

The future analysts gain the deep understanding the role of knowledge engineering and knowledge management in companies and organizations; in decision-making by members of an organization; in developing information framework. They study and are trained in practical methods mainly by doing. Attention is given both to developing inter-personal information communication skills and analytical cognitive creative abilities. The first module is targeted at essentials of informal mental modeling by presenting mind maps, concept maps, semantic networks, frames, decision tables, decision trees and other visual forms of knowledge pattern representation.

The training features short lectures, discussions, tests, quizzes and exercises. Lectures are important but the emphasis is put on learning through discussions, simulation, special games, training and case studies. A good deal of the course focuses on auto-reflection and auto-formalizing of knowledge, training of analytical and communicative abilities, discovery, creativity, cognitive styles features, and gaining new insights.

On-the-job or workplace training adds the value of the team spirit and entrained feeling. All the examples are taken just from the every day routine practice. Such approach enables the trainer for better tailoring the course to the specified needs of the company.

Normally the TKE course consists of 4 inter-related modules:

- Getting Started in KE (12 hours),
- Practical KE in depth (12 hours),
- Ontological Engineering (12 hours),
- Business Processes Modeling and mapping (12 hours).

Different combination of sub-topics is possible. Fig.1 illustrates the structure of one variant chosen by Business Engineering Group Company (Saint-Petersburg, Russia).

The main difference of TKE to existing methodologies is cognitive (not technological) bias. The topics of exercises cover categorization, observation, laddering, lateral thinking and other problem solving cognitive methods. Knowledge workers often under-value the significance of psychological background of categorization, laddering and lateral thinking. But during training some of them feel “insight” and become very enthusiastic. We try to implement the ontological approach into the teaching style and strategy. Philosophers of science define ontologism by postulating existence of the systemic hierarchical conceptual specification of any complex object.

Now ontologies help to support knowledge navigation, search and retrieval. They are also used in educational and business research (Blanchard, Mizoguchi, Lajoie, 2009; Dicheva et al, 2005). The practical knowledge workers often underestimate the impact of their cognitive styles on decision making procedure. Their verbal skills and logics really influence the information processing. It is supposed to be guided by common sense while it needs to be taught and trained.

From organizational point of view the training process consists of series of on-the-job sessions. One group was never not more than 8 persons. Each day classes do not last more then 3-4 hours including the hand-on computer practice in mind-mapping and concept mapping techniques.

3 Teaching Ontological Thinking and Design

Ontologies can be used to describe any business world. But our experience in training shows that nobody can deal with ontologies without knowledge engineering practice. How to teach ontology design? The theory differs from practical need. There are numerous well-known definitions of this milestone term (Gruber, 1993; Guarino and Jiarretta, 1998; Jasper and Uschold, 1999; Mizogushi and Bourdeau, 2000; Neches, 1991) but they may be generalized as “Ontology is a hierarchically structured set of terms for describing an arbitrary domain” (Gomez-Perez et al., 2004). In other words “ontologies are nothing but making knowledge explicit” (Guarino and Welty, 2000).

Since 2000 a major interest of researchers focuses on building customized tools that aid in the process of knowledge capture and structuring. This new generation of tools – such as Protégé, OntoEdit, and OilEd - is concerned with visual knowledge mapping that facilitates knowledge sharing and reuse. The problem of iconic representation has been partially solved by developing knowledge repositories and ontology servers where

reusable static domain knowledge is stored. But practitioners from companies and research centres still need simple and constructive algorithms for their activity.

Ontology creating also faces the knowledge acquisition bottleneck problem. The ontology developer encounters the additional problem of not having sufficiently tested practical methodologies, which would recommend what activities to perform.

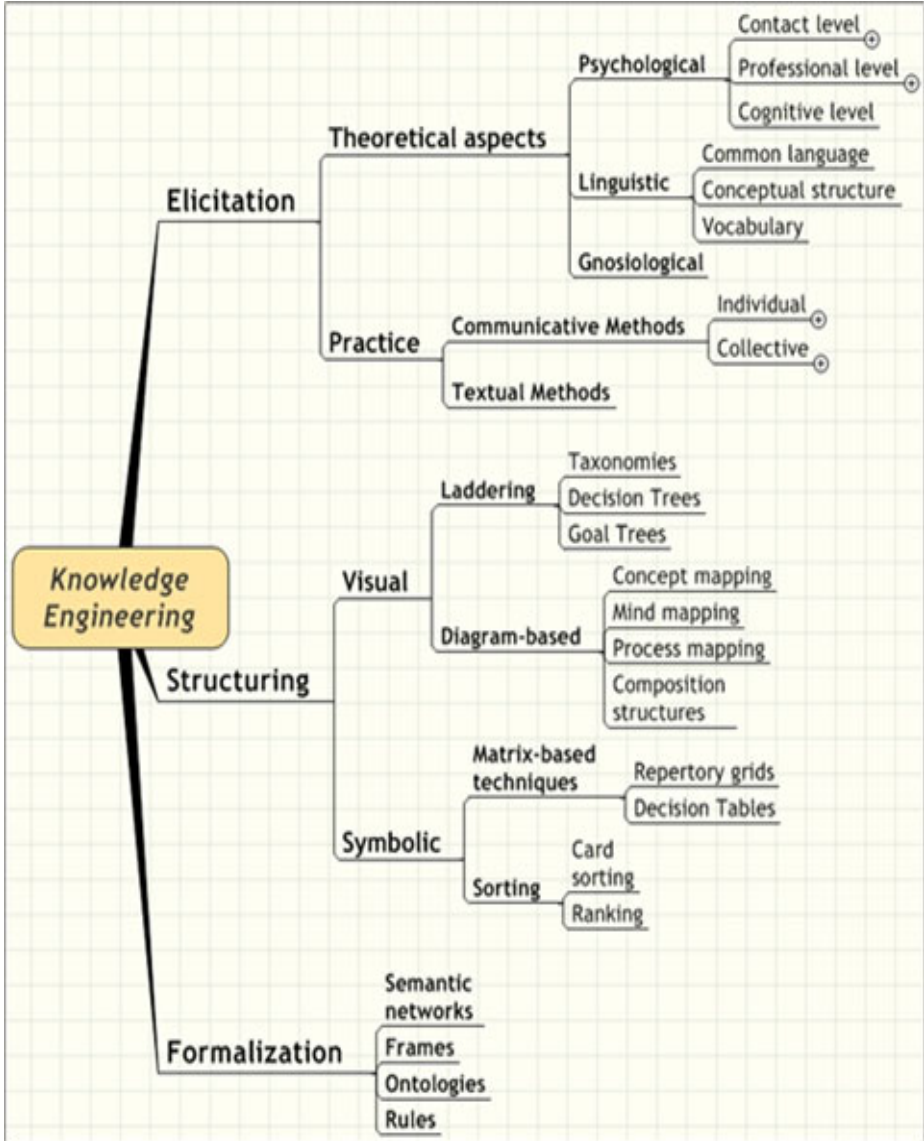


Fig. 1. Outline of training on knowledge engineering

An example of this can be seen when each development team usually follows their own set of principles, design criteria, and steps in the ontology development process. The lack of structured guidelines hinders the development of shared and consensual ontologies within and between the teams. Moreover, it makes the extension of a given ontology by others, its reuse in other ontologies, and final applications difficult (Guarino and Giaretta, 1998; Guarino and Welty, 2000; Jasper and Uschold, 1999).

Several effective methodological approaches have been reported for building ontologies (Swartout, et al 1997; Mizogusgi and Bordeau, 2000; Fridman Noy, Griffin, Musen, 2008).). What they have in common is that they start from the identification of the purpose of the ontology and the needs for the domain knowledge acquisition. However, having acquired a significant amount of knowledge, major researchers propose a formal language expressing the idea as a set of intermediate representations and then generating the ontology using translators. These representations bridge the gap between how people see a domain and the languages in which ontologies are formalized. The conceptual models are implicit in the implementation codes. A re-engineering process is usually required to make the conceptual models explicit.

The idea of using visual structuring of information to improve the quality of user learning and understanding is not new. Concept mapping has been used for more than twenty years (Sowa, 1984; Conlon, 1997; Jonassen, 1998) in system design and development for providing structures and mental models to support the knowledge sharing process. As such, the visual representation of general corporate business concepts facilitates and supports company personnel understanding both substantive and syntactic knowledge. An analyst serves as a knowledge engineer by making the skeleton of the company's data and knowledge visible, and showing the domain's conceptual structure. We try to simplify a bunch of different approaches, terms and notations for practical use and dare to propose a 4-steps recipe for practical ontology design.

3.1 Ontology Design Recipe

The existing methodologies describing ontology life cycle (Uschold and Gruninger, 1996; Mizoguchi and Bourdeau, 2000; Gomez-Perez et al, 2008; Noy, Musen, 2008) deal with general phases and sometimes don't discover the design process in details. Four simple practical steps were proposed in the training course.

Step 1. Glossary development: The first step should be devoted to gathering all the information relevant to the described domain. The main goal of this step is selecting and verbalizing all the essential objects and concepts in the domain.

Step 2. Laddering: Having all the essential objects and concepts of the domain in hand, the next step is to define the main levels of abstraction. It is also important to elucidate the type of ontology according to ontology classification, such as taxonomy, partonomy, or genealogy.

This is being done at this step since it affects the next stages of the design. Consequently, the high level hierarchies among the concepts should be revealed and the hierarchy should be represented visually on the defined levels.

Step3. Disintegration and Categorization: the main goal of this step is breaking high level concepts, built in the previous step, into a set of detailed ones where it is needed. This could be done via a top-down strategy trying to break the high level concept from the root of previously built hierarchy. At the same stage, detailed concepts are revealed in a structured hierarchy and the main goal at this stage is generalization via bottom-up structuring strategy. This could be done by associating similar concepts to create meta-concepts from leaves of the aforementioned hierarchy.

Step 4. Refinement: The final step is devoted to updating the visual structure by excluding the excessiveness, synonymy, and contradictions. As mentioned before, the main goal of the final step is try to create a beautiful ontology. We believe what makes ontology beautiful is harmony.

Using these tips the trainees developed several huge company ontologies (Gavrilova, Laird, 2005).

3.2 “Beatification” of Business Ontology

The idea of the good shape in modelling is rather common in science. Let’s try to apply this approach to the ontology design. One of substantial impulse to it was given by German psychological school of M. Wertheimer. His idea of good Gestalt (image or pattern) may be transferred into ontological engineering design guidelines. Some essential Gestalt principles of this school (Wertheimer, 1959):

- Law of Pragnanz (M. Wertheimer) - organization of any structure in nature or cognition will be as good (regular, complete, balanced, or symmetrical) as the prevailing conditions allow (law of good shape).
- Law of Proximity – objects or stimuli that are viewed being close together will tend to be perceived as a unit.
- Law of Similarity – things that appear to have the same attributes are usually perceived as being a whole.
- Law of Inclusiveness (W.Kohler) - there is a tending to perceive only the larger figure and not the smaller when it is embedded in a larger.
- Law of Parsimony – the simplest example is the best or known as Ockham’s razor principle (14-th century): “entities should not be multiplied unnecessarily”.

We suggest to use these laws for pursuing conceptual balance and clarity of corporate knowledge ontology.

3.2.1 Conceptual Balance

A well-balanced ontological hierarchy equals a strong and comprehensible representation of the domain knowledge. However, it is a challenge to formulate the idea of a well-balanced tree. Here we offer some tips to help formulate the “harmony”:

- Concepts of one level should be linked with the parent concept by one type of relationship such as is-a, or has part.

- The depth of the branches should be more or less equal (± 2 nodes).
- The general outlay should be symmetrical.
- Cross-links should be avoided as much as possible.

3.2.2 Clarity

Moreover, when building a comprehensible ontology it is important to pay attention to clarity. Clarity may be provided through number of concepts and type of the relationships among the concepts. Minimizing the number of concepts is the best tip according to Law of Parsimony. The maximal number of branches and the number of levels should follow Miller's magical number (7 ± 2) (Miller, 1956). Furthermore, the type of relationship should be clear and obvious if the name of the link is missed.

At the first stages it is possible to use any of the available graphical editors to design an ontology, e.g. PaintBrush, Visio, Inspiration. A nice layout can be reached by using mindmapping tools as Freemind™, MindManager™ or Visual Mind™. The trainees really enjoyed the process of "beatification" of their ontologies during test exercises.

As an example we may discuss the ontology presented at Fig.2. This figure maps the ontology of knowledge engineering. We try to follow all the rules described earlier, but one can see that the branch "formalization" is too short and shallow. In our case it is understandable because of the specifics of the course which was aimed at non-programmers. But in the general ontology of this field should be detailed better.

4 Discussion

Challenges have fueled opportunities for analytic tool developers, educators, and business process owners that support analytic communities in the management of knowledge, information and data sources. The field of Knowledge Management has undergone several bouts of high hopes and press-induced hype ending with grave disappointment and missed promises. All too often we see old Information Management technology repackaged and retagged as the latest KM offering. However, today we still have functioning corporate KM systems being arranged by qualified knowledge workers. It is expected that large corporations will be forced to rethink their knowledge management strategies towards the human factors assessment. We hope that training and coaching of knowledge analysts will rise the new types of business development platforms and will play a key role in the articulation of the corporate KM landscape in the next 3-5-7 years.

Any mature company needs business analysts. Analysts are super-knowledge workers, but even they enter "the world of ontologies" with some doubt. But in the training their interest grows and rather soon they begin to use ontologies in their practical work. Our experience in training of knowledge analysts in the period of 1999-2010 confirm the unique role of knowledge structuring for developing ontologies quickly, efficiently and effectively. We follow David Jonassen's idea of using concept maps as "a mind tool" (Jonassen,1998). The use of visual paradigm for the representing and supporting the training process not only helps a professional trainer to concentrate on the problem rather than on details, but also enables students to process and understand greater volume of information. After training major of the

trainees were able to map their professional knowledge using different visual forms of ontology design – from mind maps to concept maps. They developed the ontologies of the customers, suppliers, products, solutions, requirements, projects, etc.

Business is based on knowledge processing in new information age. So the skillfull knowledge workers can really increase the productivity and sustainability of modern business practice in the innovate service-oriented economy. And the use and development of ontologies help to annotate information so that diverse groups of humans and machines can process it more meaningfully.

Acknowledgements

The work was partially funded by grants of Russian Foundation of Basic Research and grant from St.Petersburg State University. Thanks to all of my students and trainees whose questions and discussion help a lot to the improvement of the course.

References

- Blanchard, E., Mizoguchi, R., Lajoie, S.: Addressing the Interplay of Culture and Affect in HCI: An Ontological Approach //Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction. In: Proceedings of 13th International Conference, HCI International 2009, Part III, USA, July 19-24. LNCS, vol. 5612, pp. 575–584. Springer, Heidelberg (2009)
- Boose, J.H.: Knowledge Acquisition Tools, Methods and Mediating Representations. In: Motoda, H., et al. (eds.) Knowledge Acquisition for Knowledge-Based Systems, pp. 123–168. IOS Press, Ohinsha Ltd., Tokyo (1990)
- Dicheva, D., Gavrilova, T., Sosnovsky, S., Brusilovsky, P.: Ontological Web Portal for Educational Ontologies. In: Proc. Of Applications of Semantic Web Technologies for E-Learning Workshop (SW-EL'05) in Conjunction with 12th Int. Conf. on Artificial Intelligence in Education (AI-ED 2005), Amsterdam, pp. 19–29 (2005)
- Conlon, T.: Visions of Change: Information Technology, Education and Postmodernism. British Journal of Educational Technology Vol 31(2), 109–116 (2000)
- Firestone, J., McElroy, M.: Doing Knowledge Management. The Learning Organization 12(2), 189–212 (2005)
- Firestone, J.: Enterprise Information Portals and Knowledge Management. KMCI Press/Butterworth-Heinemann, Burlington, MA (2003)
- Fridman Noy, N., Griffith, N., Musen, M.: Collecting Community-Based Mappings in an Ontology Repository. In: International Semantic Web Conference, pp. 371–386 (2008)
- Jasper, R., Uschold, M.: A Framework for Understanding and Classifying Ontology Applications. In: 12th Workshop on Knowledge Acquisition Modelling and Management KAW 1999(1999); Bechhofer, S., Ng, G.: OilEd (2004), <http://oiled.man.ac.uk/>
- Jonassen, D.H.: Designing constructivist learning environments. In: Reigeluth, C.M. (ed.) Instructional Design Models and Strategies, 2nd edn., Lawrence Erlbaum, Mahwah (1998)
- Gavrilova, T., Voinov, A.: Work in Progress: Visual Specification of Knowledge Bases. In: del Pobil, A.P., Mira, J., Ali, M. (eds.) IEA/AIE 1998. LNCS, vol. 1416, pp. 717–726. Springer, Heidelberg (1998)
- Gavrilova, T.A., Voinov, A., Vasilyeva, E.: Visual Knowledge Engineering as a Cognitive Tool. In: Mira, J. (ed.) IWANN 1999. LNCS, vol. 1607, pp. 123–128. Springer, Heidelberg (1999)

- Gavrilova, T.: Teaching via Using Ontological Engineering. In: Proceedings of XI Int. Conf. Powerful ICT for Teaching and Learning PEG 2003, St.Petersburg, pp. 23–26 (2003)
- Gavrilova, T., Kurochkin, M., Veremiev, V.: Teaching Strategies and Ontologies for E-learning. *Int. J. Information Theories and Applications* 11(1), 35–42 (2004)
- Gavrilova, T., Laird, D.: Practical Design of Business Enterprise Ontologies. In: Bramer, M., Terzayan, V. (eds.) *Industrial Applications of Semantic Web*, pp. 61–81. Springer, Heidelberg (2005)
- Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering with examples from the areas of Knowledge Management. In: *e-Commerce and the Semantic Web*. Springer, Heidelberg (2004)
- Gruber, T.: A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 199–220 (1993)
- Guarino, N., Giaretta, P.: Ontologies and Knowledge Bases: Towards a Terminological Clarification. In: *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*, pp. 25–32. IOS Press, Amsterdam (1998)
- Guarino, N., Welty, C.: A Formal Ontology of Properties. In: Dieng, R., Corby, O. (eds.) *EKAW 2000. LNCS (LNAI)*, vol. 1937, pp. 97–112. Springer, Heidelberg (2000)
- Leydesdorff, L.: *The Knowledge-Based Economy Modeled, Measured, Simulated*. Universal-Publishers (2006)
- Miller, G.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63, 81–97 (1956)
- Mizogushi, R., Bourdeau, J.: Using Ontological Engineering to Overcome Common AI-ED Problems. *International Journal of Artificial Intelligence in Education* 11, 1–12 (2000)
- Neches, et al.: Enabling Technology for Knowledge Sharing. *AI Magazin*, Winter, 36– 56 (1991)
- Protégé, Stanford Medical Informatics (2007), <http://protege.stanford.edu/> (accessed)
- Scott, A., Clayton, J.E., Gibson, E.L.: *A Practical Guide to Knowledge Acquisition*. Addison-Wesley, Reading (1994); Swartout, B., Patil, R., Knight, K., Russ, T.: Toward Distributed Use of Large-Scale Ontologies. In: *Ontological Engineering, AAAI 1997 Spring Symposium Series*, pp. 138–148 (1997)
- Sowa, J.F.: *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley, Reading (1984)
- The CIO's Guide to Semantics, Semantic Arts©, Inc. (2004), <http://www.semantic-conference.com>
- Tu, S., Eriksson, H., Gennari, J., Shahar, Y., Musen, M.: Ontology-Based Configuration of Problem-Solving Methods and Generation of Knowledge-Acquisition Tools. In: *Artificial Intelligence in Medicine, N7*, pp. 257–289 (1995)
- Wertheimer, M.: *Productive Thinking*. HarperCollins, New York (1959)
- Wielinga, B., Schreiber, G., Breuker, J.: A Modelling Approach to Knowledge Engineering. *Knowledge Acquisition, Special Issue* 4(1), 23–39 (1992)
- Wiig, E.H., Wiig, K.M.: *On Conceptual Learning* Knowledge Research Institute, Inc. Working Paper 1999-1 (1999), http://www.krii.com/downloads/conceptual_learning.pdf
- Uschold, M., Gruninger, M.: *Ontologies: Principles Methods and Applications*. *Knowledge Engineering Review* 1(1) (1996)

Attribute Exploration Algorithms on Ontology Construction*

Qin Ping^{1,2}, Zhang Zhongxiang¹, Gao Hualing¹, and Wang Ju¹

¹ Guangxi Normal University, College of Computer Science and Information Technology, Guilin, 541004, China

² Guangxi Arts Institute, Postgraduate Affairs Office, Nanning, 530022, China
qpapple_5716@163.com

Abstract. Attribute exploration in FCA is proposed by Baader etc. in the past decade and it is an effective tool applying to description logics to construct ontology on Semantic Web. The authors firstly introduce attribute exploration algorithm, then investigate different cases in which the redundant computation may occur. As new results, the improved attribute exploration algorithm is proposed in terms of relevancy. We also give the proof of the completeness of the improved algorithm, and show how the proposed algorithm avoids redundancy and simplifies computation in some certain cases. We finally present the method to construct an ontology based on attribute exploration algorithm(AEOCM) on the open formal context, and specify the implementation procedure of this method in terms of instantiation.

Keywords: Attribute Exploration Algorithms, Description Logics, Formal Concept Analysis, Pseudo Intent, Ontology Construction.

1 Introduction

With the development of the semantic web, we have to confront the problem on searching useful information from the mass data. The way to acquire the information which we focus on is making good use of ontology to improve the veracity and validity on searching. An ontology, the backbone of the Semantic Web, is an explicit and formal specification of a conceptualization and is consisted of finite list of terms and the relationships between these terms. A well-defined ontology should be constructed by promising approaches and assures the soundness and completeness. Therefore, how to construct a well-defined ontology becomes a hotspot on the field of intelligent information processing.

Attribute exploration, a simple but useful knowledge acquisition technique, is an important tool from Formal Concept Analysis (FCA), a mathematical theory for concepts and conceptual hierarchies. It is used to acquire knowledge from a domain expert by asking successive questions, find implication between attributes which can

* Supported by The National Natural Science Foundation of China under Grant No.60573010, 60663001;The Innovation Project of Guangxi Graduate Education under Grant No.2008M025.

express the knowledge on inclusion between object sets, and get all of the intent and stem base. After the method, applying formal concept analysis to description logics presented by Franz Baader in 2004-2007, there are some preferable results to modify the attribute exploration to determine a minimal set of implications[1], extend both the terminological and the assertional part of a description logic knowledge base[2], extend the definition on conceptual description on the least common subsumer (lcs) to background terminology[3], and show the attribute exploration algorithm to compute the subsumption hierarchy of all lcs as well as the hierarchy of all conjunctions of subsets of description logic concepts[4].

Nowadays, many ontologies are constructed by manual or semi-automatic work. Attribute exploration is an efficient and automatic way to generate concept lattice and construct ontology, and makes its goal. With the help of it, we can improve the ontology construction methods, specially on the open semantics by asking the domain expert to get the hierarchical relationship between attributes and objects, and find out all of the knowledge in the domain. It is possible to share and reuse knowledge after the ontology has built and some new knowledge is added easily.

In this paper, we firstly introduce the attribute exploration algorithm, proposed applying formal concept analysis to description logic by Baader, etc. during the past decade. After studying on it, we find out some redundant computation. We propose an improved attribute exploration algorithm in terms of relevancy simplify the computing, and prove its completeness, finally propose an ontology construction method based on attribute exploration algorithm(AEOCM), and specify the implementation procedure of this method in terms of instantiation.

2 Basic Definitions in Formal Concept Analysis

Formal concept analysis (FCA) is a field of applied mathematics that aims to formalize the notions of a concept and a conceptual hierarchy by means of mathematical tools. It facilitates the use of mathematical reasoning for conceptual data analysis and knowledge processing. There are some basic definitions and theorems used in the following.

Definition 2.1[4,6](Formal context). A formal context is a triple $\mathcal{K}=(\mathcal{O},\mathcal{P},\mathcal{S})$, where \mathcal{O} is a set of objects, \mathcal{P} is a set of attributes, and $\mathcal{S}\subseteq\mathcal{O}\times\mathcal{P}$ is a relation that associates each object o with the attributes satisfied by o .

Definition 2.2[4,6] (Derivation operator). Let $\mathcal{K}=(\mathcal{O},\mathcal{P},\mathcal{S})$ be a formal context. For a set of objects $A\subseteq\mathcal{O}$, we define the intent A' of A as follows: $A':=\{p\in\mathcal{P}\mid\forall a\in A.(a,p)\in\mathcal{S}\}$.

Similarly, for a set of attributes $B\subseteq\mathcal{P}$, we define the extent $B':=\{o\in\mathcal{O}\mid\forall b\in B.(o,b)\in\mathcal{S}\}$.

Definition 2.3[6] (Formal concept). Let $\mathcal{K}=(\mathcal{O},\mathcal{P},\mathcal{S})$ be a formal context. A formal concept of \mathcal{K} is a pair (A,B) where $A\subseteq\mathcal{O}$, $B\subseteq\mathcal{P}$ such that $A'=B$ and $B'=A$. We call A the extent, and B the intent of (A,B) .

Definition 2.4[7] (Subconcept-superconcept-relation). The subconcept-superconcept-relation is mathematized by $(A_1, B_1) \leq (A_2, B_2) : \Leftrightarrow A_1 \subseteq A_2 (B_1 \supseteq B_2)$.

The set of all formal concepts of \mathcal{K} together with the defined order relation is denoted by $\mathfrak{B}(\mathcal{K})$.

Definition 2.5[7] (Concept lattices). Let $\mathcal{K}=(\mathcal{O}, \mathcal{P}, \mathcal{S})$ be a formal context. Then $\mathfrak{B}(\mathcal{K})$ is a complete lattice, called the concept lattice of $(\mathcal{O}, \mathcal{P}, \mathcal{S})$, for which infimum and supremum can be described as follows:

$$\bigwedge_{i \in T} (A_i, B_i) = (\bigcap_{i \in T} A_i, \bigcup_{i \in T} B_i)'' , \quad \bigvee_{i \in T} (A_i, B_i) = ((\bigcup_{i \in T} A_i)'', \bigcap_{i \in T} B_i) .$$

Definition 2.6[6] (Implication between attributes). An implication $L \rightarrow R$ holds in \mathcal{K} if every object of \mathcal{K} that has all of the attributes in L also has all of the attributes in R , i.e., if $L' \subseteq R'$. We denote the set of all implication that hold in \mathcal{K} by $Imp(\mathcal{K})$.

Definition 2.7[8] (Pseudo-intent). $P \subseteq M$ is called the pseudo-intent of $(\mathcal{O}, \mathcal{P}, \mathcal{S})$ if and only if $P \neq P''$ and $Q'' \subseteq P$ holds for every pseudo-intent $Q \subseteq P, Q \neq P$.

Definition 2.8[8] (Duquenne-Guigues-Base). The set of implications $\{P \rightarrow P' \setminus P \mid P \text{ pseudo-intent}\}$. We call this the Duquenne-Guigues-Basis or simply the stem base of the attribute implications.

Theorem 1[8]. The set of implications $\mathcal{L} := \{P \rightarrow P' \setminus P \mid P \text{ pseudo-intent}\}$ is non-redundant and complete.

Definition 2.9[4,8] (Lectic order). Let an arbitrary linear order on the set of attributes $\mathcal{P} = \{p_1, \dots, p_n\}$, say $p_1 < \dots < p_n$. For all $j, 1 \leq j \leq n$, and $B_1, B_2 \subseteq \mathcal{P}$ we define $B_1 < B_2$ iff $p_j \in B_2 \setminus B_1$ and $B_1 \cap \{p_1, \dots, p_{j-1}\} = B_2 \cap \{p_1, \dots, p_{j-1}\}$.

Theorem 2. For a given attribute set $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$, which basic order is $p_1 < p_2 < \dots < p_n$ and $B \subset \mathcal{P}$, we denote the first element after B on lectic order by B^+ . It satisfies:

- (i) existing a positive j is maximal such that $p_j \notin B$;
- (ii) $B^+ = B \cap \{p_1, p_2, \dots, p_{j-1}\} \cup \{p_j\}$.

Definition 2.10[4] (Implication Pseudo-hull). For a subset B of \mathcal{P} , the implication pseudo-hull of B with respect to \mathcal{J} is denoted by $\mathcal{J}^*(B)$. It is the smallest subset H of \mathcal{P} such that

- (i) $B \subseteq H$;
- (ii) $B_1 \rightarrow B_2 \in \mathcal{J}$ and $B_1 \subset H$ (strict subset) imply $B_2 \subseteq H$.

Definition 2.11[8] (respect). A subset $T \subseteq M$ respects an implication $A \rightarrow B$ if $A \not\subseteq T$ or $B \subseteq T$. T respect a set \mathcal{L} of implication if T respect every single implication in \mathcal{L} .

Definition 2.12. In the concept lattice, the length of maximal chain from concept (A,B) to maximum is \mathcal{N} , (A,B) is the \mathcal{N} level node of lattice.

3 Ontology and Its Construction by the Way of FCA

Ontologies, often defined as an explicit specification of conceptualization, are necessary for knowledge representation and knowledge exchange. Generally ontology can be formally defined by (C,P,I,S,E) , where C refers to Class; P refers to property of Class; I refers to instance of Class; S refers to subsumption relation and E refers to other Enriched relation. In the concept lattice, the concept can be mapped to class in the ontology definition; the elements of objects in each concept can be mapped to the instance of ontology; the element of attributes will be mapped to the property of ontology and finally subsuper concept relation is equivalent to subsumption relation in ontology. [9]

A good design ontology means that they should adequately capture the modeled domain, be understandable for human user and provide good support for machine processing. By a good definition means not only the syntax, but also the semantics. An automated reasoning over ontologies enables to support the ontology design, integrating and sharing ontologies automatically, determining and establishing relationships among ontologies etc.[10]

Formal Concept Analysis (FCA) is adopted in the ontology construction process to establish the taxonomic hierarchy correctly and properly, because it yields the mathematization of concepts and component ontology are explained briefly. The way to applying the formal concept analysis to ontology construction, first of all, we should model context in a research domain by topical similarity and subsumption which should expand user's search context and improve the interactive capability of traditional search engines, and develop information context. In the concept lattice, a concept will be described by a set of attributes. In this case, we can get a formal context which is a crucial step. Then we use attribute exploration based on formal context and with the help of the domain expert to construct ontology on the concept lattice and get Duquenne-Guigues Base on knowledge inference.

4 Attribute Exploration Algorithm

Attribute exploration is a method of acquiring knowledge from a domain expert by asking successive questions to expand ABox and TBox. In many application fields where the formal context is not explicitly given, but only "known" to some domain expert, it has proved to be a successful method for efficiently capturing the expert's knowledge. The superiority of attribute exploration is not only computing intents, pseudo-intents and Duquenne-Guigues Base, but also expanding counterexamples to it by domain expert to acquire complete knowledge about the application domain instead of being restricted by the given formal context.

4.1 Attribute Exploration Algorithm Proposed by Baader etc.

For a given formal context, implications between attributes can express the important knowledge of subsumption between object sets. In some cases, especially impossible to list all of the subsumption, we try to compute the base of these subsumption. Using it, we can compute all of intents and stem base, which can inference all of the subsumption. The attribution exploration algorithm proposed by Baader etc. is as follows:

Algorithm 1(Attribution exploration)[4]

Initialization: One starts with the empty set of implication, i.e., $\mathcal{J}_0:=\emptyset$, the empty set of concept intents $\mathcal{C}_0:=\emptyset$, and the empty subcontext \mathcal{K}_0 of \mathcal{K} , i.e., $\mathcal{O}_0:=\emptyset$. The lectic smallest subset of \mathcal{P} is $B_0:=\emptyset$.

Iteration: Assume that $\mathcal{K}_i, \mathcal{J}_i, \mathcal{C}_i$, and $B_i(i \geq 0)$ are already computed. Compute B_i'' with respect to the current subcontext \mathcal{K}_i . Now the expert is asked whether the implication $B_i \rightarrow B_i'' \setminus B_i$ holds in \mathcal{K} .

If the answer is “no”, then let $o_i \in \mathcal{O}$ be the counterexample provided by the expert. Let $B_{i+1} := B_i, \mathcal{J}_{i+1} := \mathcal{J}_i$, and let \mathcal{K}_{i+1} be the subcontext of \mathcal{K} with $\mathcal{O}_{i+1} := \mathcal{O}_i \cup \{o_i\}$. The iteration continues with $\mathcal{K}_{i+1}, \mathcal{J}_{i+1}, \mathcal{C}_{i+1}$, and B_{i+1} .

If the answer is “yes”, then $\mathcal{K}_{i+1} := \mathcal{K}_i$ and

$$(\mathcal{C}_{i+1}, \mathcal{J}_{i+1}) := \begin{cases} (\mathcal{C}_i, \mathcal{J}_i \cup \{ B_i \rightarrow B_i'' \setminus B_i \}) & \text{if } B_i'' \neq B_i \\ (\mathcal{C}_i \cup \{B_i\}, \mathcal{J}_i) & \text{if } B_i'' = B_i \end{cases}$$

To find the new set B_{i+1} , we starts with $j=n$, and test whether

$$B_i <_j \mathcal{J}_{i+1}^* ((B_i \cap \{ p_1, \dots, p_{j-1} \}) \cup \{ p_j \}) \tag{*}$$

holds. The index j is decreased until one of the following cases occurs:

(1) $j=0$: In this case, \mathcal{C}_{i+1} is the set of all concept intents and \mathcal{J}_{i+1} the Duquenne-Guigues base of \mathcal{K} , and the algorithm stops.

(2) (*) holds for $j > 0$: In this case, $B_{i+1} := \mathcal{J}_{i+1}^* ((B_i \cap \{ p_1, \dots, p_{j-1} \}) \cup \{ p_j \})$, and the iteration is continued.

By computing on this algorithm, we find out that there are some redundancies in the iterative steps. This redundancy happens in computing formula(*). When computing B_{i+1} , we work out the implication pseudo-hull, i.e. $\mathcal{J}_{i+1}^* ((B_i \cap \{ p_1, \dots, p_{j-1} \}) \cup \{ p_j \})$ and lectic order $<_j$. It appear this case that the same set of implication pseudo-hull is computed repeatedly when j don't satisfy formula(*) and then $j := j-1$. For avoiding the complex computation repeatedly, we try to improve the efficiency on the following.

4.2 The Improved Attribute Exploration Algorithm(I)

It is evident that computation the implication pseudo-hull and lectic order is rather complex. In order to resolve this problem, we can modify the way to compute B_{i+1} in terms of relevancy between attribute sets and implication sets, i.e. asking that does B_i^+ (the next element of B_i on lectic order) respect an implication \mathcal{J}_{i+1} , if the answer is “Yes”, let $B_{i+1} := B_i^+$, the iteration is continued; if the answer is “No”, let $B_i := B_i^+$, continue to compute the next B_i^+ . Modify the improved algorithm as follows:

Table 1. Improved Attribute Exploration Algorithm

Algorithm 2. Improved attribute exploration	
1:	Input : A formal context $\mathcal{K} (\mathcal{O}, \mathcal{P}, \mathcal{S})$
2:	Initialization
3:	$\mathcal{I}_0 := \emptyset$ {initial empty set of implications}
4:	$\mathcal{C}_0 := \emptyset$ {initial empty set of intents}
5:	$\mathcal{K}_0 := \emptyset$ {initial formal context, possibly empty set of objects}
6:	$\mathcal{O}_0 := \emptyset$ {initial empty set of objects}
7:	$B_0 := \emptyset$ {initial empty set of lectic order}
8:	while $B_i \neq \mathcal{P}$ do {assume $\mathcal{K}_i, \mathcal{I}_i, \mathcal{C}_i, B_i (i \geq 0)$ are already computed}
9:	Compute B_i'' w.r.t. \mathcal{K}_i
10:	Ask the expert if $B_i \rightarrow B_i'' \setminus B_i$ holds w.r.t. \mathcal{K} ?
11:	if no then the expert provides a counterexample $o_i \in \mathcal{O}$
12:	$\mathcal{O}_{i+1} := \mathcal{O}_i \cup \{o_i\}$
13:	$B_{i+1} := B_i$
14:	$\mathcal{I}_{i+1} := \mathcal{I}_i$
15:	$\mathcal{C}_{i+1} := \mathcal{C}_i$
16:	$\mathcal{K}_{i+1} \{ \mathcal{K}_i (\mathcal{O}_i, \mathcal{P}_i, \mathcal{S}_i), \mathcal{O}_i \subseteq \mathcal{O}, \mathcal{P}_i \subseteq \mathcal{P}, \mathcal{S}_i \subseteq \mathcal{S} \cap (\mathcal{O}_i \times \mathcal{P}_i); \mathcal{K}_{i+1}$ is a subcontext of \mathcal{K}
17:	else {yes}
18:	$\mathcal{K}_{i+1} := \mathcal{K}_i$
19:	if $B_i'' \neq B_i$ then { B_i is a pseudo-intent}
20:	$(\mathcal{C}_{i+1}, \mathcal{I}_{i+1}) := (\mathcal{C}_i, \mathcal{I}_i \cup \{B_i \rightarrow B_i'' \setminus B_i\})$
21:	else { $B_i'' = B_i$, B_i is an intent}
22:	$(\mathcal{C}_{i+1}, \mathcal{I}_{i+1}) := (\mathcal{C}_i \cup \{B_i\}, \mathcal{I}_i)$
23:	end if
24:	Compute B_{i+1}
25:	while $B_i \neq \mathcal{P}$ then
26:	Does B_i^+ respect an implication \mathcal{I}_{i+1} ?
27:	if yes then
28:	$B_{i+1} := B_i^+$
29:	break
30:	else {no}
31:	$B_i := B_i^+$
32:	end if
33:	end while
34:	end if
35:	end while
36:	Output: $\mathcal{C}_i, \mathcal{I}_i$

Theorem 3. The improved attribute exploration algorithm is complete.

Proof: According to the algorithm, we can obtain the set $\{B_0 = \emptyset, B_1, B_2, \dots, B_m = \mathcal{P}\}$, and $B_0 < B_1 < B_2 < \dots < B_m$.

(1) Initialization: Take $B_0 = \emptyset$ as beginning, classified discussion as follows:

If $\emptyset'' = A_0 (A_0 \neq \emptyset)$, then \emptyset is a pseudo-intent, $\mathcal{I}_1 = \{\emptyset \rightarrow A_0\}$. The first element respecting with \mathcal{I}_1 after \emptyset is A_0 , so $B_1 = A_0$ is intent. On the other hand, as \emptyset is a pseudo-intent, any non-empty pseudo-intent must include A_0 , similarly, A_0 must be included in any non-empty intent. It's known that there is any element neither intent nor pseudo-intent.

If $\emptyset'' = \emptyset$, \emptyset is a intent ($B_0 = \emptyset$), $\mathcal{J}_1 = \emptyset$, the first element after \emptyset is $\{p_n\}$, $\{p_n\}$ respects \mathcal{J}_1 , $B_1 = \{p_n\}$. As $\{p_n\}$ is the next element of \emptyset on the lectic order. there is not any element between B_0 and B_1 . Now we only have to prove B_1 is intent or pseudo-intent: if $\{p_n\}'' = \{p_n\}$, $\{p_n\}$ is intent; if $\{p_n\}'' \neq \{p_n\}$, $\{p_n\}$ is pseudo-intent.

Above all, B_0 and B_1 is intent or pseudo-intent. Moreover there is no other intent or pseudo-intent between B_0 and B_1 .

(2) Inductive argument: On the assumption that all of the intents and pseudo-intents less than or up to B_i on lectic order have been computed, denoting that B_1, B_2, \dots, B_i and ensure there is no intent or pseudo-intent between adjacent elements. Now $\mathcal{J}_{i+1} = \{P \rightarrow P'' \setminus P \mid P \text{ is pseudo-intent, } P \leq B_i\}$. According to the algorithm, testing from the first element after B_i on lectic order one by one, until acquiring the first element B respecting \mathcal{J}_{i+1} , let $B_{i+1} = B$, now we should prove that : (i) B_{i+1} is intent or pseudo-intent; (ii) There is no intent or pseudo-intent between B_i and B_{i+1} .

Case 1: If $B = B_i^+$, i.e. B_i^+ , the first element after B_i on lectic order, respects \mathcal{J}_{i+1} .

Considering pseudo-intent Q , satisfying $Q \subseteq B_i^+$, we know that \mathcal{J}_{i+1} includes the implication $Q \rightarrow Q'' \setminus Q$ according to the inductive assumption above, i.e. $Q = B_k, k \leq i, B_i^+$ respects \mathcal{J}_{i+1} . It means that $Q \subseteq B_i^+$ imply $Q'' \setminus Q \subseteq B_i^+$, i.e. $Q'' \subseteq B_i^+$. Now if $(B_i^+)'' = B_i^+, B_i^+$ is intent; if $(B_i^+)'' \neq B_i^+, B_i^+$ is pseudo-intent. So (i) and (ii) hold.

Case 2: If $B \neq B_i^+$, i.e. B_i^+ doesn't respects \mathcal{J}_{i+1} .

We should prove B_i^+ is neither intent nor pseudo-intent. Using reduction to absurdity, considering pseudo-intent Q , satisfying $Q \subseteq B_i^+$, we know that \mathcal{J}_{i+1} includes the implication $Q \rightarrow Q'' \setminus Q$ according to the inductive assumption above, i.e. $Q = B_k, k \leq i, B_i^+$ respects \mathcal{J}_{i+1} . It is evident that it contradicts with the assumption that B_i^+ doesn't respect \mathcal{J}_{i+1} . So B_{i+1} is neither intent nor pseudo-intent.

Now testing the next element of B_i^+, B_i^+ either doesn't respect \mathcal{J}_{i+1} , or is neither intent nor pseudo-intent. There is a minimal element B on lectic order, respecting \mathcal{J}_{i+1} . Now let $B_{i+1} = B$, there is no intent or pseudo-intent between B_i and B_{i+1} . We should prove B_{i+1} is intent or pseudo-intent, considering pseudo-intent A , satisfying $A \subseteq B_{i+1}$, i.e. $A \leq B_{i+1}$. We know that and \mathcal{J}_{i+1} includes the implication $A \rightarrow A'' \setminus A$ according to the inductive assumption above, i.e. $A = B_l, l \leq i, B_{i+1}$ respects \mathcal{J}_{i+1} . It means that $A \subseteq B_{i+1}$ imply $A'' \setminus A \subseteq B_{i+1}$, i.e. $A'' \subseteq B_{i+1}$. Now if $(B_{i+1})'' = B_{i+1}$, B_{i+1} is intent; if $(B_{i+1})'' \neq B_{i+1}$, B_{i+1} is pseudo-intent. So (i) and (ii) hold.

Therefore, the improved attribute exploration algorithm is complete. □

4.3 Analysis and Comparison of the Two Attribute Exploration Algorithms

Comparing with the two different algorithms, we can make this conclusion that the improved attribute exploration algorithm neither needn't compute implication pseudo-hull, nor judge whether the computed pseudo-hull satisfy formula(*) based on the definition of lectic order, but judges the next attribute set on lectic order from the view of respect to acquire the next intent or pseudo-intent. The computation with

respect, implementing on the way of subsumption between sets, effectively simplify the complexity. It also avoids redundancy computation the same set repeatedly, which sometimes happened in computing the next B_{i+1} in the attribute exploration algorithm proposed by Baader etc. We also prove the improved one is complete.

5 Ontology Construction Based on Attribute Exploration Algorithm

According to the results of research, we propose a model of the method of ontology construction based on attribute exploration algorithm (AEOCM). This model includes four modules: preprocessing module, attribute exploration module, generating concept lattice module, ontology construction module. The figure of this model is as following Fig. 1.

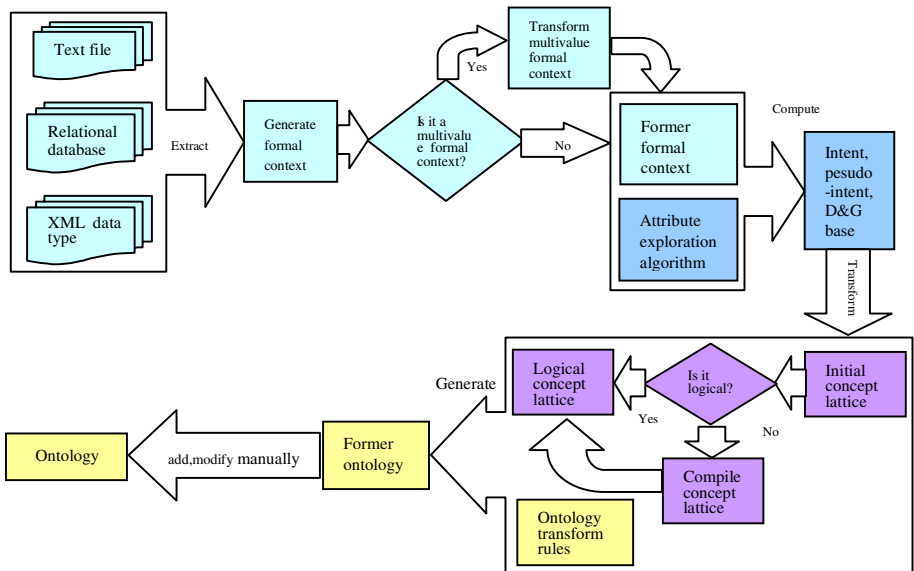


Fig. 1. The model of ontology construction based on attribute exploration algorithm (AEOCM)

Figure legends: :preprocessing module : attribute exploration module
 :generating concept lattice module :ontology construction module

5.1 Preprocessing Module

In practice, formal context is generally not given, but extracted from data sources. As a result, we need to analysis data resources and resort to respective strategies and algorithms for different data resources to extract formal context. The main processes

of preprocessing module are extracting the multi-value formal context which contains attributes and objects from data source files including text files, relational databases, XML data types and so on, and transforming multi-value formal context to single one. The preprocessing module includes two submodules: the submodule of extracting formal context, the submodule of transforming formal context .

5.2 Attribute Exploration Module

The main processes of attribute exploration module are computing intent, pseudo-intent, Duquenne- Guigues base by the ways of the improved attribute exploration algorithm after asking the expert successive questions, in order to generate formal concept lattice. The way of this module to distinguish with other algorithms computing intent and pseudo-intent, is asking the expert whether the implication holds in the given context, if the answer is “yes”, then doing the next computing in the formal context \mathcal{K}_i ; if the answer is “no”, then providing counterexample to \mathcal{K}_i . In this case, it ensures that the new objects are dynamically added to the formal context and what is never expressed explicitly but on the experts’ mind, making the experts expand the knowledge base easily.

5.3 Generating Concept Lattice Module

This module is mainly to generate concept lattice by the structure concept lattice algorithm according to the computed intents and draw the respective hasse graph. The up to down structure concept lattice algorithm is as follows:

Input: All the formal concept $\mathfrak{B}(\mathcal{K})$.

Initialization: $\forall (X,Y) \in \mathfrak{B}(\mathcal{K})$ access to the set $C_{|Y|}$ according to the size of $|Y|$, then we will get the sets, such as $C_0 \parallel C_1 \parallel \dots \parallel C_{|P|}$, $C_0 = (P, P')$ is maximum.

Iteration: Assume $C_k = (A_k, B_k)$ is already computed.
 Start from the k ($k \geq 0$) level, compute as following method:
 $\forall (X,Y) \in C_k$, if finding (X_1, Y_1) in C_{k+1} , satisfy $Y \subset Y_1$, connect (X,Y) with (X_1, Y_1) ; if finding (X_2, Y_2) in C_{k-i} ($1 \leq i < k$), satisfy $Y_2 \subset Y$ and $X_2 \prec X$, connect (X,Y) with (X_2, Y_2) . The iteration is continued.
 Until $k = |P|$, the algorithm stops.

Output: Concept lattice $\langle \mathfrak{B}(\mathcal{K}), \leq \rangle$.

5.4 Ontology Construction Module

This module mainly contains constructing an ontology from above concept lattice according to the ontology generating rules, and adding other properties and restrictions practically, finally generating the ultimate ontology.

The respective ontology generating rules are as follows:

From the table1, we can make a conclusion that the classes, individuals and their hiberarchy can be found in the lattice directly, however, object properties and data properties are added manually by the constructor.

Table 2. The respective ontology generating rules

Lattice	Description Logic	Ontology
Attributes	Concepts	Classes
Objects	Instances	Individuals
Hierarchy of lattice	Subsumptions and equivalences between concepts	Hierarchy between classes and individuals
Can't be gotten directly	Hierarchy among Roles	The relations among Object Properties
Can't be gotten directly	Number restriction operations	Data Properties define datatype

6 Conclusion

In recent years, the automatic methods on ontology construction become a research hotspot. Attribute exploration in FCA is an efficient tool to generate concept lattice and construct ontology, and assure to share and reuse knowledge. Through asking a domain expert question, we can add new knowledge to generate the concept lattice structure on concept hierarchy to construct ontology. In this paper, we firstly have investigated the existing attribute exploration algorithm and find its redundancy computation, then have proposed an improved attribute exploration algorithm to simplify the complexity effectively. Applying this algorithm to ontology construction, we also have proposed a model of ontology construction based on attribute exploration algorithm (AEOCM), and elaborate the whole implement procedures in detail according to each module. Our experience in the research has convinced us that the study in this direction is indeed significant and should have a good prospect both in the application and the theory. We will continue our works to find more efficient algorithms and try to implement it in practice to construct ontology and its automatic inference.

References

1. Stumme, G.: Attribute exploration with background implications and exceptions. In: Bock, H.-H., Polasek, W. (eds.) *Data Analysis and Information Systems. Proceedings of the 19th Annual Conference of Gesellschaft für Klassifikation e.v. University of Basel, March 8-10 (1995)*, pp. 457–469. Springer, Heidelberg (1996)
2. Baader, F., Ganter, B., Sattler, U., et al.: *Completing Description Logic Knowledge Bases using Formal Concept Analysis*, LTCS-Report 06-02[R]. Germany, Dresden University of Technology (2006)
3. Baader, F., Sertkaya, B., Turhan, A.-Y.: Computing the least common subsumer w.r.t. a background terminology. *Journal of Applied Logic* 5, 392–420 (2007)
4. Baader, F., Sertkaya, B.: Applying formal concept analysis to description logics. In: Eklund, P. (ed.) *ICFCA 2004. LNCS (LNAI)*, vol. 2961, pp. 261–286. Springer, Heidelberg (2004)
5. Baader, F.: Computing a minimal representation of the subsumption lattice of all conjunctions of concepts defined in a terminology[J]. In: *Proc. Intl. KRUSE Symposium*, pp. 168–178 (1995)
6. Sertkaya, B.: *Formal Concept Analysis Methods for Description Logics(D)*

7. Wille, R.: Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies. In: Ganter, B., Stumme, G., Wille, R. (eds.) Formal Concept Analysis. LNCS (LNAI), vol. 3626, pp. 1–33. Springer, Heidelberg (2005)
8. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations, pp. 79–90. Springer, Heidelberg (1999)
9. Jia, H., Newman, J., Tianfield, H.: A new Formal Concept Analysis based learning approach to Ontology building,
<http://www.mtsr.ionio.gr/proceedings/haibo.pdf>
10. Obitko, M., Snasel, V., Smid, J.: Ontology Design with Formal Concept Analysis. In: CLA 2004, pp. 111–119 (2004)
11. Wenxiu, Z., Ling, W., Jianjun, Q.: The attribute reduction theory and approach of concept lattice. Science in China ser.E Information Sciences 35(6), 628–639 (2005)

Intelligent Business Transaction Agents for Cross-Organizational Workflow Definition and Execution

Mohammad Saleem¹, Paul W.H. Chung¹, Shaheen Fatima¹, and Wei Dai²

¹ Computer Science Department, Loughborough University, Loughborough, LE11 3TU, UK
{M.Saleem, P.W.H.Chung, S.S.Fatima}@lboro.ac.uk

² School of Information Systems, Victoria University, Melbourne City MC, Victoria, Australia
Wei.Dai@vu.edu.au

Abstract. Business organizations seldom work in isolation; effective interaction and cooperation between different organizations is essential for the success of any organization. In order to work together, organizations must have compatible workflows. This paper proposes a framework for automatically creating compatible workflows from high level requirements given by organizations that intend to work together. The framework uses intelligent agents [1] for the creation and execution of workflows. The proposed framework is different from existing systems as they focus on collaboration by modifying existing workflows. The proposed framework will enable organizations to avoid the time consuming task of creating a workflow and then ensuring that it is compatible with every organization it interacts with.

Keywords. Cross-Organizational Workflow Collaboration, Web Service Composition, Workflow Generation.

1 Introduction

A business process represents successful work practice and is a crucial part of corporate asset. It refers to a set of connected and ordered activities to achieve a business goal within the context of an organizational structure [2]. In this internet age more and more organizations are embracing electronic commerce and there is increasing demand for automatic business process management. Also, when two organizations interact with each other, the need for automatic cross-organizational business process collaboration arises. Since workflow technology is commonly used for business process management within a single organization, there is a need for support for cross-organizational workflow collaboration.

For business collaboration to work, workflows of business partners should be compatible at the business level [3]. Compatible means that there should be an agreed sequence of exchanging collaborative messages and information. The points where collaborative messages and information is exchanged is called interface activity [4]. The set of all interface activities when extracted from a process is called an interface process. Considerable amount of effort is needed to ensure that workflows are compatible in the first place [5, 6].

Fig. 1 shows a workflow for a vendor and the corresponding interface process, which only models the points where interaction takes place with another organization, the customer in this case. In this paper, an activity name followed by [s] or [r] means sending and receiving collaboration message respectively [4]. For example ‘Advance Payment[r]’ means an activity ‘Advance Payment’ is required to receive a message from the collaborating organization.

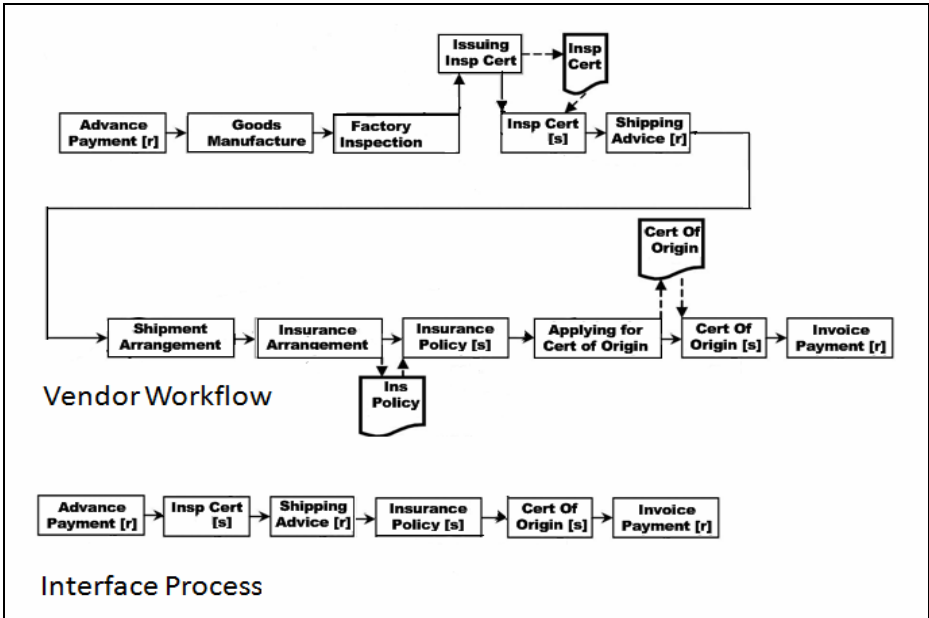


Fig. 1. Workflow and Interface Process for a Vendor

Any incompatibility between two workflows has to be reconciled before proceeding with the business. This can be a very time consuming process. To overcome this problem a new framework for automated workflow generation is presented.

Workflow generation is considered as an AI planning problem. AI planning creates an execution plan that reaches the required goal state from a given initial state. Since web services can be organized in a workflow to support the execution plan so AI planning can be applied to automatic web service composition and hence workflow generation [7]. Given a high level description of a goal, a planner can reason with all available services in terms of their capabilities. A required service that can achieve a desirable state will be identified and added into a workflow. Executing the complete workflow will result in the goal state [8].

Section 2 explains the proposed framework. Section 3 describes the implementation. Section 4 summarizes some related work. Conclusions are drawn in Section 5.

2 Proposed Framework

Fig. 2 shows the architecture of the proposed framework. Although the number of organizations is not limited to two, but for clarity the figure only depicts two

organizations. Each organization has its own high level goals and OWLS process definition. Also for each organization a separate intelligent agent is created which has its own instance of SHOP2 planner.

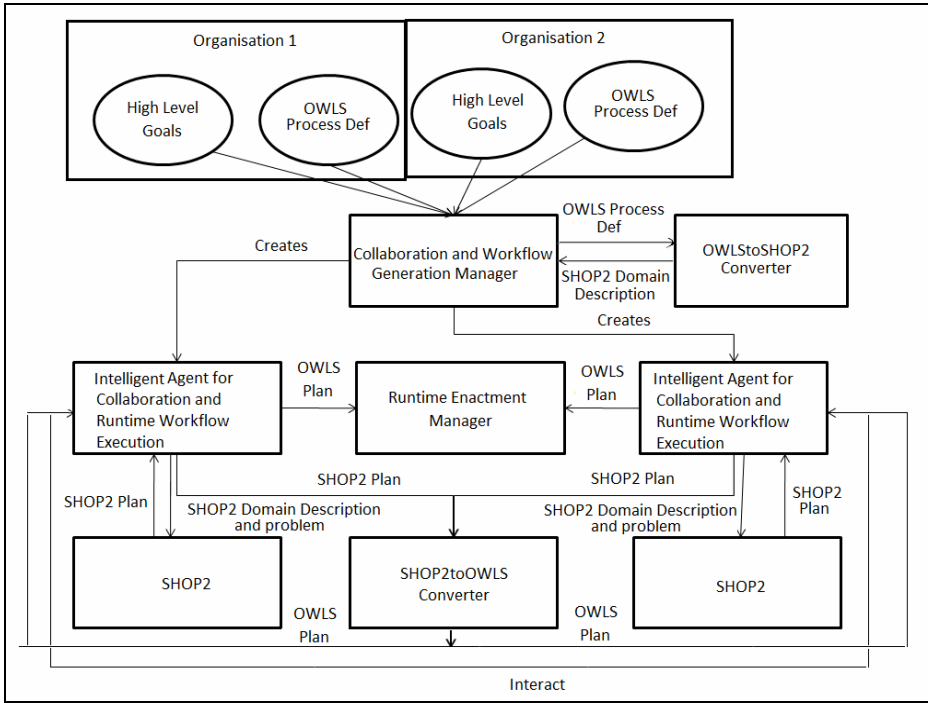


Fig. 2. Architectural Diagram

The following text discusses some of the basic components shown in Fig 2. Simple Hierarchical Ordered Planner 2 (SHOP2) [9, 10] is a domain independent hierarchical task network (HTN) planner which is used in the proposed system. It decomposes task into subtasks in the order they are to be executed. It requires domain description for planning. Domain description consists of operators and methods. Operators are primitive tasks while methods are specifications for decomposing composite tasks into subtasks [11].

As tasks are planned in the order in which they will be performed, this makes it possible for SHOP2 to know the current state of the world at every step. This in turn enables the precondition evaluation mechanism of SHOP2 to do reasoning and call external functions and sources [11]. A mechanism for querying external sources and backtracking in SHOP2 in case of failures is described in [12]. SHOP2 also allows tasks and subtasks to be ordered, unordered and partially ordered which makes it possible to interleave certain tasks with other tasks. These functionalities make SHOP2 an appropriate choice for cross-organizational workflow collaboration.

The OWL-Service (OWL-S) language provides a mechanism for describing web services [13]. It is used to describe the functionality, access point and execution

mechanism of web services. OWL-S is a set of ontologies and OWL-S process ontology describes web services composition based on ‘action’ or ‘process’ metaphor. It describes simple tasks as simple actions or simple processes and complex tasks as composite actions or composite processes. This similar way of modeling makes it possible to translate OWL-S web services descriptions to SHOP2 domain descriptions.

Each organization has high level rules for carrying out business which are stated in its process definition. Process definition is actually the OWLS domain ontology of the respective organization. High level goals describe the final goals of the organization which have to be achieved. OWLS process definition and high level goals together express the business interest, requirements and rules of an organization to carry out business with another organization.

As shown in Fig. 2, the interacting organizations pass their OWL-S process definitions and high level goals to Collaboration and Workflow Generation Manager (CWGM). CWGM passes the process definitions to OWLS-SHOP2 translator which translates them into SHOP2 domain description. The translation algorithm was put forward and proved for soundness and correctness by Sirin *et al.* [11]. The simple processes are translated into operators while composite processes are translated into methods. OWLS-SHOP2 translator also translates high level goals into SHOP2 problem.

Intelligent agents are autonomous problem solving entities that can take input about the state of their environment and fulfill a specific role by acting on the environment in which they are situated [1]. Intelligent agents decentralize complex problems, define a simplified model to focus on specific details and define and manage the inter-relationships between different problem solving entities; hence making it easier to build complex software systems. CWGM creates an autonomous intelligent agent for each organization, which acts on behalf of its respective organization. Each agent has an instance of SHOP2 planner. The problem and the translated domain description are passed to the respective agents. Each agent will add a function call to itself in the precondition of each operator in the domain description of its respective organization. This way a step can only be added in a plan if the agent permits it to be added in the plan. The permission by the agent is based on its interaction with all other agents working on behalf of the interacting organizations. The agent only permits a step to be added in the plan if it does not make the plan incompatible with the plans of other interacting organizations. If a specific step makes the plan incompatible with the other plans then an alternative step is tried. The querying and backtracking mechanism presented by Au *et al.* is followed for this purpose [12].

The generated compatible SHOP2 plans are then transferred to SHOP2toOWLS converter to convert the plans from SHOP2 format into OWL-S format. OWL-S plans are passed back to the respective agents for execution with the help of Runtime Enactment Manager, which ensures that the transfer of information and files between the interacting organizations happens smoothly.

3 Implementation

The current prototype is able to create SHOP2 plans from high level user requirements and goals. The system takes a process definition and high level goals as input from user, translates the process definition into SHOP2 domain description and high

level goals into SHOP2 problem. Then the system creates all possible plans from the SHOP2 domain description to solve the generated SHOP2 problem. The OWLS-SHOP2 module from Transplan [14] is used in the system. Transplan is further based on the algorithm proposed by Sirin *et al.* [11]. The Java version of SHOP2 (JSHOP2) is used as a planner. Currently work is underway on extending the system to handle more than one organization so that based on high level requirements of interacting organizations, compatible workflows can be generated.

4 Related Work

Sirin *et al.*[15] presented a semi automatic system for web service composition. Sirin *et al.* [11] later extended his work to propose a fully automated system for web services composition. The system takes high level tasks, uses SHOP2 to create a plan to achieve them and then executes the plan using web services from the web. The authors have argued that web service composition is an AI planning problem. The framework proposed in this paper also generates all possible plans from an OWLS process definition and execute them using web services from the net. But unlike the proposed framework, system developed by Sirin *et al.* only targets web service composition for a single organization and does not take cross-organizational collaboration into account.

Chen and Chung [4] proposed a framework for cross-organizational workflow collaboration. The framework proposed by Chen and Chung detects incompatibilities between workflows, suggest changes to the workflows to remove the detected incompatibilities and ask the users in turn to see whether they would accept the suggested changes. Although their work saves considerable amount of time by automating workflow collaboration but users are still required to model workflows beforehand. Wang *et al.* [16] have also proposed a system for cross-organizational workflow collaboration and dynamic workflow composition. They use intelligent agents for discovering, executing and monitoring web services. They also use intelligent agents for dynamically composing workflows and negotiating over the net. Work done by Wang *et al.* is closely related to the work reported in this paper because it uses OWLS ontology to compose workflow at runtime and uses agents for negotiation. Wang *et al.* target collaboration among organization to take an alternative path in the workflow when certain service fails to achieve the desired goal in the workflow while the system proposed in this paper targets collaboration to generate compatible workflows from high level requirements.

5 Conclusion

This paper proposes a framework for cross-organizational workflow collaboration. It is different from existing systems because it automatically creates compatible workflows from high level goals and requirements from organizations that intend to collaborate. Other approaches aim to modify existing workflows of interacting organizations to make them compatible. This paper identifies workflow generation as AI planning problem and focuses on collaboration while generating the workflows.

References

- [1] Jennings, N.R.: An agent-based approach for building complex software systems. *Communications of the ACM* 44(4), 35–41 (2001)
- [2] Workflow Management Coalition: Terminology & Glossary. Technical Report WFMCTC-1011 (1999)
- [3] Yang, J., Papazoglou, M.: Interoperation Support for Electronic Business. *Communication of the ACM* 43(6), 39–47 (2000)
- [4] Chen, X., Chung, P.W.H.: Facilitating B2B E-Business by IT-Supported Business Process Negotiation Services. In: *Proceedings of the 2008 IEEE International Conference on Service Operations and Logistics and Informatics*, pp. 2800–2805 (2008)
- [5] Schulz, K., Orłowska, M.: Facilitating cross-organizational workflows with a workflow view approach. *Data and Knowledge Engineering* 51(1), 109–147 (2004)
- [6] Chiu, D.K.W., Cheung, S.C., Karlapalem, K., Li, Q., Till, S., Kafeza, E.: Workflow View Driven Cross-Organizational Interoperability in a Web-Services Environment. *Information Technology and Management* 5, 221–250 (2004)
- [7] Dong, X., Wild, D.: An Automatic Drug Discovery Workflow Generation Tool using Semantic Web Technologies. In: *Proceedings of Fourth IEEE International Conference on eScience*, pp. 652–657 (2008)
- [8] Chen, X., Yang, L.: Applying AI Planning to Semantic Web Services for Workflow Generation. In: *Proceedings of the First International Conference on Semantics, Knowledge and Grid*. IEEE Computer Society, Washington (2005)
- [9] Nau, D., Muñoz-Avila, H., Cao, Y., Lotem, A., Mitchell, S.: Total-order planning with partially ordered subtasks. In: *IJCAI 2001*, Seattle (2001)
- [10] Nau, D., Au, T., Ilghami, O., Kuter, U., Murdock, J., Wu, D., Yaman, F.: SHOP2: An HTN planning system. *Journal of Artificial Intelligence Research*, 379–404 (2003)
- [11] Sirin, E., Parsia, B., Wu, D., Hendler, J., Nau, D.: HTN planning for web service composition using SHOP2. *Journal of Web Semantics* 1(4), 377–396 (2004)
- [12] Au, T.C., Nau, D., Subrahmanian, V.S.: Utilizing Volatile External Information during Planning. In: *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pp. 647–651 (2004)
- [13] OWL Services Coalition: OWL-S: Semantic markup for web services (2003), <http://www.daml.org/services/owl-s/1.0/>
- [14] Transplan, <http://sourceforge.net/projects/transplan/>
- [15] Sirin, E., Hendler, J., Parsia, B.: Semi-automatic composition of Web services using semantic descriptions. In: *Proceedings of Web Services: Modeling, Architecture and Infrastructure Workshop in Conjunction with ICEIS* (2003)
- [16] Wang, S.Y., Shen, W.M., Hao, Q.: An agent-based Web service workflow model for inter-enterprise collaboration. *Expert System with Applications*, 787–799 (2006)

Knowledge Granularity and Representation of Knowledge: Towards Knowledge Grid

Maria A. Mach and Mieczyslaw L. Owoc

Wroclaw University of Economics
Komandorska 118/120
53-345 Wroclaw, Poland
{maria.mach,mieczyslaw.owoc}@ue.wroc.pl

Abstract. Knowledge granularity, usually identified with the size of knowledge granules, seems to be real challenge for knowledge consumers as well as for knowledge creators. In this paper, relationships between knowledge granularity as a result of different ways of a knowledge representation are considered. The paper deals with the problem of developing knowledge grid in the context of encapsulation of knowledge including different dimensions and measures. The origin of the problem is discussed in the first section stressing flexibility of knowledge interpretations. Concepts of knowledge granularity (also from formal point of view) are presented in the next section. The nature of well represented knowledge is considered in the next chapter with references to granularity of knowledge. In the last part of the paper the question of formulating knowledge granularity in the context of knowledge grid concepts is discussed. This document comprising guidelines for authors is divided into several sections.

Keywords: Knowledge grid, representation of knowledge, knowledge granularity.

1 Introduction

Knowledge as the essential term defined in many ways (stressing its philosophical or domain roots and aspects) represents structured information with the ability of interpretation and application for different purposes. The most general knowledge definition states, that knowledge may be perceived as: (i) expertise, and skills acquired by a person through experience or education; the theoretical or practical understanding of a subject; (ii) what is known in a particular field or in total; facts and information; or (iii) awareness or familiarity gained by experience of a fact or situation. In the field of artificial intelligence, the second meaning of the notion is common. Nevertheless of the context discussed, knowledge should be considered as the crucial value creating rational base for the future acting for potential users (individual or organizational). This is worth to stress that in different areas of knowledge applications we mainly use some pieces of knowledge for performing particular tasks. On the other hand there is a natural tendency to define whole knowledge which is useful in supporting selected

activities in the broadly understood decision-making processes. Therefore a quest of knowledge scopes essential and useful in such a context seems to be obvious.

The main goal of the paper is to investigate solutions in defining knowledge pieces (as a part of the whole relevant knowledge) in order to support knowledge grid concepts. Considering explicit knowledge, which is available for different users, we tend to focus on elaborated knowledge representation techniques acceptable in the computer infrastructure. Knowledge grid identified with intelligent and available via computer network platform covering explicit knowledge resources and supporting on-demand services oriented on problem solving, decision-making and the like processes (comp. [Zhuge, 2002]). Therefore knowledge granularity, knowledge representation and knowledge grid are the crucial categories in this research.

2 Concepts of Knowledge Granularity

The concept of granularity (in general) comes from photography, and describes accuracy of pictorial presentation on film. In a more specific way, one may speak also of information resources granularity and knowledge granularity.

Granularity of information resources refers to size, decomposability, and extent to which a resource is intended to be used as a part of a larger one. A typical example of information resources granularity concerns hierarchy of data, as depicted in Fig. 1.

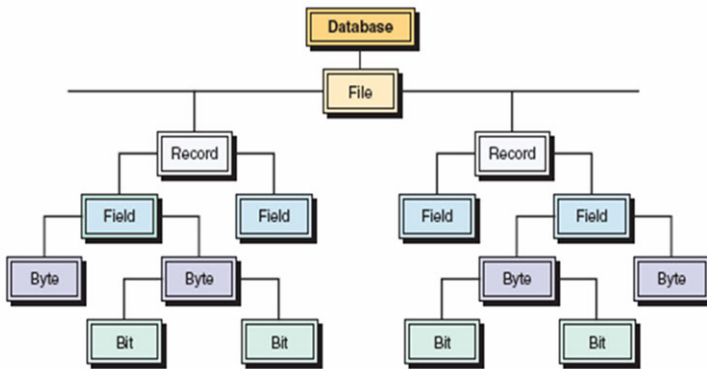


Fig. 1. Hierarchy of data as granularity of data

Source: [Turban et al., 2008], p. 88.

Knowledge granularity, in turn, is linked with dividing knowledge into pieces (so-called knowledge chunks), i.e. perceiving different levels of detail in knowledge structures.

The origins of knowledge granularity phenomenon are threefold. First, knowledge in modern information systems comes from many sources and is mapped in many ways. Second, there are many different forms of presentation of knowledge (e.g. semantic networks, rules, frames etc.). Finally, there are different domains of knowledge application: for example, different knowledge is needed on different management

levels – a CEO needs more “concentrated” knowledge, while a bottom-level user needs more details.

To the best of our knowledge, there is no formal specification of knowledge granularity in the literature. For the purposes of this paper, we adopted Bettini’s formalization of time granularity (may be found in [Bettini et al., 1998]). Not all the concepts of time granularity apply to knowledge granularity, because knowledge is of different nature than time. Nevertheless, some of Bettini’s concepts may be used.

By *knowledge domain* we understand a set of primitive knowledge entities concerning a particular problem. These entities may be for example raw data. All coarser knowledge chunks are generated from these primitive entities.

Knowledge granularity may be formally defined as the resolution power of qualification of a knowledge piece. A granularity is a mapping G from integers to subsets of knowledge domain, such that (see also [Bettini et al., 1998b]):

1. If $i < j$ and $G(i), G(j)$ are non-empty, then each element of $G(i)$ is less than all elements of $G(j)$. This is to state that granularities do not overlap;
2. If $i < k < j$ and $G(i), G(j)$ are non-empty, then $G(k)$ is non-empty. This rule means that the subset of the index set that maps to non-empty subsets of knowledge domain is contiguous. Each non-empty subset $G(i)$ is called a granule of granularity G . A granule may be composed e.g. of data.

Granularities and granules are interrelated in many ways. We may adopt three kinds of relationships from Bettini’s framework:

1. *Groups into* – a granularity G groups into a granularity H ($G \blacktriangleleft H$) if for each index j there exists a subset S of the integers such that $H(j) = \cup_{i \in S} G(i)$.

Example: granules of type “data” and “data label” group into a granule of type “information” (where information = data + label)

2. *Finer than* – a granularity G is finer than a granularity H ($G \leq H$) if for each index i , there exists an index j such that $G(i) \subseteq H(j)$.

Example: information on levels of sugar measured in blood (daily measure) is finer than information on HbA1c (level of sugar in a three-month period); a list of individual salaries is finer than average salaries grouped by department.

3. *Partitions* – a granularity G partitions a granularity H , if $G \blacktriangleleft H$ and $G \leq H$.

Having granularities and granules of knowledge, one may perform some kind of granular reasoning. This concerns using some particular functions that operate on granules. These functions enable:

- creating queries concerning main granular components (e.g. granular levels),
- extracting contents of levels,
- moving across levels, e.g. if we compute average salary of each department, we move from a more detailed (i.e. finer) knowledge level onto a coarser one.

We are aware that many knowledge representation methods are based on hierarchical classification, for example description logic based knowledge representation. Nevertheless using knowledge granularity concepts one may not only move across levels

and use different “pieces” of knowledge, but may also enrich the representation with the notion of time, a very important aspect of changing world.

Presented in brief the concept of knowledge granularity should convince us about great importance of implementing “knowledge pieces” in reality.

3 Selecting Knowledge Granularity in the Knowledge Representation Context

There is a natural tend to represent any part of reality (perhaps anything) sometimes called as upper ontology where generalized view of the world consists of intuitively formulated objects. Figure 2 depicts one of possible hierarchies representing of anything.

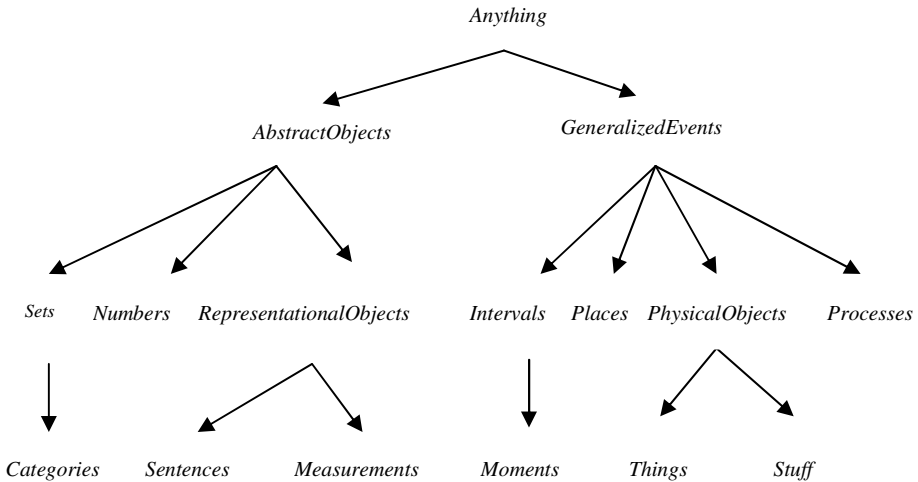


Fig. 2. The upper ontology of the world

Source: [Russell and Norvig, 2003], p.321

Concepts introduced in Figure 2 express in a simplified way a hierarchy of components existing in any domain covering static (*AbstractObjects*) as well as dynamic aspects of its functioning (*GeneralizedEvents*). At the upper level of ontology discussed we focus on unification of represented knowledge which allows for applying reasoning and problem solving techniques in more universal way.

Let us remind the main assumptions of properly represented knowledge. Nevertheless the techniques of knowledge representation we need to organize objects or events into categories which are interrelated in some way. Both sides of the hierarchy expressed in Figure covering selected details referring to *AbstractObjects* and *GeneralizedEvents*. This way materialized knowledge can be used through inheritance of features from higher levels of abstraction (e.g. *Sentences* and *Measurements* in case of

RepresentationalObjects or *Things* and *Staff* for *PhysicalObjects*). At any level of knowledge representation process we have to decide about knowledge granularity.

Knowledge granularity concept defined earlier can be used in practice for example in search agent purpose (see: Yiming and Tsotsos, 2001). Agents performing determined tasks use certain kind of knowledge recall single granularity or a set of hierarchical granularities in various situations. In some circumstances an agent is able to select the best knowledge granularity according to maximizing its performance. In the complex agent environment particular agent is able to apply granularity representing the corresponding knowledge from the demanding environment-granularity hash table.

The presented example of the “working” knowledge granularity in agent systems confirms usability of this approach. Keeping in memory some features of knowledge granularity we are able to select proper representation schema for the defined purpose.

4 Knowledge Grid and Knowledge Granularity

The concept of knowledge granularity can be applied in these approaches where we need to represent knowledge pieces in more flexible way. Therefore natural references of knowledge granularity can be found in the knowledge grid context.

According to quoted earlier definition proposed by H. Zhuge we assume that knowledge grid “is an intelligent and sustainable Internet application environment that enables people or virtual roles to effectively capture, coordinate, publish, understand, share and manage knowledge resources” (see: [Zhuge 2004]). Ultimate goal of knowledge grid systems is to provide services to support innovation, problem solving and decision making in a distributed environment, which varies in scale and stability.

For this purpose the need of developing a new architecture seems to be obvious. The Knowledge Grid as a new architecture based on the existing methods and technologies such as the Grid, the Semantic Web, Web Services Peer-to-Peer, AI, proper data modelling, information processing technologies and system methodology. Such system can be formulated as a three-layer model (see: [Hengsham, Liqun 2005]) consisting of the following layers:

- knowledge storage layer,
- knowledge services layer,
- knowledge grid application with interface for the end-user.

The *knowledge storage layer* corresponds mostly to the particular knowledge domain database. The pieces of knowledge are stored and can be acquired from there. This layer's function is to provide secure access to the knowledge items. Therefore all the mentioned before relationships in terms of knowledge granularity are actual: groups into, finer than and partitions. We create specific repository of knowledge assuming expecting level of knowledge granularity. Of course all necessary functions assuring knowledge storage functionality should be involved: syntactic analysis, searching, inquiring and extending knowledge bases.

The *knowledge services layer* supplies one view of heterogeneous knowledge sources and software systems, together with suitable software for knowledge discovery and reduction of redundant information. Furthermore, knowledge in this layer is used to improve query precision and to explain results to the end-user. The processes,

which help to intelligently eliminate, create and discover organisational knowledge, happen in this middle layer. All potential services are direct oriented on knowledge granules in terms of performing the itemized processes.

The *knowledge grid application layer* is responsible for delivering knowledge to a particular user. As before processes formulated by users recall some pieces of knowledge from the storage level. Again relationships among the formulated earlier knowledge granules determine ways of performing tasks by a user.

The presented in the paper concepts of knowledge granularity and knowledge grid create intersection of the research streams and confirm intuitively formulated relationships. Let us stress the placement of both concepts in broadly understood research areas of knowledge grid (compare: Zhuge 2004 and Owoc 2009):

- **Theories and methods for supporting knowledge management.** Defining the core phases of knowledge management: gathering, representation and sharing knowledge we should express the expecting level of knowledge granularity in the knowledge management process. The value of the knowledge grid approach in knowledge management solutions is common accepted from many reasons.
- **Ontology and semantic aspects of knowledge grid.** Nevertheless of the knowledge grid application area there are huge of problems with common understanding of the whole concept; domain terminology, interpretation of interrelationships, principles and references to other disciplines including flexible understanding of knowledge granularity concepts.
- **Knowledge grid in different institutions.** Propagation and management of knowledge within a virtual organization is one of the suggested hot topics. There are problems how implement knowledge grid in global organizations and on the other hand what kind of information infrastructure could be effective in case of hybrid and multilevel companies. One can assume specific approaches to formulating knowledge granularity in different companies.
- **knowledge grid and effectiveness of knowledge management.** Taking into consideration particular knowledge management phases: organization, evaluation, and improvement we are looking for efficient tools and techniques in order to support the whole cycle. The knowledge grid seems to be very promising in eliminating redundant knowledge and improving knowledge so that quality of useful knowledge pieces and knowledge as a whole should be better and better. It covers also such processes like: creation new knowledge from existing well-represented knowledge, from case histories, and from raw knowledge materials like texts. The role of knowledge granularity in these processes is very important.

Knowledge integration in the grid architecture. One of the promising features of the presented approach is gathering information and knowledge pieces from many sources. For example integrating knowledge resources could support analogies, problem solving, and scientific discovery so standards in this area are welcome. In every case we should define a proper lever of knowledge granularity.

5 Conclusions

This paper presents intuitively observed correlation between knowledge granularity and knowledge grid. Main research findings can be formulated in the following way:

1. Adopted from Bettini's framework, concept of granularity can be applied to formulating of knowledge granularity concepts. This approach can be compared with other proposals of knowledge granularity. Especially relationships formulated in this framework are useful in different domains.
2. Knowledge granularity is the very important component of broadly understood knowledge representation process. In more universal concept of the upper ontology of the world knowledge granularity plays the essential role.
3. In knowledge grid approaches a concept of knowledge granularity is present in many research areas. It is very difficult to create and manage knowledge grid without properly defined knowledge granularity concepts.

The further research can embrace investigation on determining standards and dimensions of knowledge granularity in knowledge grid architectures applied in different areas.

References

- Bettini, C., Wang, X., Jajodia, S.: A General Framework for Time Granularity and its Application to Temporal Reasoning. *Annals of Mathematics and Artificial Intelligence* 22(1,2), 29–58 (1998)
- Bettini, C., Dyreson, C.E., Evans, W.S., Snodgrass, R.T., Wang, X.S.: A Glossary of Time Granularity Concepts. In: Etzion, O., Jajodia, S., Sripada, S. (eds.) *Dagstuhl Seminar 1997*. LNCS, vol. 1399, pp. 406–413. Springer, Heidelberg (1998)
- Bittner, T., Smith, B.: *A Theory of Granular Partitions*. In: *Foundations of Geographic Information Science*
- Duckham, M., Goodchild, M.F., Worboys, M.F. (eds.): Taylor & Francis Books, London (2003)
- Benett, B.: Space, Time, Matter and Things. In: *Proc. FOIS 2001*. ACM Press, USA (2001)
- Bettini, C., Jajodia, S., Wang, S.X.: Time Granularities in Databases. In: *Data Mining, and Temporal Reasoning*, Springer, Berlin (2000)
- Duncan, C.: Granularisation. In: Littlejohn, A. (ed.) *Reusing Online Resources: A Sustainable Approach to eLearning*. Kogan Page, London (2003)
- Feng, Q., Miao, D., Zhou, J., Cheng, Y.: A Novel Measure of Knowledge Granularity in Rough Sets
- Goralwalla, I.A., Leontiev, Y., Ozsu, T.M., Szafron, D.: Temporal Granularity: Completing the Puzzle. *Journal of Intelligent Information Systems* 16, 41–46 (2001)
- Huang, F., Zhang, S.: Clustering Web Documents Based on Knowledge Granularity. *Asian Journal on Information Technology* 5(1) (2006)

- Keet, C.M.: A Formal Theory of Granularity. PhD Thesis, KRDB Research Centre, Faculty of Computer Science, Free University of Bozen-Bolzano, Italy (2008)
- Keet, C.M.: A taxonomy of types of granularity. In: IEEE Conference in Granular Computing (GrC 2006), Atlanta, USA, May 10-12 (2006)
- Kamble, A.S.: A Data Warehouse Conceptual Data Model for Multidimensional Information. PhD thesis, University of Manchester, UK (2004)
- Mach, M.A., Owoc, M.L.: Granularity of Knowledge from Different Sources. In: Intelligent Information Processing IV, IFIP - The International Federation for Information Processing, vol. 288 (2009)
- Mach, M., Owoc, M.L.: About Dimension and Measures of Knowledge Granularity. In: Tadeusiewicz, R., Ligeza, A., Szymkat, M. (eds.) Computer Methods and Systems, Kraków, vol. I (2009)
- Mani, I.: A theory of granularity and its application to problems of polysemy and underspecification of meaning. In: Cohn, A.G., Schubert, L.K., Shapiro, S.C. (eds.) Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning, KR 1998. Morgan Kaufmann, San Mateo (1998)
- Owoc, M.L.: Research trends in knowledge grid. In: Nowicki, A. (ed.) Business Informatics 13. Research Papers No 55 of Wroclaw University of Economics (2009)
- Pawlak, Z.: Granularity of knowledge, indiscernibility and rough sets. In: IEEE World Congress on Computational Intelligence, Fuzzy Systems Proceedings, vol. 1 (1998)
- Russell, S., Norvig, P.: Artificial Intelligence. A Modern Approach. Pearson Education International (2003)
- Turban, E., Leidner, D., Mclean, E., Wetherbe, J.: Information Technology for Management. In: Transforming Organizations in the Digital Economy, John Wiley & Sons, Chichester (2008)
- Yiming, Y., Tsotsos, J.: Knowledge granularity spectrum, action pyramid, and the scaling problem. *International Journal of Pattern Recognition and Artificial Intelligence* 15(3) (2001)
- Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets and System* 19(1) (1997)
- Zhao, M., Yang, Q., Gao, D.: Axiomatic Definition of Knowledge Granularity and its Constructive Method. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 348–354. Springer, Heidelberg (2008)
- Zhugue, H.: The Knowledge Grid. World Scientific, Singapore (2004)

Combining the Missing Link: An Incremental Topic Model of Document Content and Hyperlink

Huifang Ma^{1,2}, Zhixin Li^{1,2}, and Zhongzhi Shi¹

¹ Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, 100080, Beijing, China

² Graduate School of the Chinese Academy of Sciences, 100039, Beijing, China
{mahf, lizhixin, shizz}@ics.ict.ac.cn

Abstract. The content and structure of linked information such as sets of web pages or research paper archives are dynamic and keep on changing. Even though different methods are proposed to exploit both the link structure and the content information, no existing approach can effectively deal with this evolution. We propose a novel joint model, called Link-IPLSI, to combine texts and links in a topic modeling framework incrementally. The model takes advantage of a novel link updating technique that can cope with dynamic changes of online document streams in a faster and scalable way. Furthermore, an adaptive asymmetric learning method is adopted to freely control the assignment of weights to terms and citations. Experimental results on two different sources of online information demonstrate the time saving strength of our method and indicate that our model leads to systematic improvements in the quality of classification.

Keywords: Topic model; Link-IPLSI; Incremental Learning; Adaptive Asymmetric learning.

1 Introduction

Obtaining multi-side semantic information from a topic report containing dynamic online data streams is useful both from a theoretical point of view, as there are many complex phenomena to be addressed, and from purely practical applications such as topic modeling. A variety of techniques for automatically extracting thematic content of a set of documents are proposed, such as latent semantic indexing(LSI)[1], probabilistic latent semantic indexing (PLSI)[2]. The topics learned by a topic model can be regarded as themes discovered from documents sets, while the topic-term distributions focus on the high probability words that are relevant to a theme.

With lots of electronic documents connected with hyperlinks/citations can be easily and readily acquired through the Internet, scholars demonstrate an increasing academic interest concerning how to effectively construct models for these correlated hypertexts. Automatic techniques to analyze and mine these document collections are at the intersection of the work in link analysis [3, 4], hypertext and Web mining [5, 6]. The most well known algorithms in link mining are PageRank [7] and HITS [8]. Both algorithms exploit the hyperlinks of the Web to rank pages based on their levels of

“prestige” or “authority”. Link mining encompasses a wide range of tasks [9] and we focus on the core challenges addressed by a majority of ongoing research in the field of topic modeling.

There are many noteworthy works. Cohn and Chang [10] introduced PHITS as a probabilistic analogue of the HITS algorithm, attempting to explain the link structure in terms of a set of latent factors. One of the first efforts in applying topic models to modeling both citation and content came from Cohn and Hoffman [11], they constructed Link-PLSI to integrate content and connectivity together. Erosheva et al. [12] defined a generative model for hyperlinks and text and thereby modeled topic specific influence of documents. We refer to this model as Link-LDA. Nallapati et al. [13] addressed the problem of joint modeling of text and citations in the topic modeling framework and presented two different models. Gruber et al. [14] recently presented a probabilistic generative model LTHM for hypertext document collections that explicitly models the generation of links.

These methods, however, are not suitable to be applied to the changing situation as the links and documents are probably only valid at a certain time. When new documents and a set of inter-connections are added, the existing model should be updated correspondingly. A similar situation happens when part of old documents and citations are deleted. A naïve approach to catch the update of links and contents is to rerun the batch algorithm from scratch on all existing data each time new data comes in, which is computationally expensive. Another obvious shortcoming for the naïve approach is that after re-running of batch algorithm, changes to the links and contents themselves can not be captured with the content of latent topics maintained.

As for incremental learning of topic modeling, Chien et al. [15] proposed an incremental PLSI learning algorithm which efficiently updates PLSI parameters using the maximum a posterior. Chou et al. [16] introduced another Incremental PLSI (IPLSI), aiming to address the problem of online event detection. Although these models capture the basic concept of incremental learning for PLSI, their weakness is that they do not take additional link information into consideration. Even so, these models offer excellent foundations on which to build our model.

In this paper, we present a new model Link-IPLSI, which extends the existing Link-PLSI by modeling interactions between document and link structure incrementally. In contrast to PLSI and Link-PLSI, the new model processes incoming online documents incrementally for each time period, discards out-of-date documents, terms and links not used recently, and folds in new terms, links and documents with the latent semantic indices preserved from one time period to the next. To the best of our knowledge, there is no previous work constructing the interconnected documents incrementally.

This paper has the following technical contributions:

- Present an incremental Link-PLSI model, owning the ability to identify meaningful topics while reducing the amount of computations by maintaining latent topics incrementally.
- Extend Link-PLSI model for updating two modalities simultaneously, which supports addition/deletion for both terms and citations.
- By means of integrating link information, our incremental model takes advantage of adaptive asymmetric learning method to weigh terms and links respectively.

In a word, this paper presents an incremental topic model that is applicable to a set of dynamic interconnected data. We have applied this method to both cited and hyper-linked data sets. Experiments show that our method is effective for topic modeling and works more efficiently than the corresponding batch method.

The remainder of this paper is organized as follows: In Section 2, we introduce the Link-PLSI model and its principles. In Section 3, we give detailed information on our proposed Link-IPLSI model. Section 4 describes the test corpora, the performance measures and the baseline method together with the experiment results. We conclude and discuss future work in Section 5.

Note that in the rest of the paper, we use the terms ‘‘citation’’ and ‘‘hyperlink’’ interchangeably. Likewise, the term ‘‘citing’’ is synonymous to ‘‘linking’’ and so is ‘‘cited’’ to ‘‘linked’’ [13].

2 Link-PLSI Model

Link-PLSI [11] is based on the assumption that similar decomposition of the document term co-occurrence matrix can be applied to the cite-document co-occurrence matrix in which each entry is a count of appearances of a linked-document (or citation) in a source document. Under this assumption, a document is modeled as a mixture of latent topics that generates both terms and citations. A representation of Link-PLSI model is depicted in Fig. 1.

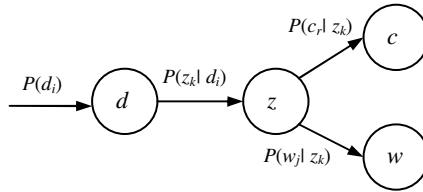


Fig. 1. Representation of Link-PLSI model

The model is defined as a generative process: a document d_i is generated with some probability $P(d_i)$, a latent topic z_k associated with documents, terms and citations is chosen probabilistically so that their association can be represented as conditional probabilities $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$. The following joint model for predicting citations/links and terms in documents is defined as:

$$P(c_r | d_i) = \sum_k P(c_r | z_k)P(z_k | d_i), \tag{1}$$

$$P(w_j | d_i) = \sum_k P(w_j | z_k)P(z_k | d_i), \tag{2}$$

where w_j represents one term and c_r indicates one citation, c and d both refer to document in the corpus and they may be identical. They are kept separate notationally to reinforce different roles they play in the model, c is conveyed by being cited and d is conveyed by citing [11].

An EM algorithm is used to compute the parameters $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$ through maximizing the following log-likelihood function with a relative weight α of the observed data:

$$L = \sum_i [\alpha \sum_j \frac{N_{ji}}{\sum_{j'} N_{j'i}} \log \sum_k P(w_j | z_k) P(z_k | d_i) + (1 - \alpha) \sum_r \frac{A_{ri}}{\sum_{r'} A_{r'i}} \log \sum_k P(c_r | z_k) P(z_k | d_i)], \tag{3}$$

where N_{ji} is the count of term w_j in document d_i and A_{ri} is the count of citation c_r from document d_i . The steps of the EM algorithm are described as follows:

E-step. The conditional distributions $P(z_k|d_i, w_j)$ and $P(z_k|d_i, c_r)$ are computed from the previous estimate value of the parameters $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$:

$$P(z_k | d_i, w_j) = \frac{P(z_k | d_i) P(w_j | z_k)}{\sum_k P(z_k | d_i) P(w_j | z_k)}, \tag{4}$$

$$P(z_k | d_i, c_r) = \frac{P(z_k | d_i) P(c_r | z_k)}{\sum_k P(z_k | d_i) P(c_r | z_k)}. \tag{5}$$

M-step. The parameters $P(w_j|z_k)$, $P(c_r|z_k)$ and $P(z_k|d_i)$ are updated with the new expected values $P(z_k|d_i, w_j)$ and $P(z_k|d_i, c_r)$:

$$P(w_j | z_k) = \sum_i \frac{N_{ji}}{\sum_{j'} N_{j'i}} P(z_k | d_i, w_j), \tag{6}$$

$$P(c_r | z_k) = \sum_i \frac{A_{ri}}{\sum_{r'} A_{r'i}} P(z_k | d_i, c_r), \tag{7}$$

along with the mixing proportions

$$P(z_k | d_i) \propto \alpha \sum_j \frac{N_{ji}}{\sum_{j'} N_{j'i}} P(z_k | d_i, w_j) + (1 - \alpha) \sum_r \frac{A_{ri}}{\sum_{r'} A_{r'i}} P(z_k | d_i, c_r). \tag{8}$$

3 A Joint Incremental Link-PLSI for Content and Hyperlink

The topic modeling process often requires simultaneous model construction and testing in an environment which constantly evolves over time. It is assumed that the most

effective topic model to be used under such environment does not stay constant over time, but varies with progression of the data stream.

For the effective update of contents and links when new documents are added or old linked-data are removed, we propose an incremental approach to Link-PLSI technique, which is referred to as Link-IPLSI. The basic idea of our updating algorithm is straightforward: the Link-IPLSI model is performed on the initial linked-documents at the beginning. When a set of new documents are added introducing new terms and citations, a cycle will be created for folding in these documents, terms and citations and the model is then updated during the cycle. For new adding of documents or removing of old ones, our model adjusts both term-topic and link-topic probabilities at the lowest cost.

3.1 Preprocessing

The preprocessing phase is the first step for the incremental learning, involving elimination of out-of-date documents, terms and hyperlinks. The corresponding parameters $P(w_{out}|z)$, $P(c_{out}|z)$ and $P(z|d_{out})$ are removed. (d_{out} is an out-of-date document, and so are w_{out} and c_{out}) We can not simply augment the model directly, as the basic principle of probability that the total probability will be equal to one should be observed, the remaining parameters need to be renormalized proportionally:

$$P(w|z) = \frac{P_0(w|z)}{\sum_{w \in W_0} P_0(w|z)}, \tag{9}$$

$$P(c|z) = \frac{P_0(c|z)}{\sum_{c \in C_0} P_0(c|z)}, \tag{10}$$

where $P_0(w|z)$ and $P_0(c|z)$ stand for the probabilities of the remaining terms and citations, whereas W_0 and C_0 are the respective sets of remaining terms and citations.

3.1 Incremental Link-PLSI Technique

In this section, we give a detailed illustration of Link-IPLSI. The novelty of our model is that it takes advantage of the existing information to handle the streaming data without retraining the model. Therefore, the model is much faster and more scalable which makes model construction easier than that of the batch system. Fig. 2 is an illustration of sequences for updating related information of Link-IPLSI, where d' and w' indicate new documents and new terms respectively.

As the figure shows, new documents should first be folded in with old terms and links fixed, and then $P(d'|z_k)$ are calculated which sets a foundation for folding in new terms and links in the followings. In this way, $P(w_{all}|z_k)$, $P(c_{all}|z_k)$ and $P(z_k|d_{all})$ are updated as better initial values for the final EM algorithm, which guarantees a faster convergence. (d_{all} is a final document in the entire document set, and so are w_{all} and c_{all}). A specified illustration is given below.

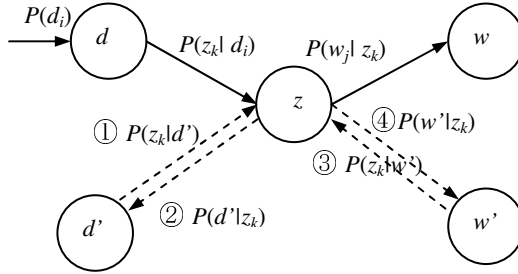


Fig. 2. Illustration of Link-IPLSI model

Fold in new document. There is a need to realize how many data have already been well explained by the existing model in order to integrate the streaming data into the model effectively. Using a partial version of EM algorithm, folding-in, we can update the unknown parameters with the known parameters fixed so as to maximize the likelihood with respect to the previously trained parameters. Obviously, documents should first be folded in, since old terms/links are well trained and the arriving documents contain old terms/links while old documents convey no corresponding information to aid the folding in of new terms and links.

For new documents d_{new} , we first randomize and normalize $P(z|d_{new})$. Thereafter $P(z|d_{new})$ are updated through fusion of $P(w|z)$ and $P(c|z)$ in the follows:

E-step. The conditional distributions $P(z|d_{new}, w)$ and $P(z|d_{new}, c)$ are obtained from the previous estimate value of the parameters $P(z|d_{new})$:

$$P(z | d_{new}, w) = \frac{P(w | z)P(z | d_{new})}{\sum_k P(w | z_k)P(z_k | d_{new})}, \tag{11}$$

$$P(z | d_{new}, c) = \frac{P(c | z)P(z | d_{new})}{\sum_k P(c | z_k)P(z_k | d_{new})}. \tag{12}$$

M-step. The parameters $P(z|d_{new})$ are updated with the new expected values $P(z|d_{new}, w)$ and $P(z|d_{new}, c)$:

$$P(z | d_{new}) \propto \alpha \sum_{j \in d_{new}} \frac{N_{j,new}}{\sum_{j \in d_{new}} N_{j,new}} P(z | d_{new}, w_j) + (1 - \alpha) \sum_{r \in d_{new}} \frac{A_{r,new}}{\sum_{r \in d_{new}} A_{r,new}} P(z | d_{new}, c_r). \tag{13}$$

Link-PLSI assumes that terms and citations make different contributions in defining the latent topic. The only potential imbalance could result from the mix parameter α between different terms and citations while these values cannot be freely controlled. Unlike Link-PLSI, our method allows to assign weights to each modality

according to the average amount of information it offers. This concretely allows modeling of a document as a mixture of latent topics that is defined by the relative importance of its terms and its citation patterns, resulting in different mixtures.

Specifically, we use entropy as a criterion for weight assignment as entropy is useful for evaluating the average amount of information needed to specify the state of a random variable. The idea is quite straightforward as the distributions over terms in each document can be good indications for their informativeness. Distributions of terms in each document that are sharply peaked around a few values will have relatively low entropy, whereas those that are spread more evenly across different values will have higher entropy. The entropy of term feature distribution of a specific document is defined as:

$$H(d_i) = -\sum_j \frac{N_{ij}}{N_i} \log \frac{N_{ij}}{N_i}, \tag{14}$$

where N_i is the total number of feature terms in document d_i . The fusion parameter α is then defined according to our empirical formula:

$$\alpha = \begin{cases} 1 & H(d_i) \leq \theta, \\ \exp(\theta - H(d_i)) & H(d_i) > \theta. \end{cases} \tag{15}$$

where θ equals to the average entropy of the entire document set.

Fold in new terms and hyperlinks. In this step we consider the problem of how to fold in new terms and citations. It is a pity that they can not be folded in by using $P(z|d_{new})$ directly, it is because the sum of all probabilities of terms/citations in old term/citation sets under z already equals to one, which means $P(w|z)$ and $P(c|z)$ have been well trained and normalized. If we randomize and normalize all $P(w_{new}|z)$ and $P(c_{new}|z)$ when new documents arrive, the sum of the probabilities of all terms/citations under z will be larger than one. This restriction makes it inapplicable to update new terms and citations directly. To avoid this, we first derive $P(d_{new}|z)$ in the following way:

$$P(z | d_{new}, w) = \frac{P(w | z)P(z | d_{new})}{\sum_k P(w | z_k)P(z_k | d_{new})}, \tag{16}$$

$$P(z | d_{new}, c) = \frac{P(c | z)P(z | d_{new})}{\sum_k P(c | z_k)P(z_k | d_{new})}, \tag{17}$$

$$\begin{aligned} P(d_{new} | z) \propto & \alpha \sum_{j \in d_{new}} \frac{N_{j,new}}{\sum_{i \in D_{new}} N_{ji}} P(z | d_{new}, w_j) \\ & + (1 - \alpha) \sum_{r \in d_{new}} \frac{A_{r,new}}{\sum_{i \in D_{new}} A_{ri}} P(z | d_{new}, c_r), \end{aligned} \tag{18}$$

where D_{new} is the set of new documents. Thereafter we need to develop a mechanism for new terms/citations update that satisfies the basic principle of topics under new terms/citations equal to one. $P(z|w_{new})$ and $P(z|c_{new})$ are randomly initialized and normalized. We then update $P(z|w_{new})$ and $P(z|c_{new})$ with the above $P(d_{new}|z)$ fixed.

E-step. The conditional distributions $P(z|d_{new}, w_{new})$ and $P(z|d_{new}, c_{new})$ are calculated from the previous estimate value of the parameters $P(z|w_{new})$ and $P(z|c_{new})$:

$$P(z | d_{new}, w_{new}) = \frac{P(z | w_{new})P(d_{new} | z)}{\sum_k P(z_k | w_{new})P(d_{new} | z_k)}, \tag{19}$$

$$P(z | d_{new}, c_{new}) = \frac{P(z | c_{new})P(d_{new} | z)}{\sum_k P(z_k | c_{new})P(d_{new} | z_k)}. \tag{20}$$

M-step. The parameters $P(z|w_{new})$ and $P(z|c_{new})$ are updated with the new expected values $P(z|d_{new}, w_{new})$ and $P(z|d_{new}, c_{new})$:

$$P(z | w_{new}) = \sum_{i \in D_{new}} \frac{N_{new,i} P(z | d_i, w_{new})}{\sum_{i' \in D_{new}} N_{new,i'}}, \tag{21}$$

$$P(z | c_{new}) = \sum_{i \in D_{new}} \frac{A_{new,i} P(z | d_i, c_{new})}{\sum_{i' \in D_{new}} A_{new,i'}}. \tag{22}$$

We can add the corresponding parameters of w_{new} , c_{new} and d_{new} reasonably at different times in this way.

Update the parameters. The third step of our incremental algorithm deals with the issues of how to calculate $P(w_{new}|z)$ and $P(c_{new}|z)$ and how to get the final normalized $P(w_{all}|z)$ and $P(c_{all}|z)$ by means of adjusting $P(w|z)$ and $P(c|z)$. Following the basic principle of the total probability of terms/citations in the entire terms/citations sets under z should equal to one, we normalize $P(w_{all}|z)$ and $P(c_{all}|z)$ as:

$$P(w_{all} | z) = \sum_{i \in D \cup D_{new}} \frac{N_{all,i}}{\sum_{j' \in d_i} N_{j',i}} P(z | d_i, w_{all}), \tag{23}$$

$$P(c_{all} | z) = \sum_{i \in D \cup D_{new}} \frac{A_{all,i}}{\sum_{r \in d_i} A_{r,i}} P(z | d_i, c_{all}). \tag{24}$$

For new terms w_{new} and new citations c_{new} , $P(z|d, w)$ and $P(z|d, c)$ are calculated according to formula (19) and (20) while for old terms and citations, we use formula (4) and (5) to get $P(z|d, w)$ and $P(z|d, c)$.

In the last step, we use the above parameters to execute the original Link-PLSI model for updating the model. As new documents arrive and old documents disappear, the above Link-IPLSI model can preserve the probability and continuity of the latent parameters during each revision of the model in a fast way.

4 Experimental Results

In this section, our empirical evaluation on the performance of our approach is presented. In all experiments, we used a PC with an Intel core2 duo p8400 3GHz CPU, 2G bytes of memory on the Windows XP Professional SP2 platform. We designed three experiments to test the viability of our model: time expenditure by comparing execution time with the Naïve Link-IPLSI; some preliminary results to demonstrate the performance of our algorithm in classification, which indicates the increased power of our adaptive learning of fusion parameter.

Data description. The performance of our model was evaluated using two different types of linked data: scientific literature from Citeseer which is connected with citations, Wikipedia webpages and WebKB dataset containing hyperlinks. We first adjust the link structure to include the incoming links and outgoing links only within each corpus, and then take advantage of these dataset for our model construction with adding new documents and citations and deleting out-of-date information.

The Citeseer data can be obtained from Citeseer collection that was made publicly available by Lise Getoor’s research group at University of Maryland. There are altogether 3312 documents using abstract, title and citation information in the corpus with the vocabulary of 3703 unique words. The Citeseer dataset only includes articles that cite or are cited by at least two other documents. Thereafter the corpus size is limited to 1168 documents, of which only 168 documents have both incoming and outgoing links. The WebKB dataset contains approximately 6000 html pages from computer science departments of four schools (Cornell, Texas, Washington, and Wisconsin). The dictionary contains 2800 words in the WebKB domain and 9843 links. The dataset of Wikipedia webpages is downloaded from Wikipedia by crawling within the Wikipedia domain, starting from the “Artificial Intelligence” Wikipedia page and the dataset is composed of 1042 documents and 4912 links with the vocabulary of 3072 words.

Experiments on time cost. To evaluate time efficiency of Naïve Link-IPLSI and Link-IPLSI, we ran these two algorithms on the subset of each database consisting of 90% of its entire documents respectively. We constructed five perturbed versions of the databases, containing a randomly deleted 10% subset of the original documents and adding of the same amount of data. Remind that the time cost depends highly on the number of topics k , we examined the impact of k in the experiment. For each k , we ran the Naïve Link-IPLSI and Link-IPLSI on each dataset mentioned above, the average number of iterations to reach convergence and the total time spent on model construction are recorded. Table 1 gives a detailed illustration on the total time and the number of iterations required to achieve convergence. (The total time of Link-IPLSI is divided into two parts: Link-PLSI time and folding time).

As seen in this table, the Link-IPLSI method can save a large amount of time compared with the naïve method. In general, the computation time of the Naïve Link-IPLSI approach is much longer than that of our model. With $k=30$ on WebKB dataset, Link-IPLSI can reduce the time cost by more than 13 times. The reason is that the

Table1. Execution time (in seconds) of Naïve Link-IPLSI and Link-IPLSI

k	WebKB				Citeseer				Wiki			
	NLI		LI		NLI		LI		NLI		LI	
	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time	Aver Iter	Total Time
10	42.11	7290	8.42	728	38.12	3072	7.87	310	42.97	3019	8.34	298
15	41.32	8598	8.13	818	42.31	3912	8.23	324	41.38	3718	8.37	307
20	43.46	11921	7.91	862	42.19	6184	8.12	373	45.23	5615	8.11	352
25	39.12	13063	8.61	943	47.81	7045	9.12	418	41.22	6412	8.01	384
30	41.32	15532	8.11	1131	55.39	8280	10.03	523	45.87	7123	8.77	472

Note: NLI stands for Naïve Link-IPLSI, LI stands for Link-IPLSI, Aver.Iter stands for Average Iterations; k indicates number of latent topics.

Naïve Link-IPLSI approach uses new random initial settings to re-estimate all relevant parameters of EM algorithm each time and requires a large number of iterations to converge to a different local optimum while Link-IPLSI has preserved a better starting point and can therefore converge much faster than the Naïve Link-IPLSI approach. The larger the dataset is, the more time our model can save. This is because the key point of Link-IPLSI is to reduce the EM iteration cost on more estimated parameters. Furthermore, when k increases, time cost increases as well, these results are consistent with our intuition.

Classification. Apart from its superior performance in time saving, another attractive feature of our model is its stability of latent topics. In this section, we use three Link-IPLSI variant models, that is, Link-PLSI, Naïve Link-IPLSI and the Link-IPLSI without learning of fusion parameter, together with Link-LDA and LTHM as baseline for comparison. Besides, we use Adaptive Link-IPLSI to denote our model for using adaptive asymmetric learning mechanism.

We perform classification on the WebKB dataset and Citeseer dataset. The task of this experiment is to classify the data based on their content information and link structure. From the original datasets, five perturbed versions of the datasets were created. We randomly split each dataset into five folds and repeat the experiment for five times, for each time we use one fold for test, four other folds for training incrementally. To give these models more advantage, we set the number of latent topics to be seven and six on WebKB and Citeseer respectively which correspond to the exact number of clusters. Classification accuracy is adopted as the evaluation metric, which is defined as the percentage of the number of correct classified documents in the entire data set. We demonstrate the average classification accuracies and its standard deviation over the five repeats as results. Since the accuracy of the Link-PLSI model depends on the parameter α , we use the average classification accuracies for Link-IPLSI.

Table 2 shows the average classification accuracies on different datasets using different methods. From Table 2 we can see that the accuracies of Naïve Link-IPLSI and Link-PLSI are worse than that of Link-IPLSI and our model. Specifically, Though Link-IPLSI performs slightly better than other variant models of Link-PLSI, our method of Adaptive Link-IPLSI clearly outperform all other models and receives the highest accuracy among all these approaches. As described above, the latent variables

generated by the Naïve- Link-IPLSI algorithm are discontinuous, whereas the latent variables generated by our algorithm are continuous. This shows that latent continuity can improve the performance of classification. The difference between the results of Link-LDA, LTHM and Adaptive Link-IPLSI is significant. This indicates that the enhanced classification performance is largely attributed to the adaptive weighing mechanism, i.e. the automatically obtained reasonable parameter α plays an important role in the improvement of classification.

Table 2. Classification accuracy (mean \pm std-err %) on WebKB data set and Citeseer data set

Method	WebKB	Citeseer
Naïve Link-IPLSI	0.332 \pm 0.90	0.453 \pm 0.68
Link-PLSI	0.358 \pm 0.88	0.478 \pm 0.75
Link-IPLSI	0.371 \pm 0.87	0.481 \pm 0.83
Link-LDA	0.382 \pm 0.77	0.501 \pm 0.90
LTHM	0.411 \pm 0.67	0.534 \pm 0.52
Adaptive Link-IPLSI	0.431 \pm 0.81	0.562 \pm 0.81

5 Conclusion

Existing topic model cannot effectively update itself when changes occur. In this paper, we developed an incremental technique to update the hyperlinked information in a dynamic environment. Our technique computes and updates corresponding parameters by analyzing changes to linked documents and by re-using the results from previous Link-PLSI computation. Besides, our model can assign weights to terms and citations by means of adaptive asymmetric learning mechanism. In addition, we have demonstrated its faster and scalable performance on three distinctive dataset and illustrated preliminary results of our model in classification. However, our model learns the asymmetric fusion parameter through empirical formula hence further theoretical analysis is needed. Meanwhile, the number of latent topics of our model is fixed which is inconsistent with human intuition. Extending the model to grow or shrink as needed that permits easier model selection is our future work.

Acknowledgments. This work is supported by the National Science Foundation of China (No. 60933004, 60903141), the National Basic Research Priorities Programme (No. 2007CB311004), 863 National High-Tech Program (No.2007AA01Z132), and National Science and Technology Support Plan (No.2006BAC08B06).

References

- [1] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic Analysis. *Journal of the American Society for Information Science* 41, 391–407 (1990)

- [2] Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 177–196 (2001)
- [3] Jensen, D., Goldberg, H.: *AAAI Fall Symposium on AI and Link Analysis*. AAAI Press, Menlo Park (1998)
- [4] Feldman, R.: Link analysis: Current state of the art. In: *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 23–26 (2002)
- [5] Chakrabarti, S.: *Mining the Web*. Morgan Kaufmann, San Francisco (2002)
- [6] Shi, Z.Z., Ma, H.F., He, Q.: Web Mining: Extracting Knowledge from the World Wide Web. In: *Data Mining for Business Applications*, pp. 197–209. Springer, Heidelberg (2008)
- [7] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University (1998)
- [8] Kleinberg, J.: Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46, 604–632 (1999)
- [9] Getoor, L., Diehl, C.P.: Link Mining: A Survey. *ACM SIGKDD Explorations Newsletter* 7, 3–12 (2005)
- [10] Cohn, D., Chang, H.: Learning to Probabilistically Identify Authoritative Documents. In: *Proceedings of International Conference on Machine Learning*, pp. 167–174 (2000)
- [11] Cohn, D., Hofmann, T.: The missing link-A probabilistic model of document content and hypertext connectivity. In: *Advances in Neural Information Processing Systems*, pp. 430–436 (2001)
- [12] Erosheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 5220–5227 (2004)
- [13] Nallapati, R., Ahmed, A., Xing, E.P., Cohen, W.W.: Joint Latent Topic Models for Text and Citations. In: *Proceedings of International Conference on Knowledge Discovery and Data Mining*, pp. 542–550 (2008)
- [14] Gruber, A., Rosen-Zvi, M., Weiss, Y.: Latent Topic Models for Hypertext. In: *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 230–240 (2008)
- [15] Chien, J.T., Wu, M.S.: Adaptive Bayesian Latent Semantic Analysis. *IEEE Transactions on Audio, Speech, and Language* 16, 198–207 (2008)
- [16] Chou, T.C., Chen, M.C.: Using Incremental PLSA for Threshold Resilient Online Event Analysis. *IEEE Transaction on Knowledge and Data Engineering* 20, 289–299 (2008)

A General Approach to Extracting Full Names and Abbreviations for Chinese Entities from the Web

Guang Jiang^{1,2}, Cao Cungen¹, Sui Yuefei¹, Han Lu^{1,2}, and Shi Wang¹

¹ Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road, Zhongguancun, Beijing 100190, China

² Graduate University of Chinese Academy of Sciences, No.19 Yuquan Road, Shi Jing Shan Distinct, Beijing 100049, China

Abstract. Identifying Full names/abbreviations for entities is a challenging problem in many applications, e.g. question answering and information retrieval. In this paper, we propose a general extraction method of extracting full names/abbreviations from Chinese Web corpora. For a given entity, we construct forward and backward query items and commit them to a search engine (e.g. Google), and utilize search results to extract full names and abbreviations for the entity. To verify the results, filtering and marking methods are used to sort all the results. Experiments show that our method achieves precision of 84.7% for abbreviations, and 77.0% for full names.

1 Introduction

Named Entity Recognition (NER) is a basic task in text mining; it is significant for information extraction, machine translation etc. in nature language processing (NLP). In 1998, MUC (Message Understanding Conference) defined seven categories of named entity task belong to three subtasks: entity names (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages). Named Entity Recognition is a challenging topic in NLP because named entities are huge and increase as time goes on.

Among the seven categories above, organizations, persons and locations are three most import types, therefore identification of these three categories become hot points of research (Maynard et al., 2000; Luo et al., 2003; Wang et al., 2007). In view of full names and abbreviations, there are two kinds of entities: full name entity and abbreviation entity. Identification of full name and abbreviation for person names and locations are easier than which of organizations, because they embody more regular laws: person name begins with a family name, e.g. (Manager Zhang), while location names are finite and their abbreviations usually are fixed usually. E.g., (Jing) is an abbreviation of (Beijing). However, organizations are more difficult to identify because they are often long, loose and flexible or sometimes contain locations.

Identification of named entities is prepared for extracting relation of different entities, which is significant in several fields such as question answering (QA),

multi-word expression (MWE) and information retrieval (IR). Automatic content extraction (ACE)¹ evaluation plan organized by the National institute of standards and technology (NIST) had defined seven types of entity relation: ART (artifact), GEN-AFF (Gen-affiliation), METONYMY, ORG-AFF (Org-affiliation), PART-WHOLE, PRE-SOC (person-social) and PHYS (physical). They also can be divided into much more subtypes.

2 Related Research

Many researchers had been focused on the identification of named entity relations (Li et al., 2007; Liang et al., 2006). A general route is to convert the problem into classification, which means relations of entities are categorized as several types and a classifier is constructed. Different machine learning algorithms are applied to solve this question (Che et al., 2005; Miller et al., 2000; Kambhalta et al., 2004). As to Chinese entity relation, (Che et al., 2005) attempted to extract relations of entities automatically using Winnow and SVM, with F-score 73.08% and 73.27% respectively. (Liu et al., 2007) used HowNet to acquire semantic knowledge, and combine both semantic sequence kernel function and KNN classifier to extract relation, the accuracy achieve about 88%. (Dong et al., 2007) divided entity relation into categories: embedding relations and non-embedding relations, and construct different syntactic features respectively. Their experiment proved the method can improve performance of Chinese entity relation extraction task.

In this paper, we investigate to extract Chinese full names and abbreviations for given entities. Entities discussed in this paper are not limited to named entities, more precisely, are lexical nominal concepts, which are discriminative and exist independently; they can be either physical (e.g. Peking University) or abstract (e.g. scientific technique). In the following, we denote an entity with a Chinese full name as full name entity, and that with an abbreviated Chinese name as abbreviation entity. Different from other foreign languages, Chinese words and Chinese characters are different, we will refer to word as Chinese word and Chinese character as character itself.

In our paper, we don't study document-specific abbreviation, which means the abbreviation is just used in a specific document or context. E.g. In sentence . . . , the abbreviation (Wuhan subway) appears just as a concise anaphora in the document, but not a common fixed abbreviation.

3 Extraction of Full Name Entity and Abbreviation Entity from the Web

3.1 Comparison and Analysis of Full Name Entity and Abbreviation Entity

There are many common architectural features for a full name entity and its corresponding abbreviation entities. We sum up several points as follows:

¹ The nist ace evaluation website: <http://www.nist.gov/speech/tests/ace/ace07/>

1. Full name entity is consisted of several words, and its abbreviation entity is composed by one Chinese character from each word. E.g. (Women’s Federation) and (Fulian).
2. Full name entity is consisted of several morphemes, delete some of them and constitute its abbreviation entity. E.g. (Tsinghua University) and (Tsinghua).
3. Full name entity is consisted of coordinating morphemes; first Chinese character of each morpheme and common morphemes constitute its abbreviation entity. E.g. (agriculture and industry) and (Gongnongye).
4. Full name entity is consisted of coordinating morphemes; numeral and the common morphemes constitute its abbreviation entity. E.g. (modernization of agriculture, industry, national defense and science and technology) and (the four modernizations).
5. There are no special laws for full name entities and abbreviation entities of countries, provinces, and transliterations etc. E.g. (Shanxi province) and (Jin); (Kentucky Fried Chicken) and KFC

In the following, we introduce our method of extracting full name entity as example, as extracting abbreviation entity is analogous. Fig 1 illustrates our overall framework.

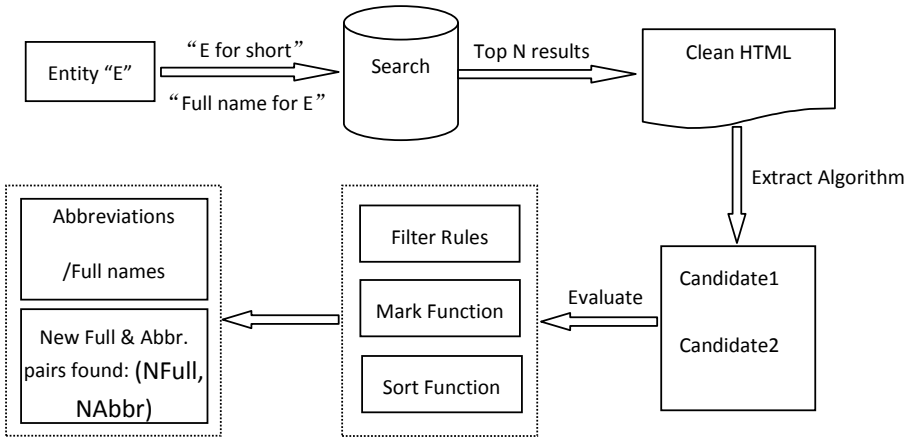


Fig. 1. Extract Abbreviations/Full names of Entity E from the Web

3.2 Obtaining Corpus

A big problem of extracting relation of entities is OOV (out of vocabulary), which is hard to identify. However, we found that full name entity and according abbreviation entity co-exist in a sentence in many cases: AB(B is short for A) and BA(A is full name of B). Inspired by this phenomenon, for a given entity, we can construct query items to search from the Web and obtain relevant corpus.

There are two kinds of query items: forward query item *Ent*, and backward query item **Ent*, in which *Ent* represents an entity. The query item contains double quotation marks to assure *Ent* and *"/* could appear in the query result consecutively. The queries are committed to Google search engine², we get only the summary of search results, not all the web pages which the result links to. In our experiment we collect top 200 search result summaries for each query item and delete the HTML tags.

3.3 Relevant Sentence Analysis

For the corpora obtained above, we use two patterns for the two query items respectively to extract sentences which contain the query items.

Pattern 1 *Ent(Common or Bracket or ...)(Colon or Bracket or is or ...)*. We call the sentence matching pattern 1 as a relevant sentence, denoted as *S*, it can be represented as *Pre+Ent+Pun++Cand+Pun*, in which *Pre* represent prefix words preceding entity *Ent*, *Pun* represent punctuation such as comma, and *Cand* represent preliminary candidate.

Pattern 2 *(Common or Bracket or ...)(Colon or Bracket or is or ...)... Ent*. Likely, relevant sentence *S* matching pattern 2 can be represented as *Cand+Pun++Pre+Pun+Ent*. Different strategies are designed according to *Pre*(show in table 1). We list some examples in table 2.

Table 1. Different actions according to *Pre*

	Relevant Sentence	Next Action	
		<i>Pre</i> = NULL	<i>Pre</i> ≠ NULL
First pattern	<i>Pre+Ent+Pun++Cand+Pun</i>	is a candidate.	(1) Extract Algorithm (FAEA)
Second pattern	<i>Cand+Pun++Pre+Pun+Ent</i>	is a candidate.	(2) New Full name and abbreviation may exist. Algorithm (FAEA)

3.4 Extracting Candidates

Pattern 1 and Pattern 2 are analogous, so we describe our method aiming at pattern 1 in the following. We segment³ relevant sentence *Pre+Ent +Pun ++ Cand+ Pun* as $P_1P_2 \dots P_m Ent Pun C_1C_2 \dots C_n Pun$. So our question converted into how to find two boundaries $i, j, 1 \leq i \leq m, 1 \leq j \leq n$, the full name entity candidate is $C_{j+1}C_{j+2} \dots C_n$, meanwhile, new possible full name & abbreviation entity pair found is $(P_{i+1}P_{i+2} \dots P_m, C_1C_2 \dots C_j)$.

² <http://www.google.cn>

³ We use Chinese segment tool ICTCLAS. <http://www.ictclas.org/>

Table 2. Some examples of relevant sentence

Entity	Query Item	Relevant Sentence	Pre=	Pre≠
			NULL	NULL
(Peking)		””...	√	
(at- tached primary school)		...		√
(four moderni- zations)	*	1963129, ””	√	
(primary school)	*			√

For a full name entity with segmentation $F = f_1f_2 \dots f_m$, $f_j(1 \leq j \leq m)$ is a word, and an abbreviation entity $A = a_1a_2 \dots a_n$, $a_i(1 \leq i \leq n)$ is a Chinese character. We define a similarity mark, which is not a strict measure, but used to decide the boundary of candidate from a sentence.

$$\text{SimMark}(A, F) = \sum_{i=1}^n \sum_{j=1}^m j * \text{issub}(a_i, f_j) \tag{1}$$

,in which $\text{issub}(a_i, f_j) = \begin{cases} 1 & \text{if } a_i \text{ appears in } f_j \\ 0 & \text{else} \end{cases}$

For example, we want to get the full names of entity (USTC), we extract a sentence (with segment and Pos-tagging) :

/n /n /j /w /d /v /a /a /n /w /v /w /u /n /w

(Central South University Patriotic Health Campaign Committee (full name of "Ai Wei Hui") issue notice about ...)

We found that $\text{SimMark}(,) \leq \text{SimMark}(,) \leq \text{SimMark}(,)$. Therefore, is more possible become a candidate than . The descriptions of algorithm are as follows.

3.5 Filtering and Sorting Candidates

After using algorithm **FAEA**, we may get several candidates; however, the Web is so open that some sentences may be irrelevant or even wrong. Therefore, it is necessary to validate and sort all the candidates.

We summarize some heuristic rules according to the commonalities between the abbreviation and full name of an entity, and they are used to filter out the candidates as follows.

Denote the set of candidates as $Candidates = \{Cand_1, Cand_2, \dots, Cand_n\}$, $Cand \in Candidates$, five filtering rules are defined as follows: (if one of the rules is satisfied, then $Cand$ is thought as an error, and thus be filtered out.)

Input: Sentence S (with segment and Pos-tagging): $P_1P_2 \dots P_m Ent Pun$
 $C_1C_2 \dots C_n Pun$ Entity: Ent ; Relation: "" (Full name)

Output: Full name Candidate of Ent : $EntFull$; New Full name-abbreviation pair ($NewFull, NewAbbr$);

- 1 Initialization: $i=m, j=n$;
- 2 If $P_1P_2 \dots P_i$ has at least a common character with $C_1C_2 \dots C_n$, then $i=i-1$; Else go to (4);
- 3 If P_i is an auxiliary or preposition word, then go to (4); Else $i=i-1$;
- 4 If P_i is a geographical entity name (discerning from Pos-tagging) or $SimMark(P_iP_{i+1} \dots P_m Ent, C_1C_2 \dots C_n) > SimMark(P_{i+1}P_{i+2} \dots P_m Ent, C_1C_2 \dots C_n)$, then $i=i-1$, go to (2);
- 5 If $C_jC_{j+1} \dots C_n$ has no common character with Ent , then $j=j-1$;
- 6 If $SimMark(Ent, C_jC_{j+1} \dots C_n) + SimMark(P_{i+1}P_{i+2} \dots P_m, C_1C_2 \dots C_{j-1}) > SimMark(Ent, C_{j+1}C_{j+2} \dots C_n) + SimMark(P_{i+1}P_{i+2} \dots P_m, C_1C_2 \dots C_j)$, then $j=j-1$; Else go to (7);
- 7 $EntFull=C_{j+1}C_{j+2} \dots C_n, NewFull=P_{i+1}P_{i+2} \dots P_m, NewAbbr=C_1C_2 \dots C_j$;
Return.

Algorithm 1. Full name/ Abbreviation Extract Algorithm (FAEA)

1. $Cand$ is a single Chinese character;
2. $len(Ent) \geq len(Cand)$, where $len(Ent)$ and $len(Cand)$ are the numbers of characters in $Cand$ and Ent respectively;
3. There are no common Chinese characters between $Cand$ and Ent ;
4. $segnum(Cand) > segnum(Ent)+3$, where $segnum(Cand)$ and $segnum(Ent)$ are the numbers of words of $Cand$ and Ent with segmentation respectively;
5. $Cand$ contains some meaningless interrogatives words;

The first four rules are straightforward, and we explain the fifth one a little. We may obtain some interrogative sentences which interrogate for the full name or abbreviation of an entity, but which do not end with any question mark, such as (What is the full name of Tsinghua). In this case, $Cand = (what)$, which is actually an error. The fifth rule can identify such errors, and filter them out.

Attributed to the fact that an entity could have more than one full name or abbreviation entities, it's necessary to sort all the candidates using the statistical information from $Sents$, the candidate with rank one is most common full name or abbreviation. We define a sort comparison function to sort all the candidates. Denote the set of relevant sentences as $Sents$, and define $C_i > C_j$, meaning C_i precedes C_j , iff

1. $SubstrFreq(C_i) \geq SubstrFreq(C_j)$
2. $LD(C_i, Ent) \leq LD(C_j, Ent)$, if $SubstrFreq(C_i) = SubstrFreq(C_j)$
3. $SubseqFreq(C_i) \geq SubseqFreq(C_j)$, if $SubstrFreq(C_i) = SubstrFreq(C_j)$ and $LD(C_i, Ent) = LD(C_j, Ent)$.

In which, $SubstrFreq(Cand)$ is the number of sentences which $Cand$ appear in the sentences of $Sent$ as a substring; $LD(Cand, Ent)$ is the Levenshtein distance of $Cand$ and Ent ; $SubseqFreq(Cand)$ is the number of sentences which $Cand$ appear in the sentences of $Sents$ as a subsequence.

4 Experiments and Discussion

We collect 300 pairs of full name entity and abbreviation entity as test data, in which about 80% are named entity, e.g. ((Peking University), (Beida)), (20). We compute abbreviation entities and full name entities respectively for each pair. The following tables illustrate our results. The precision of top k in table 3, table 4 and table 5 for full name is defined follows, the recall is either.

$$Precision_{top_k} = \frac{Count_{top_k}}{Count_{Full}} \tag{2}$$

In which, $Count_{top_k}$ represents the number of all correct full names extracted in top k, while $Count_{Full}$ represents the number of all full names.

Table 3. Performance with only *Pattern 1*

	Precision		Recall(of Top 1)	F-measure	
	Top 1	Top 3		Top 1	Top 3
Extract Abbr	87.1%	92.0%	71.7%	78.7%	80.6%
Extract Full	70.5%	79.3%	58.4%	63.9%	67.3%

Table 3 shows the performance when only pattern 1 is used. We can found that performance of extracting abbreviations is higher; partly because the pattern is more efficient for abbreviations, moreover, more boundary information (such as parenthesis and quotation marks) could be supplied when extracting abbreviations than full names. Performance of top 3 is higher than which of top 1; which illustrates effect of our ranking strategy; also prove that some entities have more than one abbreviation or full name. The recall of full names is only 58.4%, we found the reason is that pattern 1 cannot obtain sufficient corpus.

Table 4. Performance with *pattern 1* & *pattern 2*

	Precision		Recall(of Top 1)	F-measure	
	Top 1	Top 3		Top 1	Top 3
Extract Abbr	84.7%	93.6%	90.2%	87.4%	91.9%
Extract Full	77.0%	82.7%	85.3%	80.9%	84.0%

Table 4 shows that after pattern 2 is used, the performance is improved in both precision and recall, especially the recall of full names. The results also confirm us that the patterns accord with the most common expressions when full name and abbreviation entities co-appear.

Table 5 shows the performance of our method for named entities, we can find that our method is also efficient. Many organization entities end with a suffix and we called it "suffix abbreviation", because it usually follows many different

Table 5. Performance with *pattern 1* & *pattern 2* for named entity

	Precision		Recall(of Top 1)	F-measure	
	Top 1	Top 3		Top 1	Top 3
Extract Abbr	89.5%	95.3%	91.1%	90.3%	93.2%
Extract Full	82.3%	87.8%	85.9%	84.1%	86.8%

Table 6. Some sort comparison function values of abbreviations extracted

Entity	Abbreviation Candidates	Substr Freq	LD	Subseq Freq	Rank	Result correct?
(17	4	36	One	Yes
Chinese Academy of Social Sciences)		3	2	30	Two	Yes
		3	6	3	Three	No
		7	2	34	One	Yes
(Doctoral candidate)		6	3	36	Two	No

entities in sentences. E.g. (association),(university). We found that pattern 2 is very efficient for entities end with suffix abbreviation.

Table 6 shows some results with their sort compare function value; we can see that our three-level sort strategy is effective. For example, two abbreviation candidates of (Chinese Academy of Social Sciences) are (CASS) and (Siweiliangyuan); Although their SubstrFreq value is the same, we can sort them using Levenshtein Distance value correctly.

Table 7 illustrates the new pairs of full name entity and corresponding abbreviations found when we implement our algorithm for an entity (see the third column). In most cases, they are extracted when encountering with suffix abbreviation. Some pairs are partially correct and could be amended moreover. It also illustrates that our method can extend itself and get more pairs iteratively.

Furthermore, we summarize some difficulties of extracting full names or abbreviations as follows:

1. Abbreviation entities are usually OOV and difficult to segment. Therefore, the segments of abbreviation candidate supplied us are not reliable.
2. Obtain least but most useful corpora: we should have a tradeoff between more corpora and less Web search. We believe that the recall of extracting full name entities may be improved if we introduce more query items.

Table 7. New full names and abbreviations identified

Full name	Abbreviation	Trigger Entity	Result correct?
(Peking University)		(attached primary school)	Yes
(China)		(mobile)	Yes
(Beijing's Second)		(court)	No
(China Association of Trade in Service)		(customer service)	No

3. The prefix of a correct full name entity or abbreviation entity is difficult to identify when extracting candidates. In addition, the length and constituents of prefix words are hard to be determined when we figure out the left boundary of the candidate.

5 Conclusions

In this paper, we aim at extracting full names and abbreviations for a given entity from the Web. We propose a method of combining both patterns and rules to solve the problem. In addition, new pairs of full name entity and abbreviation entity can be extracted simultaneously. Our experiment shows our method is efficient. However, there is still more future research to improve this work. For example, how to construct more query items to obtain more corpora, how to use named entity identification method to validate candidates.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant No.60496326, 60573063, and 60773059; the National High Technology Research and Development Program of China under Grant No. 2007AA01Z325; the Education Commission Program of Beijing under Grant No. KM201010009004.

References

1. Che, W.X., Liu, T., Li, S.: Automatic Entity Relation Extraction. *Journal of Chinese Information Processing* 19(2), 1–6 (2005)
2. Dong, J., Sun, L., Feng, Y.Y., et al.: Chinese Automatic Entity Relation Extraction. *Journal of Chinese Information Processing* 21(4), 80–85, 91 (2007)
3. Kambhatla, N.: Combining lexicalsyntactic and semantic features with Maximum Entropy models for extracting relations. In: *Proceedings of 42th Annual Meeting of the Association for Computational Linguistics*, pp. 21–26 (2004)
4. Li, W.G., Liu, T., Li, S.: Automated Entity Relation Tuple Extraction Using Web Mining. *Acta Electronica Sinica* 35(11), 2111–2116 (2007)
5. Liang, H., Chen, J.X., Wu, P.B.: Information Extraction System Based on Event Frame. *Journal of Chinese Information Processing* 20(2), 40–46 (2006)
6. Liu, K.B., Li, F., Liu, L., et al.: Implementation of a Kernel-Based Chinese Relation Extraction System. *Journal of Computer Research and Development* 44(8), 1406–1411 (2007)
7. Luo, S.F., Sun, M.S.: Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures. In: *Proceedings of the Second SIGHAN Workshop, on Chinese Language Processing ACL*, pp. 24–30 (2003)
8. Maynard, D., Ananiadou, S.: Identifying Terms by Their Family and Friends. In: *Proceeding of COLING*, pp. 530–536 (2000)
9. Miller, S., Fox, H., Ramshaw, L., Weischedel, R.: A Novel Use of Statistical Parsing to Extract Information from Text. In: *Proceedings of 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 226–233 (2000)

10. Tian, G.G.: Research of Self-Supervised Knowledge Acquisition from Text based on Constrained Chinese Corpora (Doctor thesis). Institute of Computing Technology, Chinese Academy of Sciences (2007)
11. Wang, S., Cao, Y.N., Cao, X.Y., Cao, C.G.: Learning Concepts from Text Based on the Inner-Constructive Model. In: Zhang, Z., Siekmann, J.H. (eds.) KSEM 2007. LNCS (LNAI), vol. 4798, pp. 255–266. Springer, Heidelberg (2007)

An English-Arabic Bi-directional Machine Translation Tool in the Agriculture Domain

A Rule-Based Transfer Approach for Translating Expert Systems

Khaled Shaalan¹, Ashraf Hendam², and Ahmed Rafea³

¹ The British University in Dubai, Informatics
Dubai International Academic City,
Dubai, P.O. Box 345015, UAE

Honorary Fellow, School of Informatics, University of Edinburgh
khaled.shaalan@buid.ac.ae

² Central Lab. For Agricultural Expert Systems (CLAES), TEUES
6 El Nour St., Giza, 1123 Egypt,
a_hendam@mail.claes.sci.eg

³ American University in Cairo, SSE,
AUC Avenue, P.O. Box 74, New Cairo
1183 Egypt
rafeaa@aucegypt.edu

Abstract. The present work reports our attempt in developing an English-Arabic bi-directional Machine Translation (MT) tool in the agriculture domain. It aims to achieve automated translation of expert systems. In particular, we describe the translation of knowledge base, including, prompts, responses, explanation text, and advices. In the central laboratory for agricultural expert systems, this tool is found to be essential in developing bi-directional (English-Arabic) expert systems because both English and Arabic versions are needed for development, deployment, and usage purpose. The tool follows the rule-based transfer MT approach. A major design goal of this tool is that it can be used as a stand-alone tool and can be very well integrated with a general (English-Arabic) MT system for Arabic scientific text. The paper also discusses our experience with the developed MT system and reports on results of its application on real agricultural expert systems.

Keywords: Machine translation, transfer-based translation, rule-based analysis, rule-based generation, Arabic natural language processing, bilingual agricultural expert systems.

1 Introduction

Arabic is the fourth most-widely spoken language in the world. It is a highly inflectional language, with a rich morphology, relatively free word order, and two types of sentences (Ryding, 2005): nominal and verbal. Arabic natural language processing has been the focus of research for a long time in order to achieve an automated understanding of Arabic (Al-Sughaiyer et al., 2004). With globalisation and expanding

trade, demand for translation is set to grow. Computer technology has been applied in technical translation in order to improve speed and cost of translation (Trujillo, 1999). *Speed*: Translation by or with the aid of machines can be faster than manual translation. *Cost*: Computer aids to translation can reduce the cost per word of a translation. In addition, the use of machine translation (MT) can result in improvements in *quality*, particularly in the use of consistent terminology within a scientific text or for a specific domain.

With the recent technological advances in MT, Arabic has received attention in order to automate Arabic translations (Farghaly et al., 2009). In this paper, we follow a transfer-based MT approach. In the transfer approach (Trujillo, 1999), the translation process is decomposed into three steps: analysis, transfer, and generation. In the analysis step, the input sentence is analyzed syntactically (and in some cases semantically) to produce an abstract representation of the source sentence, usually an annotated parse tree. In the transfer step, this representation is transferred into a corresponding representation in the target language; a collection of tree-to-tree transformations is applied recursively to the analysis tree of the source language in order to construct a target-language analysis tree. In the generation step, the target-language output is produced. The (morphological and syntactic) generator is responsible for polishing and producing the surface structure of the target sentence. For each natural language processing component, i.e., analysis, transfer, and generation, we followed the rule-based approach. The advantage of the rule-based approach over the corpus-based approach is clear for (Abdel Monem et al., 2008; Shaalan, 2010): 1) less-resourced languages, for which large corpora, possibly parallel or bilingual, with representative structures and entities are neither available nor easily affordable, and 2) for morphologically rich languages, which even with the availability of corpora suffer from data sparseness.

English is a universal language that is widely used in the media, commerce, science and technology, and education. The size of the modern English content (e.g. literature and web content) is far larger than the amount of Arabic content available. Consequently, English-to-Arabic MT is particularly important. English-Arabic MT systems are mainly based on the transfer approach. For example, Ibrahim (1991) discussed the problem of the English-to-Arabic translation of embedded idioms and proverb expressions with the English sentences. Rafea et al. (1992) developed an English-to-Arabic MT system which translates sentences from the domain of political news from the Middle East. Pease et al. (1996) developed a system which translates medical texts from English-to-Arabic. El-Desouki et al. (1996) discussed the necessity of modular programming for English-to-Arabic MT. Translation of an English subset of a knowledge base to the corresponding Arabic phrases is described in (El-Saka et al., 1999). Mokhtar et al. (2000) developed an English-to-Arabic MT system, which is applied on abstracts from the field of Artificial Intelligence. Shaalan et al. (2004) developed an MT system for translating English noun phrases into Arabic that was applied to titles of theses and journals from the computer science domain. On the contrary, little work has been done in developing Arabic-to-English MT systems. Al-barhamtoshy (1995) proposes a translation method for compound verbs. Shaalan (2000) described a tool for translating the Arabic interrogative sentence into English. Chalabi (2001) presented an Arabic-to-English MT engine that allows any Arabic user to search and navigate through the Internet using the Arabic language. Othman et al. (2003) developed an efficient chart parser that will be used for translating Arabic sentence.

The proposed rule-based transfer MT tool described here is part of an ongoing research to automate the translation of expert systems between Arabic and English. This process translates the knowledge base, in particular, prompts, responses, explanation text, and advices. In CLAES¹, this tool is found to be essential in developing bilingual (English-Arabic) expert systems because both English and Arabic versions are needed for development, deployment, and usage purpose.

The next section outlines the overall architecture of the proposed English-Arabic bi-directional MT tool with illustrative examples of simple and complex transfers. In following section, we present the results of evaluation experiments. In a concluding section, we present some final remarks. Appendix I presents a classification of problems in the evaluation experiments.

2 The System Architecture

The structure of the bi-directional MT tool is shown in Figure 1. In this figure the arrows indicate the flow of information. The oval blocks indicate the basic modules of the system. Rectangular blocks represent the linguistic knowledge. This architecture describes the translation of a knowledge base in the agricultural domain, in particular, see Table 1: 1) prompts: noun phrases in the form of interrogative expressions, 2) responses: legal values in the form of noun phrases, 3) advices: in the form of imperative expressions and noun phrases, and 3) explanation text: in the form of verbal and nominal sentences.

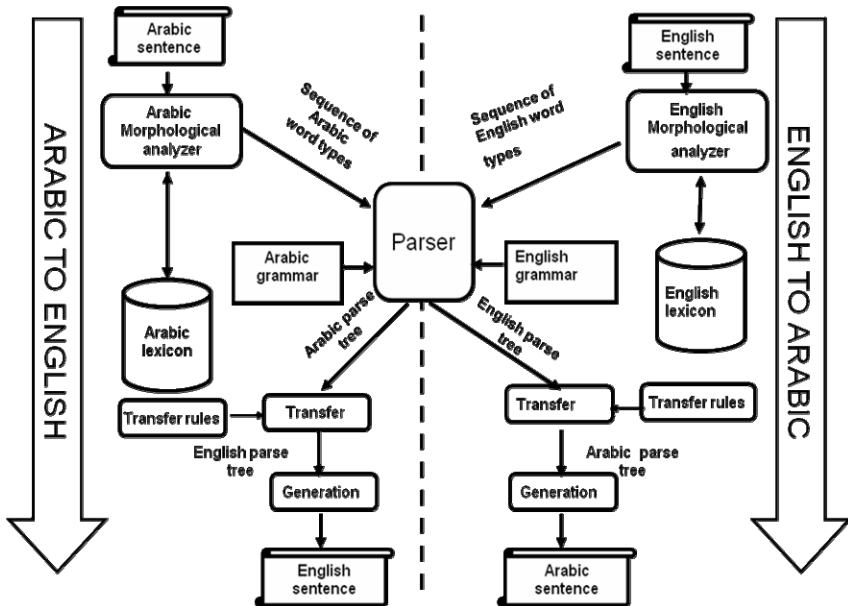


Fig. 1. Overall Structure of English-Arabic bi-directional sentence Translator

¹ Stands for Central Laboratory of Agricultural Expert Systems (CLAES), Agricultural Research Centre (ARC), Egypt, <http://www.claes.sci.eg>

Table 1. Examples of English-Arabic textual knowledge

	English	Arabic
Prompts	what is the abnormal leaves colour in the tunnel?	ما لون الأوراق الغير الطبيعي فى الصوبة؟
	what is the level of the nitrogen in the soil surface?	ما مستوى النتروجين فى سطح التربة؟
Responses (legal values)	bean mottle virus	فيروس تبقع الفول
	white growth with large black sclerotia	نمو أبيض مع أجسام حجرية سوداء
Advices (decisions)	Get rid of the remnants of the previous crop	تخلص من بقايا المحصول السابق
	spray when the number of nymphs is 3 on leaf	رش عندما يكون عدد الحوريات 3 على الورقة
Explanation	The unit for micro element for manganese during vegetative stage two	وحدة العناصر الصغرى من المنجنيز خلال مرحلة النمو الخضري الثانية
	the added fertilization elements are determined during the flowering stage by using the watery fertilization elements index	تحدد عناصر التسميد المضافة خلال مرحلة التزهير باستخدام ترتيب عناصر التسميد المائية

The proposed system is based on the transfer approach with three main components for each direction of translation: analysis, transfer, and generation. The analysis component consists of two steps morphological analysis and parsing. For accomplishing morphological analysis the lexicon is necessary, which is a repository of word stems. As Arabic is morphologically rich language, the morphological analysis of Arabic-to-English MT is an important step that is needed before we proceed with parsing the input sentence (Rafea et al., 1993). The transfer component has a collection of tree-to-tree transformations to the analysis tree of source sentence in order to construct a target analysis tree. The generation component generates the target language words according to the semantic features of the source language words. In our bi-directional English-Arabic translator, the actual translation occurs in the transfer phase. To explain how the sentence transfer process is performed by our translation system, we provide illustrative examples in Figure 2 through Figure 3 to show simple transfer of a noun phrase and compound transfer of a complete sentence, respectively. The former is an example showing that the syntactic transfer between English and Arabic noun phrase parse trees yields a representation in which word order is reversed. The later is a wider example showing the syntactic transfer between English sentence parse tree and Arabic verbal sentence parse tree yields a representation in which the Arabic VSO (verb-subject-object) order is transformed into the English SVO order.

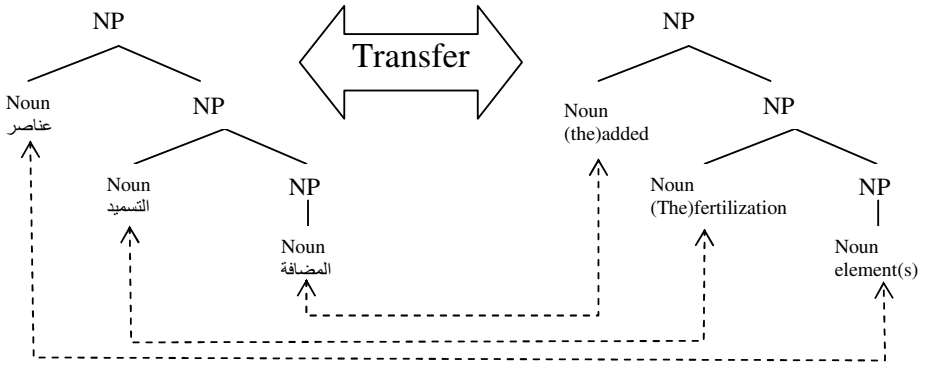


Fig. 2. Simple Transfer of Noun Phrase

3 Automatic Evaluation

To meet the demands of a rapid MT evaluation method, various automatic MT evaluation methods have been proposed in recent years. These include the BiLingual Evaluation Understudy (BLEU) (Papineni et al., 2002; Akiba et al., 2004). BLEU has attracted many MT researchers, who have used it to demonstrate the quality of their novel approaches to developing MT systems. BLEU is an automatic scoring method based on the precisions of N-grams. The precision of N-grams is calculated against reference translations produced by human translators. The results of BLEU is a score

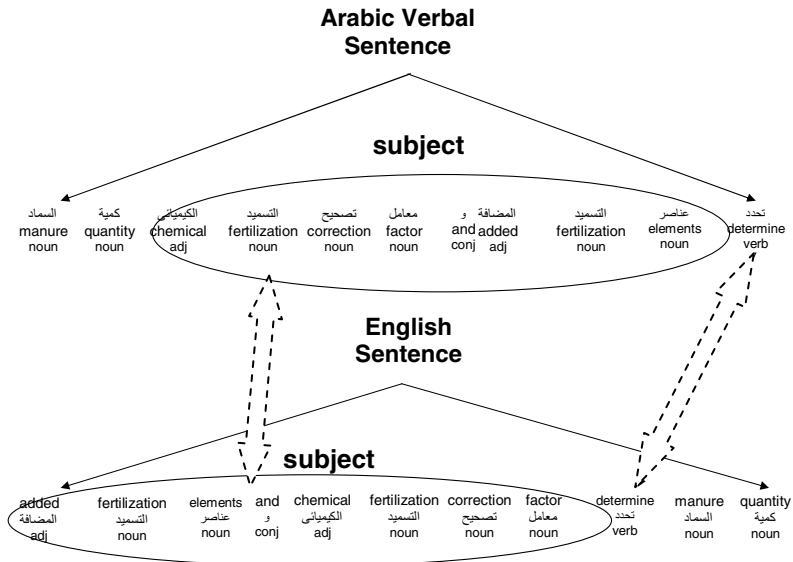


Fig. 3. Compound Transfer of verb and subject of a sentence

in the range of [0,1], with 1 indicating a perfect match. In order to evaluate the quality of our MT system by the Bleu tool we conducted two experiments in each direction of translation, i.e., from English to Arabic, and vice versa.

A set of real parallel 100 phrases and sentences from both English and Arabic versions of agricultural expert systems at CLAES, was used as a gold standard reference test data. This set consists of 23 advices, 46 prompts, and 31 explanation and responses. The evaluation methodology is performed as follows: 1) Run the system on the test data, 2) Automatically evaluate the system output against the reference translation and get results of the BLEU score, 3) Classify the problems that arise from mismatches between the two translations, 4) For problems that needs an alternative reference translation such as synonyms, prepare a second reference translation for the identified problems, and 5) Rerun the system on the same test data using both reference translations and present the results of improvements.

3.1 English to Arabic Evaluation Experiment

The automatic evaluation results of experiment I are shown in Table 2. There are 9 classifications of problems that arise from the divergences and mismatches between system output and reference translation which is shown in Table 6. As for problems 1, 4, 5, and 6, we made the changes on a second reference translation but for the remaining problems they are not solved at the moment as more research is needed to decide on their translations. Table 3 presents the automatic evaluation results of experiment IV which shows an improvement from 0.4504 to 0.6427.

Table 2. Results of automatic evaluation in Experiment I

	BLEU Score
Advices	0.5147
Prompts	0.4433
Explanation and responses	0.4703
Overall	0.4504

Table 3. Results of automatic evaluation in Experiment II

	BLEU Score
Advices	0.7673
Prompts	0.6549
Explanation and responses	0.6156
Overall	0.6427

3.2 Arabic to English Evaluation Experiment

The automatic evaluation results of experiment III are shown in Table 4. There are 4 classifications of problems that arise from the divergences and mismatches between system output and reference translation which is shown in Table 7. As for problems 1

and 4, we made the changes on a second reference translation but for problems 2 and 3 they are not solved at the moment as more research is needed to decide on their translations. Table 5 presents the automatic evaluation results of experiment IV which shows an improvement from 0.4581 to 0.8122.

Table 4. Results of automatic evaluation in Experiment III

	BLEU Score
Advices	0.4019
Prompts	0.4988
Explanation and responses	0.5616
Overall	0.4581

Table 5. Results of automatic evaluation in Experiment IV

	BLEU Score
Advices	0.8682
Prompts	0.7851
Explanation and responses	0.8169
Overall	0.8122

4 Conclusions

In this paper, we described the development of a novel English-Arabic bi-directional rule-based transfer MT tool in the agriculture domain. The translation between monolingual English and Arabic expert systems leads to rapid development and deployment of agricultural expert systems when one version is available. However, in the current version we may need to resort to minor post editing. Moreover, this tool would facilitate knowledge acquisition process to be either in English when international agricultural domain experts are available or in Arabic from local domain experts, which lead to bridging the gap of the language barrier.

A set of gold standard parallel English-Arabic phrases and sentences selected from agricultural expert systems developed at CLAES, is used to evaluate our approach, as well as the quality of the output of the MT tool. The problems found are classified, explained, and possible improvements, to some extent, are dealt with. The overall evaluation results, according to the presented evaluation methodology, were satisfactory. The automatic evaluation under one reference set achieved a BLEU score of 0.4504 for English-to-Arabic direction and 0.4581 for Arabic-to-English direction, whereas for two reference sets achieved 0.6427 for English-to-Arabic direction and 0.8122 for Arabic-to-English direction. However, more investigations are needed in order to make further improvements. One possible future direction is to use semantic processing. Another direction is to invest in building parallel corpora in the agriculture domain and employ the statistical machine translation approach.

References

- Akiba, Y., Federico, M., Kando, N., Nakaiwa, H., Paul, M., Tsujii, J.: Overview of the IWSLT 2004 evaluation campaign. In: Proceedings of the International Workshop on Spoken Language Translation, Kyoto, Japan, pp. 1–12 (2004)
- Abdel Monem, A., Shaalan, K., Rafea, A., Baraka, H.: Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework. *Machine Translation* 20(4), 205–258 (2008)
- Al-barhamtoshy, A.: Arabic to English Translator of Compound Verbs. In: Proceeding of the Annual Conference on Statistics, Computer Science, and Operations Research, Cairo University (December 1995)
- Al-Sughaiyer, I., Al-Kharashi, I.: Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology* 55(3), 189–213 (2004)
- Sakhr, C.A.: Web-based Arabic-English MT engine. In: proceeding of the ACL/EACL Arabic NLP Workshop (2001)
- El-Desouki, A., Abd Elgawwad, A., Saleh, M.: A Proposed Algorithm For English-Arabic Machine Translation System. In: Proceeding of the 1st KFUPM Workshop on Information and Computer Sciences (WICS): Machine Translation, Dhahran, Saudi Arabia (1996)
- El-Saka, T., Rafea, A., Rafea, M., Madkour, M.: English to Arabic Knowledge Base Translation Tool. In: Proceedings of the 7th International Conference on Artificial Intelligence Applications, Cairo (February 1999)
- Farghaly, A., Shaalan, K.: Arabic Natural Language Processing: Challenges and Solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, the Association for Computing Machinery (ACM) 8(4), 1–22 (2009)
- Ibrahim, M.: A Fast and Expert Machine Translation System involving Arabic Language, Ph. D. Thesis, Cranfield Institute of Technology, UK (1991)
- Mokhtar, H., Darwish, N., Rafea, A.: An automated system for English-Arabic translation of scientific texts (SEATS). In: International Conference on mMachine Translation and Multilingual Applications in the New Millennium, MT 2000, University of Exeter, British Computer Society, November 20-22 (2000)
- Othman, E., Shaalan, K., Rafea, A.: A Chart Parser for Analyzing Modern Standard Arabic Sentence. In: Proceedings of the MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches, New Orleans, Louisiana, USA (2003)
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA, pp. 311–318 (2002)
- Pease, C., Boushaba, A.: Towards an Automatic Translation of Medical Terminology and Texts into Arabic. In: Proceedings of the Translation in the Arab World, King Fahd Advanced School of Translation, November 27-30 (1996)
- Rafea, A., Sabry, M., El-Ansary, R., Samir, S.: Al-Mutargem: A Machine Translator for Middle East News. In: Proceedings of the 3rd International Conference and Exhibition on Multi-Lingual Computing (December 1992)
- Rafea, A., Shaalan, K.: Lexical Analysis of Inflected Arabic words using Exhaustive Search of an Augmented Transition Network. *Software Practice and Experience* 23(6), 567–588 (1993)
- Ryding, K.: Reference Grammar of Modern Standard Arabic. Cambridge University Press, Cambridge (2005)

Shaalán, K.: Machine Translation of Arabic Interrogative Sentence into English. In: Proceeding of the 8th International Conference on Artificial Intelligence Applications, pp. 473–483. Egyptian Computer Society (EGS), Egypt (2000)

Shaalán, K., Rafea, A., Abdel Monem, A., Baraka, H.: Machine translation of English noun phrases into Arabic. The International Journal of Computer Processing of Oriental Languages (IJCPOL) 17(2), 121–134 (2004)

Shaalán, K.: Rule-based Approach in Arabic Natural Language Processing. In: Special Issue on Advances in Arabic Language Processing, the International Journal on Information and Communication Technologies (IJICT). Serial Publications, New Delhi (June 2010) (submitted for publication)

Trujillo, A.: Translation Engines: Techniques for Machine Translation. Springer, London (1999)

Appendix I: Classification of Problems in Experiments I & III

Table 6. Classification of problems in Experiment I

1. Difference due to using a synonym of the target Arabic noun	the added fertilization elements are determined during the flowering stage by using the watery fertilization elements index	Source
	يحدد عناصر التسميد المضاف خلال مرحلة التزهير باستخدام فهرس عناصر التسميد المائي	Reference
	يحدد عناصر التسميد المضاف خلال مرحلة التزهير باستخدام ترتيب عناصر التسميد المائي	Output
2. Different translation of a preposition	The melted fertilization elements index in water for nitrogen during the second vegetative growth stage in kgm Fert/m ³	Source
	ترتيب عناصر التسميد المذابة في الماء من النيتروجين خلال مرحلة النمو الخضري الثانية في كجم تسميد/متر ³	Reference
	ترتيب عناصر التسميد المذابة في الماء من النيتروجين خلال مرحلة النمو الخضري الثانية بكجم تسميد/متر ³	Output
3. Misinterpret Arabic conjunction of words as English conjunction of phrases	The used fertilizers units total quantity determines the season length based on current and previous quantity of manure	Source
	يحدد إجمالي كمية وحدات الاسمدة المستخدمة طول العروة بناء على الكمية الحالية و الكمية السابقة للسماد	Reference
	يحدد إجمالي وحدات الاسمدة المستخدمة طول العروة بناء على الحالية و كمية سابقة للسماد	Output
4. An optional pronoun might come after the Arabic interrogative particle	What is the abnormal growth colour on the fruits?	Source
	ما هو لون النمو الغير طبيعي للثمار؟	Reference
	ما لون النمو الغير طبيعي للثمار؟	Output
5. Some words may have either sound plural feminine noun or broken (irregular) plural	What is the shape of the irregular fruits ?	Source
	ما شكل الثمار غير المنتظمة؟	Reference
	ما شكل الثمرات غير المنتظمة؟	Output
6. Non-standardization of the Arabic Written letters	“soil”, “second”, etc.	Source
	"الثاني" "تريبه"	Reference
	"الثاني" "تربة"	Output

Table 6. (continued)

7. Disagreement in present tense prefix of an Arabic verb	the added fertilization elements quantity and the chemical fertilizer correction factor determines the manure quantity and the unit during the flowering stage	Source
	تحدد كمية عناصر التسميد المضافة و معامل تصحيح السماد الكيميائي كمية و وحدة السماد خلال مرحلة التزهير	Reference
	يحدد كمية عناصر التسميد المضافة و معامل تصحيح السماد الكيميائي كمية و وحدة السماد خلال مرحلة التزهير	Output
8. Disagreement in gender between the adjective and the noun it modifies	The fertilization quantity from magnesium during the second vegetative growth stage in kg Fert/m ³	Source
	كمية التسميد من المغنسيوم خلال مرحلة النمو الخضري الثانية بكجم تسميد/متر ³	Reference
	كمية التسميد من المغنسيوم خلال مرحلة النمو الخضري الثاني بكجم تسميد/متر ³	Output
9. missing definite article in the Arabic noun	the drippers number and the dripper flow rate determine the irrigation motor time in minutes	Source
	يحدد عدد النقاطات و معدل تصرف النقاطات وقت موتور الري بالدقائق	Reference
	يحدد عدد النقاطات و معدل تصرف النقاطات وقت موتور الري بدقائق	Output

Table 7. Classification of problems in Experiment III

1. Difference due to synonyms of a target English noun	كمية السماد العضوي	Source
	the organic fertilizer quantity	Reference
	the organic manure quantity	Output
2. Selecting ambiguous category of a source Arabic word	رش المساحة المصابة فقط	Source
	spray the infected area only	Reference
	the infected area was sprayed only	Output
3. Misinterpret English conjunction of words as Arabic conjunction of phrases	حساب كمية المياه الكلية في كل مرحلة بإستخدام تاريخ البداية و النهاية	Source
	The total water quantity calculation for every stage by using the start and the end date	Reference
	the total water quantity calculation for every stage by using the start date and the end	Output
4. Variant translation without the preposition "of"	كمية السماد	Source
	The quantity of fertilizer	Reference
	The fertilizer quantity	Output

A Laplacian Eigenmaps Based Semantic Similarity Measure between Words

Yuming Wu^{1,2}, Cungen Cao¹, Shi Wang¹, and Dongsheng Wang^{1,2}

¹ Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, No.6 Kexueyuan South Road Zhongguancun, Beijing 100-190, China

² Graduate University of Chinese Academy of Sciences No. 19 Yu Quan Road, Shi Jing Shan Distinct, Beijing 100-049, China

Abstract. The measurement of semantic similarity between words is very important in many applicaitons. In this paper, we propose a method based on Laplacian eigenmaps to measure semantic similarity between words. First, we attach semantic features to each word. Second, a similarity matrix ,which semantic features are encoded into, is calculated in the original high-dimensional space. Finally, with the aid of Laplacian eigenmaps, we recalculate the similarities in the target low-dimensional space. The experiment on the Miller-Charles benchmark shows that the similarity measurement in the low-dimensional space achieves a correlation coefficient of 0.812, in contrast with the correlation coefficient of 0.683 calculated in the high-dimensional space, implying a significant improvement of 18.9%.

1 Introduction

Similarity measurement plays an important role in many areas, especially in semantic related applications [7]. So, the objective similarity measurement has to take more features of semantic level into consideration. For the purpose of attaching semantic features to words, we should have a knowledge source, from which we get semantic features and provide a flexible way to represent them, which can be extended to other knowledge sources without much modification.

In this paper, we propose a new method based on Laplacian eigenmaps [2] to define the semantic similarity between words. First, we use an online dictionary as a knowledge source, which is in semi-structured text format. Several types of interpretations can be extracted directly from the webpages. Then, these interpretations are transformed into a set of attribute-value pairs. These attribute-value pairs are used as semantic features, which are represented in a high dimensional space. After that, the Laplacian eigenmaps based method is adopted to find the intrinsic coordinates in low dimensional space. Finally the similarities are recalculated under the intrinsic coordinates in the low-dimensional space.

The remainder of the paper is organized as follows. In section 2 we describe the background. Section 3 makes further analysis to knowledge sources and features representation. In section 4 we describe the materials and methods. Section

5 gives the experimental result and compares against other methods on Miller-Charles benchmark dataset. Finally, we discuss the shortcomings of the proposed method and conclude this paper.

2 Background and Related Work

There has been a great deal of work on semantic similarity measurement. Some of them incorporate semantic features in the definition of similarity measure [6] [9] [3]. In Lins paper [6], an information-theoretic definition of similarity was proposed, the similarity of two objects, i.e. A and B only depends on the distributions of $common(A; B)$ and $description(A; B)$. This method is generic, but for some specific similarity, such as semantic similarity, Lins method may be unsuitable. As there is no an unified approach to encode the various semantic information in knowledge sources into the distribution, and when the data is sparse, the real distribution is hard to approximate. Resnik [9] presented a measure of semantic similarity in an IS-A taxonomy, based on the notion of information content. This measure depends on the structure taxonomy. In Chens [3] work, a context vector model to measure word similarity was adopted. The syntactic and semantic similarity is balanced by using related syntactic contexts only. Two major differences in these methods are the approach of using knowledge and representing semantic features. A more detail analysis on them is given in next section.

3 Knowledge Sources and Features Representation

Knowledge sources vary in their format, either structured or unstructured. WordNet [4] and HowNet [7] are structured knowledge sources, in which concepts are related in some manners. For example, the concepts in WordNet are related through several types of semantic relations, such as hypernym, antonym, holonym and meronym. The context around a word also provides strong and consistent clues to the sense of it [11]. Today, much text format materials are available on the Web. The context of a word can be relatively reliably extracted from those materials.

In this paper, we use an online dictionary¹ as knowledge source. In the online dictionary, there are plenty of items to interpret the sense of a given word. These items interpret the word from different aspects, such as interpretations in another language, sample sentences to describe the usage of the word, and further interpretations in same language and synonyms. After extractions, for each word, a collection of attribute-value pairs are obtained. For instance, the word "car" has an attribute-value pair $\langle hasSynonym, "automobile" \rangle$, which stands for that "car" and "automobile" is synonymous to each other. These attribute-value pairs are taken as the semantic features. The concrete examples are show in section 4.

¹ <http://www.dict.cn>

4 Materials and Methods

4.1 Interpretations in Online Dictionary

In a dictionary, a word or phrase is interpreted from multiple aspects in detail. These interpretations or descriptions are rich in semantic information. For the word "car", for example, we list some interpretations and descriptions, as depicted in Figure 1, to illustrate how the semantic features appear in an online dictionary.

Word	Car
Word Senses:	A wheeled vehicle adapted to the rails of railroad.
Sample Sentences:	The car in front of me stopped suddenly and I had to brake .
Synonyms:	auto, automobile, machine, motorcar, gondola.

Fig. 1. Sample interpretations for word "car"

As shown in figure 1, there are three subtitles the webpage, i.e. "word senses", "sample sentences" and synonyms. The word "car" is interpreted with several meanings, word senses, sample sentences, synonyms and so on. The words "wheel", and "vehicle", under the subtitle of "Word Senses", have close links to the word "car". We take the subtitle as attribute name, and take each word under a subtitle as an attribute value of the corresponding attribute. Note that the word itself can be also used as an attribute value. Sample attribute-value pairs for the word "car" are listed as Table 1. The context is a rich semantic source for a specific word, while there are also some stop words, such as "of", "but", and "in" in the context, which have little contribution to similarity between two words. Therefore, we leave out all these stop words in the process of measuring similarity.

Table 1. Sample attribute-value pairs for word "car"

Attribute Name	Attribute Value
Word Sense	motor
Synonyms	auto

4.2 Definition of Similarity Measure

Let W be the set of words, and f is a mapping from a word to an attribute-value pairs set.

$$f(w_i) = \{ \langle attribute_k, value_j \rangle \} \tag{1}$$

$$W(w_i, w_j) = \frac{|f(w_i) \cap f(w_j)|}{|f(w_i) \cup f(w_j)|} \tag{2}$$

where $|\cdot|$ refers cardinality of a set. As we take these semantic features independently, the dimension of space, in which these words are represented, is identical to the number of independent semantic features.

The number of attribute-value pairs is necessarily large. So the words are represented as points in the high dimensional space. Because the quantity of the words is limited, the words are very sparse in the high-dimensional space. Intuitively, there should exist a low dimensional embedding for the set of all the words.

A natural problem arises: how to find the optimal embedding in a low-dimensional space? Due to [2][5], the optimal embedding can be found by solve the following generalized eigenvalue problem.

$$Ly = \lambda Dy \tag{3}$$

where $L = D - W$ is called Laplacian matrix and $D(D_{ii} = \sum_j W_{ji})$ is the diagonal weight matrix, Give the dimension of the target space, let M be the matrix with column as the first m eigenvectors which satisfy the formula (3). Then, the optimal map should map x_i to be y_i , which is the i th row of matrix M .

Given the representations in a low-dimensional Euclidean space, the improved semantic similarity between y_i and y_j can be calculated as:

$$Sim_{improved}(y_i, y_j) = e^{-\|y_i - y_j\|^2} \tag{4}$$

where $\|\cdot\|$ refers to the 2-norm in Euclidean space. Now, we give the full process of how to measure the semantic similarity between two words. A more formal description of this process will be shown in the algorithm "semantic similarity based on Laplacian eigenmaps". The main ideas behind this algorithm are as follows. Firstly, encode the local relevance into a similarity matrix, in which each element is the basic similarity calculated by formula (2). Then we use Laplacian eigenmaps to find another representation in low dimensional space. Finally, we recalculate the semantic similarity in low dimensional space. Figure 2 show the algorithm of "Semantic Similarity based on Laplacian eigenmaps".

5 Experiments and Discussions

We take Miller-Charles dataset [8] as a benchmark dataset. The dataset has been used in several works [6][9]. The Miller-Charles dataset can be considered as a reliable benchmark for evaluating semantic similarity measures. As shown in Figure 3, the correlation increases rapidly when the dimension of the target space is from 1 to 10, and achieves the maximal value of 0.812 when the dimension is 11. After the dimension exceeds 11, the correlation coefficient decreases

Algorithm : *Semantic Similarity Based on Laplacian Eigenmaps*
Input:
 (1) a set of words $S = \{word_i\}$, (2) the dimension N of target space.
Output: the similarity matrix for all words in S .
Procedure:
 W are calculated by the formula [2];
 $D = diagonal(W)$; $L = D - W$;
 Calculate the first N eigenvectors v_1, \dots, v_N which satisfy the generalized eigenproblem $Lv=Dv$;
 Let U be the matrix with v_1, \dots, v_N as columns
 Let y_i be the i th row of matrix U ;
for each word-pair $\langle w_i, w_j \rangle \in S \times S$
 $Sim_Matrix(i, j) = Simimproved(y_i, y_j)$ as defined in formula (4) ;
end for each
Return Sim_Matrix ;

Fig. 2. Semantic Similarity Based on Laplacian Eigenmaps

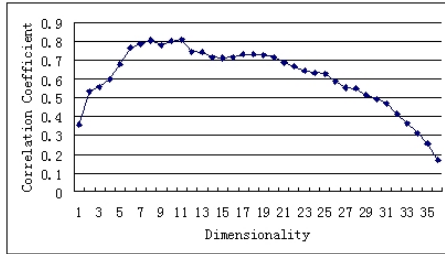


Fig. 3. Semantic Similarity Based on Laplacian Eigenmaps

steadily. These experimental results coincide with the intuition that there is a low dimensional representation for the semantic features in the high-dimensional space. Table 2 presents a comparison of the proposed method with several other methods, including CODC [3], SemSim [11], Lin [6] and Resink [9]. We get a correlation coefficient of 0.812 on Miller-Charlers dataset when the dimension of target space is 11. As shown in Table 2, the correlation coefficient of the proposed method is slightly lower than those obtained using the methods of SimSem and Lin, and higher than those obtained using the two other methods.

Table 2. Similarity measure comparison on Miller-Charles’ dataset($d = 11$)

	Miller-Charlers	CO DC	Sem	Sim Lin	Resnik	Proposed Method
Correlation Coefficient	1	0.693	0.834	0.823	0.775	0.812

6 Conclusion and Future Work

In this paper, we proposed a method based on Laplacian eigenmaps to measure the semantic similarity between words. The main contributions of our work are listed as follows.

First, our method takes semantic features into consideration in a natural way when measuring similarity between words. These semantic features are organized as attribute-value pairs. Our method is very flexible and easy to extend, because there is no dependence on the structure of semantic features.

Second, the problem of data sparseness was avoided, because the final similarities were calculated in low dimensional space.

Experimental results on the Miller-Charles dataset achieve a correlation coefficient of 0.812, showing that the proposed method outperforms the traditional corpus-based and thesauri-based measures. The future work will concentrate on the following two directions. One is to transform other knowledge sources into

Acknowledgements. This work is supported by the National Natural Science Foundation of China under Grant No. 60773059.

References

1. Bollegala, D., Matsuo, Y., Ishizuka, M.: Measuring semantic similarity between words using web search engines. In: Proc. of 16th WWW, pp. 757–766 (2007)
2. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* 15, 1373–1396 (2003)
3. Chen, K., You, J.: A study on word similarity using context vector models. *Computational Linguistics and Chinese Language Processing* 7, 37–58 (2002)
4. Fellbaum, C. (ed.): *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge (1998)
5. Chung, F.R.K.: *Spectral Graph Theory*. In: Conference Board of the Mathematical Sciences, AMS, Providence (1997)
6. Lin, D.: An information-theoretic definition of similarity. In: Proc. of 15th ICML, Madison, WI, pp. 296–304 (1998)
7. Liu, Q., Li, S.: Word Similarity Computing Based on How-net. In: *Computational Linguistics and Chinese Language Processing*, Taiwan, China, vol. (7), pp. 59–76 (2002)
8. Miller, G., Charles, W.: Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1), 1–28 (1998)
9. Resnik, P.: Using information content to evaluate semantic similarity. In: Proc. 14th IJCAI, Montreal, pp. 448–453 (1995)
10. Richardson, R., Smeaton, A., Murphy, J.: Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words, Working Paper CA-1294, Dublin City University (1994)
11. Yarowsky, D.: Unsupervised word sense disambiguation rivalling supervised method. In: Proc. of the 33rd ACL, June 26–30, pp. 189–196 (1995)

A Filter-Based Evolutionary Approach for Selecting Features in High-Dimensional Micro-array Data

Laura Maria Cannas, Nicoletta Dessì, and Barbara Pes

Università degli Studi di Cagliari, Dipartimento di Matematica e Informatica,
Via Ospedale 72, 09124 Cagliari, Italy
{lauramcannas, dessi, pes}@unica.it

Abstract. Evolutionary algorithms have received much attention in extracting knowledge on high-dimensional micro-array data, being crucial to their success a suitable definition of the search space of the potential solutions. In this paper, we present an evolutionary approach for selecting informative genes (features) to predict and diagnose cancer. We propose a procedure that combines results of filter methods, which are commonly used in the field of data mining, to reduce the search space where a genetic algorithm looks for solutions (i.e. gene subsets) with better classification performance, being the quality (fitness) of each solution evaluated by a classification method. The methodology is quite general because any classification algorithm could be incorporated as well a variety of filter methods. Extensive experiments on a public micro-array dataset are presented using four popular filter methods and SVM.

Keywords: Evolutionary algorithms, Feature Selection, Micro-array Data Analysis.

1 Introduction

Evolutionary strategies are now an active area of research and a lot of studies demonstrate the advantages of their use in several knowledge extraction tasks. In particular, recent literature [1][2][3][4] demonstrates their success on micro-array data analysis. The micro-arrays provide a view onto cellular organization of life through quantitative data on gene expression levels and it is expected that knowledge gleaned from micro-array data will contribute significantly to advances in fundamental questions in biology as well as in clinical medicine. In particular, these data may be used to extract knowledge on the molecular variation among cancer i.e. to build a model, namely a classifier, capable of discriminating between different clinical outcomes in order to make accurate prediction and diagnosis of cancer. Building such a classifier is somewhat problematic since in micro-array datasets the number of samples collected is small compared to the number of genes per sample which are usually in the thousands. Since it is highly unlikely that thousands of genes have the information related to the cancer and using all the genes results in too big dimensionality, it is necessary to select some genes highly related to particular classes for classification, which are called informative genes. This process is referred to as gene selection. It is also called feature selection in machine learning.

In this paper, we attempt to move away from strictly statistical and data mining methods that seem to dominate the current state of art in this area, and try to explore how knowledge extraction from gene expressions can be successfully carried out by an evolutionary strategy. Our approach to micro-array data classification can be viewed as a two-stage procedure.

First, we try to break the barrier of feature selection. We adopt filters, which are commonly used in the field of data mining and pattern recognition, for ranking features in terms of the mutual information between the features and the class label. Then, we combine ranking results in small subsets of predictive genes substantially reducing the number of features. These subsets are input to the second stage that adopts an evolutionary approach to further select features and precisely classify cancer. Specifically, feature selection is formulated as an optimization problem for which it is to find the genes that guarantee maximum accuracy in a given classification task. A Genetic Algorithm (GA) is used to explore the feature space defined in the first stage and look for solutions (i.e. gene subsets) with better classification performance. The quality (fitness) of each solution is evaluated by an SVM classifier (but any classification algorithm could be incorporated in our approach). As a test-bed for evaluating the proposed methodology we choose the Leukemia dataset, publicly available at [5]. We demonstrate, with results, that our approach is highly effective in selecting small subsets of predictive genes while it allows saving time and alleviating computational load.

Although a number of recent works address the problem of gene selection using a GA in conjunction with some classifier [2][6][7], our approach is innovative: instead of exploring the whole dataset, the GA looks for solutions in the small gene spaces that we defined in the first stage. This way, we can analyze the gene information very rapidly.

The paper is organized as follows. In Section 2, we discuss some related works. Section 3 describes the proposed approach, while experiments are reported in Section 4. In Section 5, we discuss the results and present some concluding remarks.

2 Related Work

Recent studies [1][2][3][4] address the problem of gene selection using a standard GA which evolves populations of possible solutions, the quality of each solution being evaluated by an SVM classifier. GAs have been employed in conjunction with different classifiers, such as k-Nearest Neighbor in [6] and Neural Networks in [7]. Moreover, evolutionary approaches enable the selection problem to be treated as a multi-objective optimization problem, minimizing simultaneously the number of genes and the number of misclassified examples [3][4][8].

Recent literature [4] shows that evolutionary approaches may benefit of a preliminary feature selection step when applied to high dimensional problems such as micro-array data analysis. A number of hybrid approaches have been proposed [2][3][4] that apply some pre-filtering technique to define suitable gene spaces to be further refined by an evolutionary algorithm. Yang et al [9] and Forman [10] conducted comparative studies on filter methods, and they found that Information Gain and Chi-square are among the most effective methods of feature selection for classification.

3 The Proposed Approach

We define extracting knowledge from micro-array data the process that selects discriminative genes related to classification, trains a classifier and then classifies new data using the learned classifier. As previously mentioned, our knowledge extraction process has two stages that we describe in the following.

First Stage: the Search Space Definition. It is common to use some techniques to generate a small list of important features in order to learn classifiers that use only a small subset of the original dataset. A popular method, which is named filter, is to define the feature selection as a preprocessing step that is independent from classification. In more detail, a filter method computes a score (ranking) for each feature and then selects features according to the scores. However, each filter method is able to point out only a peculiar character of the information contained in the data at hand, resulting in a feature list that may be not nearly informative. For overcoming this problem, we propose constructing M lists of features, that we call *Feature Pools* (FPs), via the combination of M different filter methods. The final objective is to have different lists (i.e. FPs) of candidate genes, to be further refined by a genetic algorithm. Inspired by our previous work [4], the construction of FPs is carried out according the following steps:

1. M filter processes are carried out separately on the original dataset. This results in M lists of ranked features each containing all the features in descending order of relevance.
2. According to a fixed threshold T , we cut the previous lists and consider only the T top-ranked features from each list.
3. To absorb useful knowledge from the above lists, we fuse their information by considering the features they share. Specifically, we build M nested feature pools $FP_1 \cdot FP_2 \dots \cdot FP_M$, where FP_1 contains the features shared by all the M lists, FP_2 the features shared by at least $M-1$ of the M lists, ..., FP_{M-1} the features shared by at least 2 of the M lists. Finally, FP_M contains all the features belonging to the M lists.

Second Stage: GA-based Gene Selection and Classification. The second stage considers two aspects: how the mechanism of feature selection works and how the classifier accuracy is affected by the mechanism. The evolutionary approach we propose here is intended for two distinct purposes:

1. Effective use of a GA that provides rapid local-search capabilities in the search space defined at the first stage.
2. Effective use of SVM that provides high-quality classifiers.

The key idea is to obtain the benefits from both GA and SVM: the former is used to explore the input search space and to discover promising subsets of features (i.e. genes) while the latter evaluates them by classification.

With the GA, individuals are small sets of important features, typically represented by a string or a binary array. A population of individuals is randomly initialized at the start of the GA. This population undergoes mutation (a bit in an instance is flipped) and crossover (two instances create two new instances by splitting both parent bit-strings) operators, creating a collection of new individuals. This evolution process is

repeated until a pre-defined number of generations G is reached, resulting in a “best” individual that represents the most informative feature subset.

Our evolutionary strategy considers to separately apply this process on each FP. Accordingly, each individual is a binary vector (whose maximum size is $M \cdot T$), where the values “1” and “0” respectively mean that the feature is included or not in the individual. Genetic operations are carried out by roulette wheel selection, single point crossover, and bit-flip mutation. Taking into consideration previous research [4], we assume as fitness function the accuracy of the SVM classifier learnt on the individual. With regard to SVM classifier, error estimation is made by using leave-one-out cross validation (LOOCV). This choice is justified by the will to pay great attention to the classifier accuracy, even if the required computational load is greater than using other evaluation methods.

4 Experimental Analysis

We report on the successful application of the proposed approach to Leukemia dataset [5], which contains 7129 gene expression levels from 72 samples, among which 25 samples are collected from acute myeloid leukemia (AML) patients and 47 samples are from acute lymphoblastic leukemia (ALL) patients. The overall analysis has been implemented using the Weka library [11].

First Stage. We set the number of filter methods $M = 4$, the threshold $T = 20$ and choose the following ranking methods: Information Gain, Chi-squared, Symmetrical Uncert, and One Rule. The feature selection process (section 3) results in the following feature pools: FP_1 (composed of 12 features), FP_2 (18 features), FP_3 (21 features), and finally FP_4 (29 features).

Second Stage. Each FP_i ($i = 1, 2, \dots, 4$) is used as input to the GA. In order to find an efficient setting of the algorithm in the considered domain, we operated a performance analysis by considering different values of the following parameters: (i) number of generations, (ii) population size, (iii) probability of crossover, and (iv) probability of mutation. Specifically, the analysis was carried on according to two distinct phases:

- A. We test the GA/SVM behavior as parameters (i) and (ii) change, while parameters (iii) and (iv) assume values consistent with the literature;
- B. We test the GA/SVM behavior as parameters (iii) and (iv) change, while parameters (i) and (ii) assume the best results found in the previous phase A.

This pairing is justified because, in the literature, wide discordances can be found between the values chosen for parameters (i) and (ii). As well, parameters (iii) and (iv) typically assume values in a range that we consider in our analysis. Since the evolutionary algorithm performs a stochastic search, in both phases we consider the average results over a number $P = 10$ of trials.

Phase A. We tested the performance of GA/SVM as the parameters (i) number of generations and (ii) population size change, by considering each combination of the values of these two parameters. Specifically, values considered for parameters are as follows: (i) number of generations: 10, 20, 30, 50, and 100; (ii) population size: 10,

20, 30, and 50; (iii) probability of crossover = 1; (iv) probability of mutation = 0.01. Tables (1-4) show results on each FP_i , in terms of classification accuracy and feature subset size (in brackets). Derived from Tables (1-4), Figures (1-4) show the interpolation surface expressing the global trend of the average accuracy and average subset size (y-axis) vs. the number of generation (x-axis) and the population size (z-axis). Different colours indicate different ranges of values (shown in the enclosed legends) in order to better evaluate changes respectively occurring in the average accuracy and in the average subset size. With regard to computational load, we don't show the relative results explicitly, but we consider them in the subsequent discussion.

Table 1. Performance of GA/SVM on feature pool FP_1

		FP1			
Avg. accuracy (avg size)		Population size			
		10	20	30	50
Generations	10	0.9639 (6)	0.9778 (5.4)	0.9806 (4.8)	0.9861 (4.6)
	20	0.9694 (6.2)	0.9778 (5.6)	0.9819 (4.5)	0.9861 (4)
	30	0.9694 (6.2)	0.9778 (5.4)	0.9819 (4.3)	0.9861 (4)
	50	0.9681 (5.1)	0.9806 (4.7)	0.9819 (4.3)	0.9861 (4)
	100	0.9694 (6.2)	0.9778 (5.4)	0.9833 (4)	0.9861 (4)

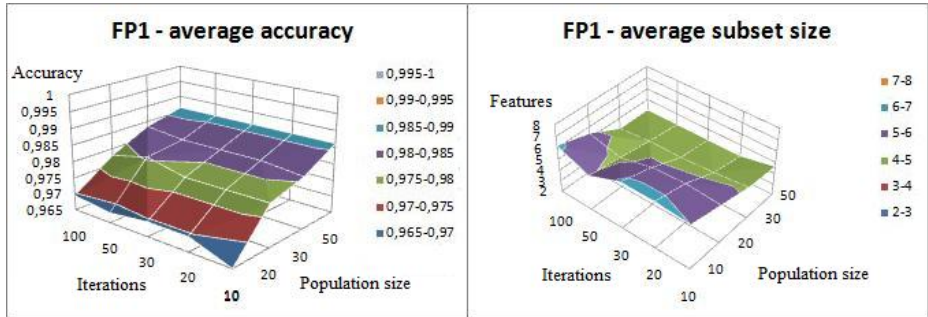


Fig. 1. Performance of GA/SVM on feature pool FP_1

As regards the number of generations, we notice significant results for 30-50 iterations. Going on up to 100 iterations causes some improvement only in one run out of 10, while computational load increases in accordance to the number of generations. Considering the population size, we observe that best results are obtained when the value assumed by this parameter is 30 or 50. Values less than 30 make the algorithm to converge to a local optimum, while values greater than 50 were not considered for two reasons: the average accuracy and average size of the subset seem to stabilize when the value assumed by this parameter is 30 as well, exceeding 30, computational load increases considerably.

Table 2. Performance of GA/SVM on feature pool FP₂

		FP2			
Avg. accuracy (avg size)		Population size			
		10	20	30	50
Generations	10	0.9778 (8)	0.9889 (7.6)	0.9875 (6.3)	0.9861 (5.8)
	20	0.9778 (8)	0.9889 (6.6)	0.9889 (5.5)	0.9889 (5.6)
	30	0.9778 (8)	0.9889 (6.2)	0.9903 (5.1)	0.9889 (5.6)
	50	0.9778 (8)	0.9875 (5.3)	0.9917 (4.7)	0.9903 (5)
	100	0.9778 (8)	0.9917 (5)	0.9917 (4.6)	0.9917 (4.8)

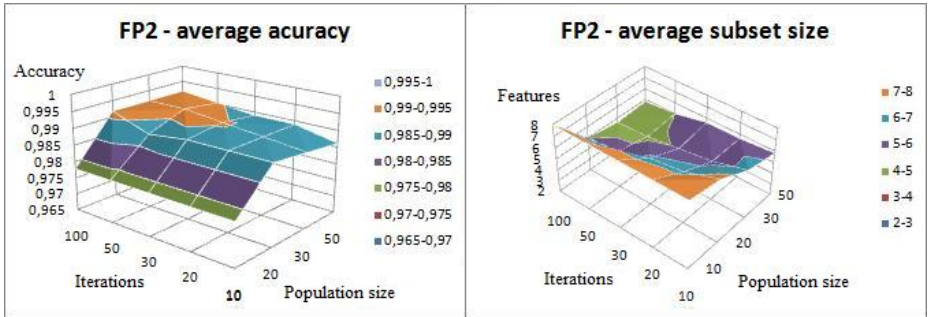


Fig. 2. Performance of GA/SVM on feature pool FP₂

Table 3. Performance of GA/SVM on feature pool FP₃

		FP3			
Avg. accuracy (avg size)		Population size			
		10	20	30	50
Generations	10	0.9833 (9.6)	0.9889 (7.8)	0.9986 (6.3)	0.9972 (5.8)
	20	0.9833 (9.2)	0.9889 (7)	0.9986 (5.5)	1 (5.8)
	30	0.9833 (9.2)	0.9889 (6.8)	0.9986 (5)	1 (5.4)
	50	0.9806 (8.5)	0.9875 (5.8)	0.9986 (4.3)	0.9972 (4.3)
	100	0.9833 (9.2)	0.9889 (6.6)	1 (4.3)	1 (3.6)

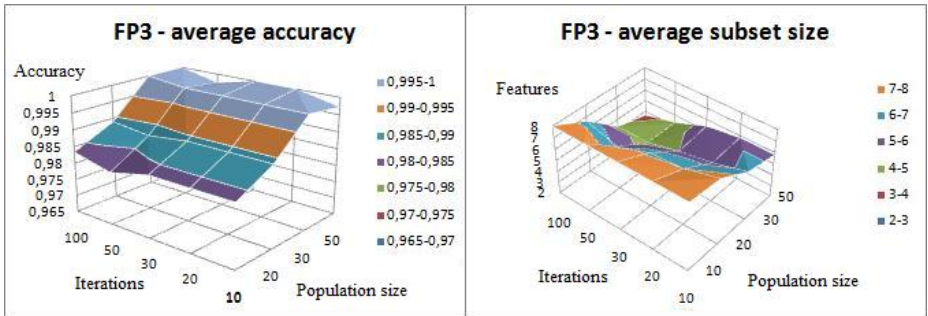


Fig. 3. Performance of GA/SVM on feature pool FP₃

Table 4. Performance of GA/SVM on feature pool FP₄

FP4					
Avg. accuracy (avg size)		Population size			
		10	20	30	50
Generations	10	0.9861 (9.4)	0.9861 (7.4)	0.9889 (9.2)	1 (9.2)
	20	0.9889 (9.2)	0.9861 (7.2)	0.9889 (9.2)	1 (7.6)
	30	0.9889 (9.2)	0.9861 (7)	0.9889 (9.2)	1 (6.8)
	50	0.9875 (10.1)	0.9889 (9.2)	0.9903 (5.9)	0.9958 (6.1)
	100	0.9889 (9.2)	0.9889 (6.6)	0.9903 (5.7)	1 (4.8)

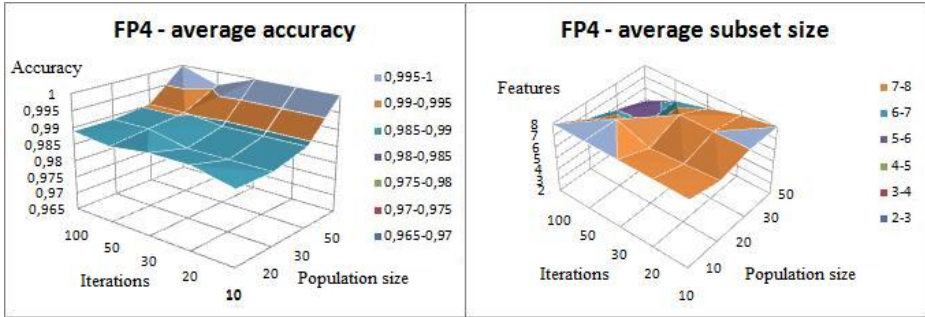


Fig. 4. Performance of GA/SVM on feature pool FP₄

Phase B. We tested the performance of GA/SVM as the parameters (iii) probability of crossover and (iv) probability of mutation change, by considering each combination of the values of these two parameters. Values considered for parameters (iii) and (iv) are respectively: (iii) 0.6, 0.8, 1 and (iv) 0.005, 0.01, 0.02, 0.03. According to the results obtained in the phase A, we set (i) number of generations = 50 and (ii) population size = 30.

Tables (5-8) show results on each FP, in terms of classification accuracy and feature subset size (in brackets). Again, figures (5-8) show the same results using charts (average accuracy and average subset size on the y-axis, probability of mutation on the x-axis, and probability of crossover on the z-axis).

Considering the parameter probability of crossover, we did not achieve significant variations as values change; however we find the best results in correspondence to value 1. Finally, as regards the parameter probability of mutation, we notice that increasing values correspond to better results on average. In particular, the value 0.02 gives good results considering both accuracy and dimensionality and, in addition, exceeding 0.02 computational load increases considerably.

Table 5. Performance of GA/SVM on feature pool FP₁

FP1				
Avg. accuracy (avg size)		Probability of crossover		
		0.6	0.8	1
Prob. of mutation	0.005	0.9778 (4.2)	0.9833 (5.4)	0.9833 (4.2)
	0.01	0.9806 (4)	0.9861 (4.6)	0.9819 (4.3)
	0.02	0.9861 (4)	0.9861 (4)	0.9861 (4)
	0.03	0.9861 (4)	0.9861 (4)	0.9861 (4)

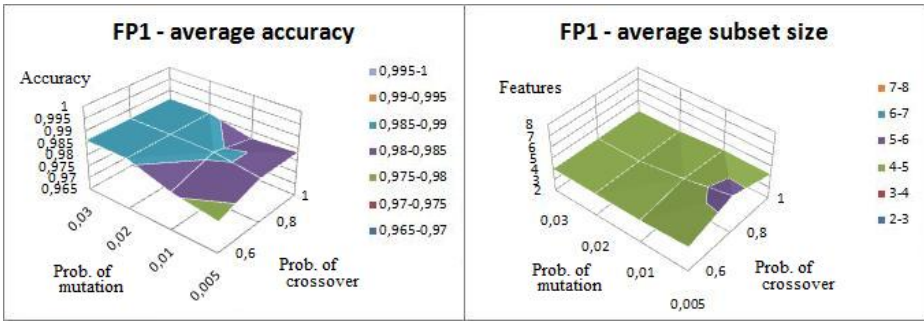


Fig. 5. Performance of GA/SVM on feature pool FP₁

Table 6. Performance of GA/SVM on feature pool FP₂

FP2				
Avg. accuracy (avg size)		Probability of crossover		
		0.6	0.8	1
Prob. Of mutation	0.005	0.9861 (6.4)	0.9889 (7.8)	0.9889 (5)
	0.01	0.9917 (6.4)	0.9889 (5.8)	0.9917 (4.7)
	0.02	0.9917 (5)	0.9944 (5)	0.9972 (4.6)
	0.03	0.9944 (4.2)	0.9972 (5.8)	0.9944 (4.4)

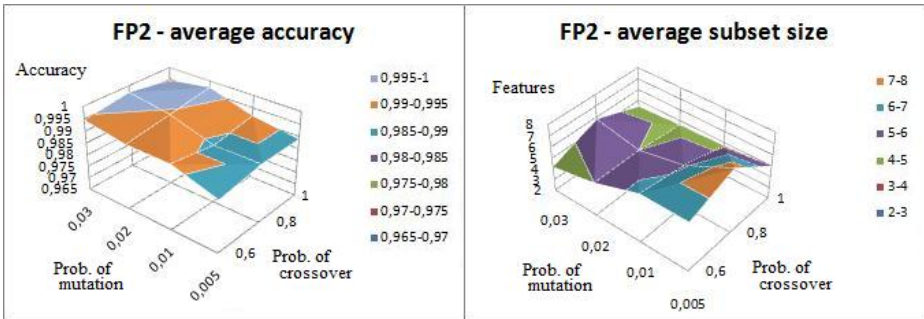


Fig. 6. Performance of GA/SVM on feature pool FP₂

Table 7. Performance of GA/SVM on feature pool FP₃

FP3				
Avg. accuracy (avg size)		Probability of crossover		
		0.6	0.8	1
Prob. of mutation	0.005	0.9944 (5.8)	0.9917 (5.4)	0.9917 (5.2)
	0.01	0.9972 (4.4)	0.9972 (4.6)	0.9986 (4.3)
	0.02	0.9972 (4.4)	0.9972 (5.4)	0.9944 (5.4)
	0.03	0.9972 (5.2)	0.9944 (4.4)	0.9972 (4.2)

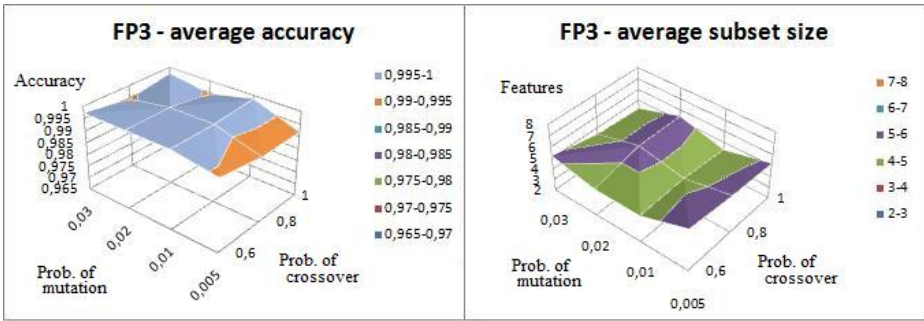


Fig. 7. Performance of GA/SVM on feature pool FP₃

Table 8. Performance of GA/SVM on feature pool FP₄

FP4				
Avg. accuracy (avg size)		Probability of crossover		
		0.6	0.8	1
Prob. of mutation	0.005	0.9917 (8.2)	0.9944 (6.8)	0.9944 (7.6)
	0.01	0.9944 (6.6)	0.9972 (6.2)	0.9903 (5.9)
	0.02	0.9972 (6.4)	0.9944 (5.8)	0.9972 (5.6)
	0.03	0.9972 (6.2)	1 (5.6)	0.9972 (6.2)

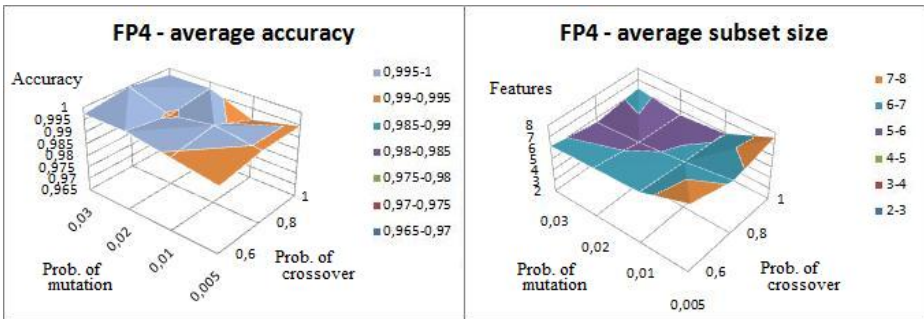


Fig. 8. Performance of GA/SVM on feature pool FP₄

5 Discussion and Concluding Remarks

First, it is important to notice that the parameter values we consider as optimal, especially regarding the number of generations and the population size, are smaller than the values commonly used in other methods discussed in the literature, with consequent time saving and computational load saving. Because our experiments result in excellent fitness, we can assert that the evolutionary approach we propose allows us to use a GA in a both effective and efficient manner: small subsets of predictive genes are selected with a reduced computational load. This validates the process of building FPs that reduce the dimensionality of the initial problem by discarding redundant or irrelevant features.

With regard to FPs construction, a basic question is how defining the most effective search space for the GA. Combining valuable results from different ranking methods allows us to achieve good results. However, features common to all ranking methods (i.e. the features belonging to FP_1) define a search space that is too small: the performance of GA/SVM achieves 98,6% of accuracy and does not increase when the search is refined by an additional number of generations. When this search space is enlarged by adding genes selected by three, two and just one method, our approach shows an excellent performance, not only at providing a very good average accuracy, but also with respect to the number of selected features and the computational cost. In particular, the pool FP_3 seems to define the most effective search space for the GA.

A further question we want to point out is that, as presented in Table 1-8, we consider the average results obtained in the analysis. But, during our study, we noticed that the difference between average values and best values was very scanty, and it means that results are not outcomes of a particularly lucky run, but they derive from a valid and effective behavior of the evolutionary method.

Table 9 summarizes the results we obtained using the proposed approach with the results of three state-of-art methods that use a GA as feature selection technique. To evaluate the results we use the conventional criteria, that is the classification accuracy in terms of the rate of correct classification (first number) and the size of the subset i.e. the number of selected genes (the number in parenthesis). For our approach, we choose to present the data obtained using FP_3 . The maximum classification rate we obtain is 1 using 3 genes while the corresponding average classification rate is 1 and the corresponding average dimension is 3.6 (see Table 3 for details). The same performance is achieved by [1] [2] [8], even if the number of genes selected by [1] [2] [8] is greater than the one obtained by our method.

As feature work, we plan to extend our study by considering different ranking methods, as well as different values of the threshold used to cut-off each ranked list, in order to gain more insight on the evolutionary search space definition. Moreover, the proposed approach will be validated on different micro-array datasets.

Table 9. The proposed method versus three state of the art methods

Studies	Classification rate	Subset size
The proposed method	1	(3)
[1]	1	(6)
[8]	1	(4)
[2]	1	(25)

References

1. Peng, S., et al.: Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letter* 555(2), 358–362 (2003)
2. Huerta, E.B., Duval, B., Hao, J.K.: A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data. In: Rothlauf, F., Branke, J., Cagnoni, S., Costa, E., Cotta, C., Drechsler, R., Lutton, E., Machado, P., Moore, J.H., Romero, J., Smith, G.D., Squillero, G., Takagi, H. (eds.) *EvoWorkshops 2006*. LNCS, vol. 3907, pp. 34–44. Springer, Heidelberg (2006)

3. Tan, F., Fu, X., Zhang, T., Bourgeois, A.G.: Improving Feature Subset Selection Using a Genetic Algorithm for Microarray Gene Expression Data. In: IEEE Congress on Evolutionary Computation, Vancouver, BC, Canada, July 16-21 (2006)
4. Dessi, N., Pes, B.: An Evolutionary Method for Combining Different Feature Selection Criteria in Microarray Data Classification. *Journal of Artificial Evolution and Applications* 2009, Article ID 803973, 10 pages, doi:10.1155/2009/803973
5. <http://datam.i2r.a-star.edu.sg/datasets/krbd/>
6. Li, L., Weinberg, C.R., Darden, T.A., Pedersen, L.G.: Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* 17(12), 1131–1142 (2001)
7. Bevilacqua, V., et al.: Genetic Algorithms and Artificial Neural Networks in Microarray Data Analysis: a Distributed Approach. *Engineering Letters* 13(3), EL_13_3_14 (2006)
8. Reddy, A.R., Deb, K.: Classification of two-class cancer data reliably using evolutionary algorithms. Technical Report. KanGAL (2003)
9. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML 1997 (1997)
10. Forman, G.: An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research* (2003)
11. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Elsevier, Amsterdam (2005)

A Novel Distribution of Local Invariant Features for Classification of Scene and Object Categories

LiJun Guo¹, JieYu Zhao², and Rong Zhang²

¹ Institute of Computer Technology, CAS
Graduate University of Chinese Academy of Sciences
Faculty of Information Science & Engineering NingBo University
818, Fenghua, Ningbo City,
Zhejiang, China
guolijun@nbu.edu.cn

² NingBo University
Faculty of Information Science & Engineering
818, Fenghua, Ningbo City,
Zhejiang, China
Zhao_jieyu@nbu.edu.cn,
zhangrong@nbu.edu.cn

Abstract. A new image representation based on distribution of local invariant features to be used in a discriminative approach to image categorization is presented. The representation which is called Probability Signature (PS) is combined with character of two distribution models Probability Density Function and standard signatures. The PS representation retains high discriminative power of PDF model, and is suited for measuring dissimilarity of images with Earth Mover's Distance (EMD), which allows for partial matches of compared distributions. It is evaluated on whole-image classification tasks from the scene and category image datasets. The comparative experiments show that the proposed algorithm has inspiring performance.

Keywords: Image classification, distribution representation, probability signature, kernel method.

1 Introduction

Image Categorization is one of the most challenging problems in computer vision, especially in the presence of scale variation, view variation, intra-class variation, clutter, occlusion, and pose changes. Generally, performance of an image categorization system depends mainly on two ingredients, the image representation and the classification algorithm. Ideally these two should be well matched so that the classification algorithm works well with the given image representation.

Local features [3, 4] are very powerful and efficient image representation for categorization problems, as seen by the state of the art performance of [1, 4]. However, the image representation produced by local features is an unordered set of feature vectors, one for each interest point found in the image. Although this kind of

representation can be used for image categorization directly by comparing image similarity with voting method and gets nice performance [5], there exist two problems aroused by the representation: The first question is that most machine learning algorithms expect a fixed dimensional feature vector as input; The other is efficiency problem, because an unordered set of feature vectors from an image includes thousands of points, each of which, i.e. local features, is a high dimension vectors usually.

Distributions of local invariant features are more constraining image representation relative to modeling image with local features directly. It can solve the problems aroused by the unordered set of feature vectors effectively. Histograms, signatures and PDF (Probability Density Function) are three mainly manners to model the representation of images with distributions of local features for classification. They all suit for discriminative classification algorithm such as Support Vector Machines (SVM), but with different matching kernel respectively.

Histograms representation with distribution of local invariant features can be got by Vector-quantized method (called as bag-of-keypoints [1]), the simplest and most popular methods in text classification and image classification, which corresponds to a histogram of the number of occurrences of particular image patterns in a given image. Here the particular image patterns are seen as keypoints which are found using a simple k-means unsupervised learning procedure over the local invariant features of the train set. Because the histogram is just a kind of coarse form of distribution description, the process of histogram quantification must lose a lot of discriminative information from local features. At the same time, just as earlier global methods based on color or gradient histograms, it cannot achieve a good balance between expressiveness and efficiency because of fixed-size bin structures [6].

A signature $\{S_j = (p_j, w_j)\}$ represents a set of feature clusters. Each cluster is represented by its mean (or mode) p_j , and by the fraction w_j of features that belong to that cluster. Since the definition of cluster is open, a histogram can be viewed as a special signature with a fixed priori partitioning of the underlying space. In contrast with histogram representation with Vector-quantized method, in which all images are limited to have the same Bin structures, the number of clusters in the signatures can vary with the complexity of images. It means that the signature is a more flexible representation of distributions. Signature feature's good performance in image categorization benefits from its categorization method with EMD (Earth Mover's Distance) kernels[6][11].

The EMD is a cross-bin dissimilarity measure and can handle variable-length representation of distributions [6][13]. It allows partial matches in a very natural way, which is important, for instance, in dealing with occlusions and clutter in image categorization applications, and in matching only parts of an image. In addition, if the ground distance is a metric and the total weights of two signatures are equal, the EMD is a true metric, which allows endowing image spaces with a metric structure. Meanwhile, the EMD can be computed efficiently by a streamlined simplex algorithm Mathematical Programming [8]. However, the signature is still not enough to retain more discriminative information for categorization task.

Among the above three modes, PDF is a kind of the most direct description of distribution and can encode more discriminative features during modeling representation

of image for categorization or recognition. However, generally, the complex form of PDFs can lead to heavy computation when applied to image categorization [2][9].

Herein we propose a novel image representation combined with characters of two distribution models, Probability Density Function and standard signatures. We call it as Probability Signatures (PS). Images categorization is completed by learning a SVM classifier with EMD (Earth Mover's Distance) kernels [6][11] based on PS. The paper evaluates the classification method by scene recognition and image categorization on different image databases. Our experimental results demonstrate that the proposed approach in this paper is superior to vector-quantized and standard signature method.

2 Improved Distribution Representation

The PS is an improving to standard signature distribution by introducing generation model. First, in the PS, the initial distribution models based on local features for each image are created by Gaussian Mixture Models. Second, the mean vector of every single model of GMMs is viewed as the center of a cluster and the summation of posteriori probability reflecting all the local features to the same single model as the weights of corresponding cluster in the PS. Therefore, the PS combines the merits of PDF with that of signature. On one hand, compared with standard signature, as each component has its own covariance structure, a point is not based solely on the Euclidean distance to the clusters but upon some local measure of the importance of different feature components. Thus different clusters can emphasize different feature components depending on the structure they are trying to represent. Finally, we obtained a much smoother approximation to the input sets density model. On the other hand, by using the probability, this approach can encode more discriminative information and capture more perceptual similarity between distributions as a local feature is allowed to respond to multi clusters. Consequently, compared with PDF, the PS retain the same discriminative information with PDF for categorization, moreover, it allows for partial matches that the SVM categorizing with PDF kernel does not possess.

2.1 Local Invariant Feature Selection

Local invariant features include detector and descriptor. Some researches[4][11] have shown that the discriminative power of local features for classification can be raised by combining multi types of detector and descriptor efficiently. This kind of integration must have some complementary in invariance such as scale and affine or in patch types such as salience or texture. We use two complementary local region detector types to extract salient image structures: The Harris-Laplace detector responds to corner-like regions, while the Laplacian detector extracts blob-like regions. In order to raise the efficiency to generate probability model, we employ the low-dimensional gradient-based descriptor called PCA-SIFT [12] as a descriptor for patches extracted at these interest regions.

2.2 Probability Signature Generation

In order to get the PB representation of image, first, we need to establish the initial distribution models based on distribution of PCA-SIFT features for each image by PDF model. We use GMMs model and its maximum likelihood parameters are estimated by EM algorithm. Given an image, its PCA-SIFT feature vectors set $X = \{x_1, x_2, \dots, x_n\}$ is extracted from detected regions, and GMMs model:

$$p(x | \theta) = \sum_{i=1}^m k_i N(x | \mu_i, \Sigma_i) \tag{3.1}$$

is estimated by EM, where $(k_i, \mu_i, \Sigma_i)_{i=1}^m$ are parameter vectors, $N(x | \mu_i, \Sigma_i)$ means a normal distribution and $k_i \geq 0, \sum_{i=1}^m k_i = 1$. Then, we generate the initial PS representations of the image:

$$S = \{(p_1, w_1), \dots, (p_i, w_i), \dots, (p_m, w_m)\} \tag{3.2}$$

where p_i is the mean vector of i th single mode of GMMs, w_i means the weights of i th mode, $w_i = \sum_{j=1}^n p(x_j)N(\mu_i, \Sigma_i), p(x_j)N(\mu_i, \Sigma_i) > \alpha$, and α is called as correlation threshold to filter some noises from local features which is a little relation with the mode. The initial PS's length is m . The final PS is formed with compression process by a compression threshold. It is noted that different images have their PSes of different length. Two thresholds and compression process will be introduced in section 3.2.

2.3 EMD Kernel-Based Classification

Supposed $S_1 = \{(p_1, w_{p_1}), \dots, (p_i, w_{p_i}), \dots, (p_m, w_{p_m})\}$ and $S_2 = \{(q_1, w_{q_1}), \dots, (q_j, w_{q_j}), \dots, (q_n, w_{q_n})\}$ are two image Probability Signatures (having the same form with standard signature). The EMD is defined as follows:

$$EMD(S_1, S_2) = \frac{\sum_{i=1}^m \sum_{j=1}^n d_{ij} f_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \tag{3.3}$$

where f_{ij} is a flow value that can be determined by solving a linear programming problem[6], and d_{ij} is the Euclidean distance between cluster centers p_i and q_j . As the EMD is a measure of dissimilarity of two signatures, to incorporate EMD into the SVM framework, we use extended Gaussian kernels[7]:

$$K(S_i, S_j) = \exp(-\frac{1}{A} EMD(S_i, S_j)) \tag{3.4}$$

The $K(S_i, S_j)$ is called the EMD kernel. A is a scaling parameter which is set to the mean value of the EMD distances between all training images to reduce the computational cost[11].

3 Experiments

We have applied our method to two domains which belong to whole image categorization: scene recognition, object categorization.

3.1 Methodology

For each categorization task, we compare our algorithm's performance with two other techniques: Vector-quantized method using linear SVM classifier in[1] and standard signature using the same classifier based on EMD kernel in[11]. All three methods share the idea of representing images based on their distribution of local invariant features and discriminative classification algorithm SVM, but they vary in distribution form and corresponding kernel in SVM. Multi-class classification is done with a SVM trained by using the one-versus-all rule.

For the Vector-quantized method, considering that classification effect is sensitivity to the size of a Bin in histogram distribution representation as image content, Bins with two sizes are selected in our experiment. We call them fine Vector-quantized(1000)and coarse Vector-quantized(200) respectively. For the standard signature scheme, we use signatures of fixed length by extracting 40 clusters with k-means for each image, although EMD can handle variable-length representation of distributions. The ground distance d_{ij} of EMD is computed by Euclidean distance in standard signature and PS.

3.2 Compression Probability Signature

A simply and flexible method to determine the length of PS is used in our experiment. The initial PS with uniform length 50 is generated according to the steps in section 2.2. Then the PS is compressed in compression procedure by setting a compression threshold which depends on the correlation threshold to some extent. If the ratio of the number of local features which posteriori probability responding a component from PS is larger than the correlation threshold to the number of total local features from the image is larger than the compression threshold, this component will be deleted from the PS of the image. The compression procedure can improve not only the performance of categorization but also efficiency of computing EMD distance between signatures.

We learn the two thresholds of PS from the same databases with scene recognition experiment introduced as next section. The correlation threshold and the compression threshold are determined respectively in two phrases: First, without executing compression procedure (i.e. under no compression threshold), determine the change curve corresponding to the correlation threshold and the recognition rate, as shown in Fig. 1. In the following phrase, according to the result of Fig.1, draw the change curve between compression threshold and recognition rate under the correlation threshold 0.4, as shown in Fig.2. By comparing the recognition rates in Fig.1 and Fig.2, it is indicated that when the compression threshold is 0.02, the recognition rate is 0.83 which has exceeded the highest recognition rate in Fig.1 and when the compression

threshold is 0.03, the recognition rate reaches its peak. However, if the compression threshold is too large, over compression of PS can cut down the recognition rate.

In the above experiments, we select 50 images per class for the training set and 10 images from the remaining in each class for the test set. The results are reported as the average recognition rate.

Experiments show when correlation threshold and compression threshold are taken with 0.4 and 0.03 respectively, the average length of all category images reduce to 29 and the performance of recognition increase 4 percentage points. Because in PS, one local feature can simultaneously contribute to multi components by probability, the compression will not lead to delete local feature directly. However, it reduces noise affection in computing EMD. so the correlation and compression threshold of PS will be set as 0.4 and 0.03 respectively in our next experiments.

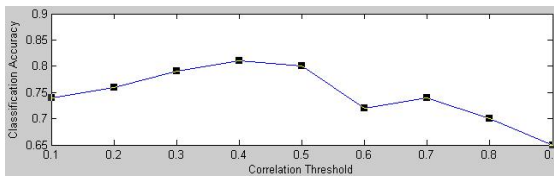


Fig. 1. Recognition accuracy with different choices of correlation threshold

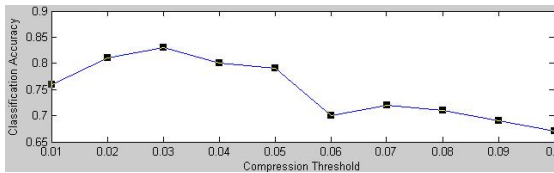


Fig. 2. Recognition accuracy with different choices of compression threshold

3.3 Scene Recognition

Our recognition task dataset is composed of eight scene categories provided by Oliva and Torralba[10]. Each category has about 300 images, and each image size is 256×256 pixels. Figure 3 shows the average recognition accuracy for a varying number of training examples per class, over 10 runs with randomly selected training examples. These are recognition rates that have been normalized according to the number of test examples per class. We can observe that our method works best among the four methods, while fine Vector-quantized approach works better than coarse and standard signature. Overall, the improved performance of our method over standard signature and two kind of Vector-quantized shows that more discriminative information are learn and more perceptual similarity between distributions are captured in our method.

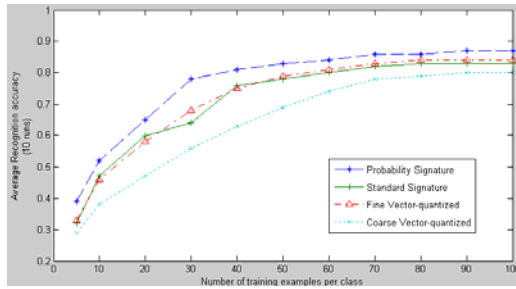


Fig. 3. Recognition results on the scene data set

3.4 Object Categorization

We evaluated four method on an object categorization task using two dataset with different styles. ETH-80[14] contains images of 80 objects from eight different classes in various poses against a simple background. All objects are almost full of the images. While Xerox7[1] includes total 1776 images which belongs 7 category object: faces, buildings, trees, cars , phones, bikes and books. These images are all of the objects in natural settings and thus the objects are in highly variable poses with substantial amounts of background clutter.

Table 1. Average classification accuracy rates in two objects datasets with four methods

Methods datasets	vector-quantized		standard signature	probability signature
	fine- 1000	coarse- 200		
Eth-80	86.2	81.8	83.3	87.6
Xer ox7	83.5	80.7	84.9	89.5

Table 1 shows that: Under object categorization tasks in images with simple background and object covering the whole image, although PS can obtain the best categorization rate, the categorization discrimina-tion power of signature is not superior to that of histogram features. However, under object category-zation tasks in natural images with complex background, signature representation has better categorization discrimination power than histogram representation, and PS categorization accuracy rate is higher than the standard Signature by 5 percentage points.

4 Conclusion

This paper proposes a novel representation of image: probability signature formed by improving the distribution of local features. The representation can capture more discriminative information for categorization in discriminative method with EMD kernel. We evaluate our method on three image databases in scene recognition and image categorization tasks. And our experiments demonstrate that the proposed approach in this paper is superior to vector quantization and standard signature method.

Acknowledgments

This work was supported by Scientific Research Fund of Zhejiang Provincial Education Department (Y200803738) and Ningbo Natural Science Foundation (2008A610027).

References

- [1] Csurka, G., Dance, C., Fan, L., Williamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 59–74. Springer, Heidelberg (2004)
- [2] Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J.: Improving Bag-of-Keypoints Image Categorisation: Generative Models and PDF-Kernels. LAVA report, 118, 141 (February 2005), http://www.ecs.soton.ac.uk/_jdrf99r/
- [3] Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 2(60), 91–110 (2004)
- [4] Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 1615–1630 (2005)
- [5] Gionis, A., Indyk, P., Motwani, R.: Similarity Search in High Dimensions via Hashing. In: Proc. of the 25th Intl Conf. on Very Large Data Bases, pp. 518–529 (1999)
- [6] Rubner, Y., Tomasi, C., Guibas, L.: The Earth Mover’s distance as a metric for image retrieval. *International Journal of Computer Vision* 40(2), 99–121 (2000)
- [7] Chapelle, O., Haffner, P., Vapnik, V.: Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks* 10(5), 1055–1064 (1999)
- [8] Hillier, F.S., Liberman, G.J.: *Introduction to Mathematical Programming*. McGraw-Hill, New York (1990)
- [9] Moreno, P.J., Ho, P.P., Vasconcelos, N.: A kullback-leibler divergence based kernel for svm classification in multimedia applications. In: *Neural Information Processing Systems*, pp. 430–441 (2004)
- [10] Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV* 42(3), 145–175 (2001)
- [11] Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *International Journal of Computer Vision* 73(2), 213–238 (2007)
- [12] Ke, Y., Sukthankar, R.: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In: Proc. CVPR, Washington, D.C., June 2004, vol. 2, pp. 506–513 (2004)
- [13] Grauman, K., Darrell, T.: Efficient Image Matching with Distributions of Local Invariant Features. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 627–634 (2005)
- [14] <http://www.vision.ethz.ch/projects/categorization/>

Adult Image Detection Combining BoVW Based on Region of Interest and Color Moments

Liu Yizhi^{1,2,3}, Lin Shouxun¹, Tang Sheng¹, and Zhang Yongdong¹

¹ Laboratory of Advanced Computing Research, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190, China

² Graduate University of the Chinese Academy of Sciences, Beijing 100039, China

³ Institute of Computer Science and Engineering, Hunan University of Science and Technology,
Xiangtan 411201, China

{liuyizhi, sxlin, ts, zhyd}@ict.ac.cn

Abstract. To prevent pornography from spreading on the Internet effectively, we propose a novel method of adult image detection which combines bag-of-visual-words (BoVW) based on region of interest (ROI) and color moments (CM). The goal of BoVW is to automatically mine the local patterns of adult contents, called visual words. The usual BoVW method clusters visual words from the patches in the whole image and adopts the weighting schemes of hard assignment. However, there are many background noises in the whole image and soft-weighting scheme is better than hard assignment. Therefore, we propose the method of BoVW based on ROI, which includes two perspectives. Firstly, we propose to create visual words in ROI for adult image detection. The representative power of visual words can be improved because the patches in ROI are more indicative to adult contents than those in the whole image. Secondly, soft-weighting scheme is adopted to detect adult images. Moreover, CM is selected by evaluating some commonly-used global features to be combined with BoVW based on ROI. The experiments and the comparison with the state-of-the-art methods show that our method is able to remarkably improve the performance of adult image detection.

Keywords: Adult image detection, bag-of-visual-words (BoVW), region of interest (ROI), soft-weighting, color moments.

1 Introduction

With the rapid penetration of the Internet into every part of our daily life, it is crucial to protect people, especially children, from exposure to objectionable information. Content-based adult image detection is one of the most powerful approaches of filtering pornography. It is mostly based on global features, whereas the false positive rate (FPR) is higher than people's expectation.

Bag-of-visual-words (BoVW) based adult image detection [1, 2] has been applied to reduce FPR because it is robust to within-class variation, occlusion, background

clutter, pose and lighting changes. Its goal is to automatically mine the local patterns of adult content, such as pornographic parts or poses, by certain clustering algorithm. These patterns are described as visual words. Therefore, visual words are the kernel of BoVW.

The usual BoVW method clusters visual words from the patches in the whole image and adopts the weighting schemes of hard assignment. Another name of patches is keypoints in some literatures [3, 4]. Hard assignment means hardly assigning the patches to the nearest visual words while generating the BoVW histogram for training or testing. However, there are many background noises in the whole image. And soft-weighting scheme is better than hard assignment.

Aiming at detecting adult images accurately, we propose a novel method of adult image detection which combines BoVW based on region of interest (ROI) and color moments (CM). There are two differences between the method of BoVW based on ROI and the usual BoVW method. Firstly, we propose to create visual words in ROI for adult image detection. The patches in ROI are more indicative to adult contents than those in the whole image. So it can improve the representative power of visual words for adult image detection. Secondly, soft-weighting scheme is adopted to improve the performance of BoVW further. Soft-weighting scheme, recently proposed by Jiang et al. [3], is better than hard assignment on both PASCAL-2005 and TRECVID-2006 datasets because of assigning a patch to *top-N* nearest neighbors. Moreover, CM is selected by evaluating some commonly-used global features. The experiments and the comparison with the state-of-the-art methods show that our method is able to remarkably improve the performance of adult image detection.

The remainder of the paper is organized as follows: section 2 introduces related works, section 3 illustrates our method in detail, section 4 shows the experiments and section 5 concludes the paper.

2 Related Works

The traditional approach of content-based adult image detection is based on the global low-level features which include color, shape, texture and etc. Forsyth et al. [5] construct a human figure grouper after detecting skin regions, but consuming too much time and low detection accuracy are the two shortcomings. Zeng et al. [6] implement the image guarder system to detect pornographic images by different kinds of global features. Kuan and Hsieh [7] use image retrieval technique and extract visual features from skin regions. Q. F. Zheng et al. [8] use an edge-based Zernike moment method to detect harmful symbol objects. Rowley et al. [9] adopt 27 visual features for Google to filter adult-content images, including color, shape, texture, face and etc. Tang et al. [10] employ latent Dirichlet allocation to cluster images into some subsets and then combine SVM classifiers on one subset to rate adult images. Nevertheless, it is difficult to detect adult images in the presence of within-class variation, occlusion, pose and lighting changes.

To cope with the difficulty, BoVW approaches have been applied. Wang et al. [1] explore an algorithm to reduce the number of visual words and integrate it with spatial distribution to detect adult images. Deselaers et al. [2] combine BoVW with color histogram to classify images into different categories of adult content. Both of them use difference of Gaussian (DoG) detector and scale-invariant feature transform (SIFT) descriptor.

Visual words are usually clustered from the patches in the whole image. The patches are always assigned to the nearest visual words, as it called hard assignment. We name the preceding procedures “the usual BoVW method”. However, visual words are clustered from the patches in the whole image and these patches are full of background noises. Furthermore, it has not been discussed in-depth that the effects of weighting schemes and combination with global features on adult image detection.

3 Our Method

To detect adult images accurately, we combine BoVW based on ROI with color moments (CM). In this section, we will illustrate our method in detail.

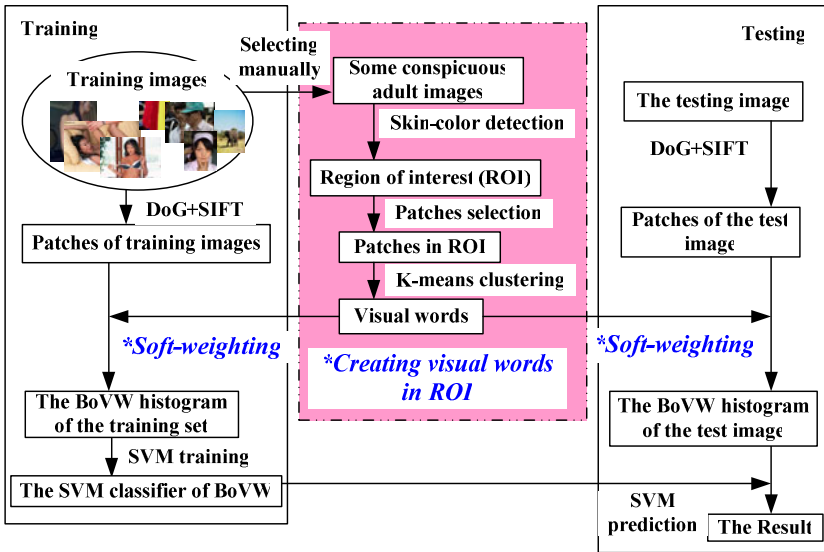


Fig. 1. BoVW based on ROI

3.1 BoVW Based on ROI

As shown in Fig. 1, we propose the method of BoVW based on ROI to improve the performance. We use DoG detector and SIFT descriptor, too. Two differences between

BoVW based on ROI and the usual BoVW method are marked with the blue and bold-faced words and with a bar nearby. Firstly, visual words are created in ROI. Secondly, soft-weighting scheme is adopted. The training and testing procedures are depicted respectively on the left and the right column in Fig. 1. The details of ROI detection and soft-weighting scheme are introduced in the subsections respectively.

3.1.1 ROI Detection

ROI means the subareas containing pornographic parts or poses in the field of adult image detection. Skin-color regions include these pornographic subareas. Thus, we apply the kind of skin-color models [11] to capture ROI.

In Fig. 2, some examples of SIFT patches in ROI are given. To avoid the objectionable information, we transform the skin-color region into white and other regions into black. SIFT patches are represented by the red points in the images. We can observe clearly that many background noises are removed after selecting patches from ROI.



Fig. 2. Some examples of SIFT patches in ROI

Garcia and Tziritas [11] have shown that skin-like pixels are more correlated with C_r and C_b components than Y component. Thus, the input image is transformed from the RGB color model to the YC_bC_r color space. A pixel is considered to be skin-like if its C_r and C_b components meet the following constraints:

$$C_r = \max\{-2(C_b + 24), -(C_b + 17), -4(C_b + 32), 2.5(C_b + \theta_1), \theta_3, 0.5(\theta_4 - C_b)\}; \quad (1)$$

And

$$C_r = \min\{(220 - C_b) / 6, 4(\theta_2 - C_b) / 3\}. \quad (2)$$

There are two constraints for θ_1 , θ_2 , θ_3 , and θ_4 :

If $Y > 128$, then

$$\begin{aligned} \theta_1 &= -2 + (256 - Y) / 16; \\ \theta_2 &= 20 - (256 - Y) / 16; \\ \theta_3 &= 6; \\ \theta_4 &= -8. \end{aligned} \quad (3)$$

Otherwise,

$$\begin{aligned}
 \theta_1 &= 6; \\
 \theta_2 &= 12; \\
 \theta_3 &= 2 + Y / 32; \\
 \theta_4 &= -16 + Y / 16.
 \end{aligned} \tag{4}$$

3.1.2 Weighting Scheme

Weighting schemes play an important role in generating the BoVW histogram and thus have great effects on the performance of BoVW. Generally speaking, the patches are hardly assigned to the nearest visual words. These schemes are named hard assignment, such as binary weighting, term frequency (TF) and the product of term frequency and inverse document frequency (TF×IDF). Among them, binary weighting always produces top or close-to-top performance on the datasets of PASCAL-2005 and TRECVID-2005 [4]. Binary weighting indicates the presence and absence of a visual word with values 1 and 0 respectively.

Soft-weighting scheme, recently proposed by Jiang et al. [3], outperforms the preceding schemes of hard assignment on both PASCAL-2005 and TRECVID-2006 datasets. Instead of assigning directly a patch to its nearest neighbor, soft-weighting assigns a patch to $top - N$ ($N = 4$ empirically) nearest neighbors to weight the significance of visual words. Suppose that there are M visual-words in a vocabulary and the $top - N$ nearest visual-words are selected for each patch in an image, each component v_k of a M -dimensional vector $V = [v_1, v_2, \dots, v_k, \dots, v_M]$ represents the weight of a visual word k in an image such that

$$t_k = \sum_{i=1}^N \sum_{j=1}^{L_i} \frac{1}{2^{i-1}} sim(j, k) \tag{5}$$

where L_i represents the number of patches whose i th nearest neighbor is visual word k . The nearest $sim(j, k)$ means the similarity between the patch j and the visual word k .

Weighting schemes are also relative to the number of visual words. Jiang et al. conclude that the performances of BoVW are similar using soft-weighting on TRECVID-2006 when the number of visual words ranges from 500 to 10,000 [3]. If the number of visual words becomes large, the cost increases in clustering, computing the BoVW histogram and running the classifier.

Therefore, we adopt soft-weighting as the weighting scheme and employ the K-means algorithm based on DBSCAN [12] to create visual words whose number is around 500. The clustering algorithm has some advantage, such as rapidness, robustness to noises, finding any shape of clusters in spaces, and adjusting clusters adaptively.

3.2 Classification and Combination

To integrate the advantage of global features, we combine BoVW based on ROI with color moments (CM). After evaluating some global color features — CM, color

correlogram and color histogram, we find that CM is the best one. Then we concatenate it with some global texture features, such as texture co-occurrence, Haar wavelet and edge histogram. But the performances of its concatenation are not as good as CM alone. CM provides a measurement for color similarity between images. In CM, we calculate the first 3 moments of 3 channels in Lab color space over 5×5 grid partitions, and aggregate the features into a 225-dimension feature vector.

The supported vector machines (SVM) classifier has been one of the most popular classifiers for BoVW-based image classification [2, 3, 4] and adult image detection based on global features [7, 9, 10, 11]. We employ SVM with Gaussian radial basis function (RBF) kernel to obtain good performance. The form is as follows:

$$g(x) = \sum_i \alpha_i y_i k(x_i, x) - b = \sum_i \alpha_i y_i e^{-\rho d(x_i, x)} - b = \sum_i \alpha_i y_i e^{-\rho \sum_i |x_i - x|^2} - b \quad (6)$$

In the formula (6): $k(x_i - x)$ is the response of a kernel function for the training sample x_i and the test sample x , which measures the similarity between the two data samples; y_i is the class label of x_i ; α_i is the learned weight of the training sample x_i , and b is a learned threshold parameter.

We combine the classification results of BoVW based on ROI and CM with “average fusion”. Average fusion is one of the commonly-used “late fusion” methods.

4 Experiments

In this section, our method is evaluated step by step. (1) Our dataset and the baseline are reported. (2) We do experiments to show the effect of soft-weighting scheme on BoVW in subsection 4.2. (3) The improvement of creating visual words in ROI is evaluated in subsection 4.3. (4) Subsection 4.4 compares some commonly-used global features. (5) In subsection 4.5, the combination of BoVW based on ROI and color moments is compared with the baseline and many previous works.

4.1 Our Dataset and the Baseline

As Table 1 shows, we provide statistics of our dataset. We collect 90,000 images from Internet. The training set is made up of 10,000 adult images and 40,000 non-adult images. The testing set has 10,000 adult images and 30,000 non-adult images. Both sets include 10,000 non-adult images containing body parts, such as faces, hands, feet and trunks. We do all these experiments in the visual studio 2003 environment with the machine of 1.86 GHz Duo CPU and 2GB memory. We evaluate our method with receiver operating characteristic (ROC) curves. A ROC space is defined by false positive rate (FPR) and true positive rate (TPR) as x and y axes respectively. There is no common dataset in the field of adult image detection. Consequently, we build up the baseline named CH+EH. The baseline uses the features concatenating color histogram (CH) and edge histogram (EH) and is classified by the SVM classifier with the Gaussian RBF kernel.

Table 1. Statistics of our dataset

	The training set	The testing set
Adult images	10,000	10,000
Non-adult images	40,000	30,000

4.2 The Usual BoVW Method with Different Weighting Schemes

In this subsection, we estimate the effect of soft-weighting scheme on BoVW. ROC curves of BoVW with binary weighting and soft-weighting are respectively abbreviated to SIFT-BW and SIFT-SW. The numbers in the parentheses are the size of visual words. According to Fig. 3, soft-weighting scheme is a little better than binary weighting.

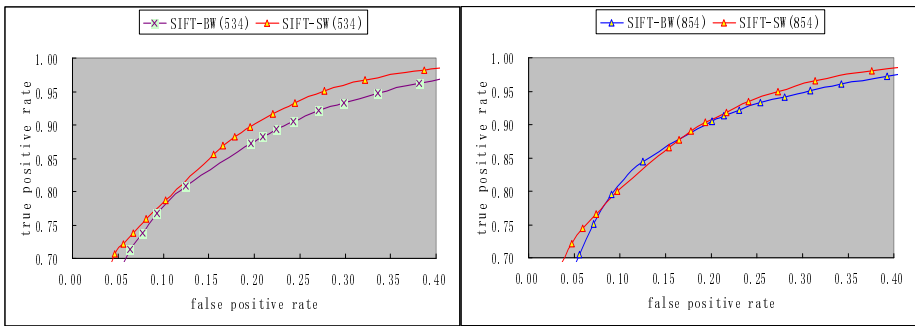


Fig. 3. The ROC curves of the usual BoVW method

4.3 Evaluation of BoVW Based on ROI

We evaluate the advantage of creating visual words in ROI. SIFT-ROI-SW is on behalf of the ROC curve of BoVW based on ROI. All the curves in Fig. 4 adopt soft-weighting scheme. So we can conclude that the method of BoVW based on ROI can remarkably improve the performance of BoVW based adult image detection.

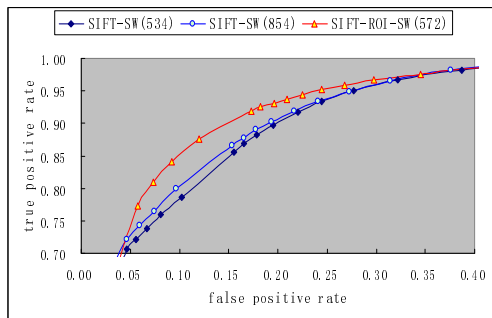


Fig. 4. The performance of BoVW based on ROI

4.4 The Performance of Color Moments

After evaluating some commonly-used global color features — color moments (CM), color correlogram (CC) and color histogram (CH), we find that CM is the best one. Then we concatenate it with some global texture features, such as texture co-occurrence (TC), Haar wavelet (HW) and edge histogram (EH). We can infer from Fig. 5 that the performances of the concatenations of CM are worse than CM alone.

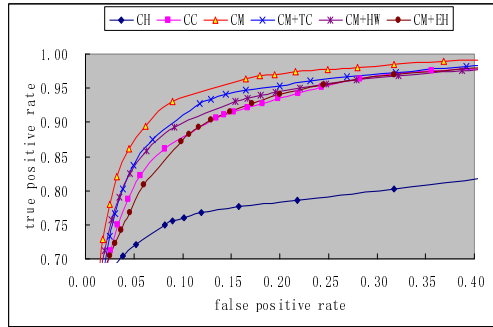


Fig. 5. Adult image detection based on the global features

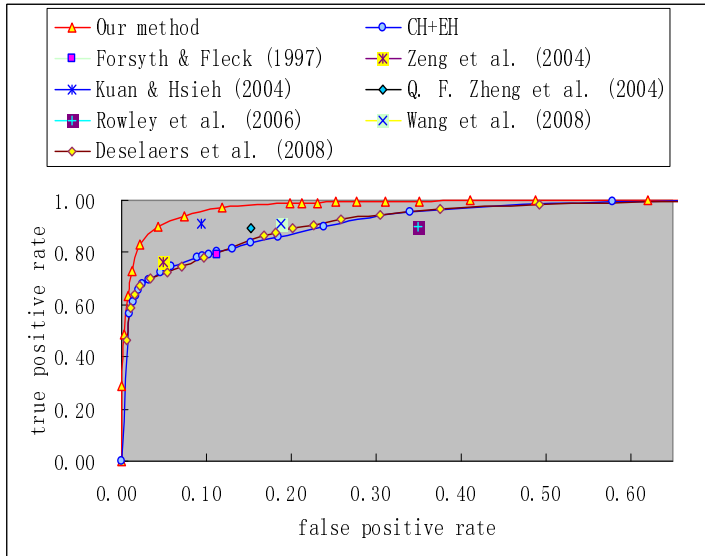


Fig. 6. Performances compared with many previous works

4.5 Evaluation of Our Method

To detect adult images accurately, we use “average fusion” to combine BoVW based on ROI and color moments. As shown in Fig. 6, our method outperforms the baseline

and the state-of-the-art method (Deselaers et al. [2]) on the same dataset. [2] combines BoVW with color histogram to classify images into different categories of adult content. The ROC curves of our method, CH+EH and Deselaers et al. [2] are represented respectively as the red, blue and purple curve. The points in Fig. 6 are on behalf of the performance of some previous works in section 2. The results experimentally show that our method is able to remarkably improve the performance and outperforms many previous works, including the state-of-the-art method (Deselaers et al. [2]).

5 Conclusions

To filter adult contents proliferating on the Internet effectively, we propose a novel method of adult image detection which combines BoVW based on ROI and CM. We create visual words in ROI and adopt soft-weighting scheme to improve the performance of BoVW. The patches in ROI are more indicative than those in the whole image. Furthermore, soft-weighting scheme is better than that of hard assignment and CM is selected by evaluating some commonly-used global features. The experiments and the comparison with the state-of-the-art methods show that our method is able to remarkably improve the performance.

Acknowledgments. This work is supported by the National Basic Research Program of China (973 Program, 2007CB311100); National High Technology and Research Development Program of China (863 Program, 2007AA01Z416); National Nature Science Foundation of China (60873165, 60802028); Beijing New Star Project on Science & Technology (2007B071); Co-building Program of Beijing Municipal Education Commission.

References

- [1] Wang, Y.S., Li, Y.N., Gao, W.: Detecting pornographic images with visual words. *Transactions of Beijing Institute of Technology* 28, 410–413 (2008) (in Chinese)
- [2] Deselaers, T., Pimenidis, L., Ney, H.: Bag-of-visual-words models for adult image classification and filtering. In: 19th International Conference on Pattern Recognition, Tampa, USA, pp. 1–4 (2008)
- [3] Jiang, Y.G., Ngo, C.W., Yang, J.: Towards optimal bag-of-features for object categorization and semantic video retrieval. In: 6th ACM International Conference on Image and Video Retrieval, Amsterdam, Netherlands, pp. 494–501 (2007)
- [4] Yang, J., Jiang, Y.G., Hauptmann, A.G., et al.: Evaluating bag-of-visual-words representations in scene classification. In: 9th ACM SIGMM International Workshop on Multimedia Information Retrieval, Augsburg, Germany, pp. 197–206 (2007)
- [5] Fleck, M.M., Forsyth, D.A., Bregler, C.: Finding naked people. In: 4th European Conference on Computer Vision, Cambridge, UK, pp. 593–602 (1996)
- [6] Zeng, W., Gao, W., Zhang, T., et al.: Image guarder: an intelligent detector for adult. In: 6th Asian Conference of Computer Vision, Jeju Island, Korea, pp. 198–203 (2004)
- [7] Kuan, Y.H., Hsieh, C.H.: Content-based pornography image detection. In: International Conference on Imaging Science, System and Technology, Las Vegas, USA (2004)

- [8] Zheng, Q.F., Zeng, W., Gao, W., et al.: Shape-based adult images detection. In: 3th International Conference on Image and Graphics, Hong Kong, China, pp. 150–153 (2004)
- [9] Rowley, H.A., Yushi, J., Baluja, S.: Large scale image-based adult-content filtering. In: 1st International Conference on Computer Vision Theory and Applications, pp. 290–296 (2006)
- [10] Tang, S., Li, J., Zhang, Y., et al.: PornProbe: an LDA-SVM based Pornography Detection System. In: ACM International Conference on Multimedia, Beijing, China (2009)

Multimedia Speech Therapy Tools and Other Disability Solutions as Part of a Digital Ecosystem Framework

David Calder

Curtin University of Technology,
Bentley, Perth,
West Australia
Tel.: 61-8-9266 2875
david.calder@cbs.curtin.edu.au

Abstract. Curtin University has developed a multimedia prototype system for use by speech therapists to assist in the rehabilitation of motor speech impaired patients. These are often stroke victims who have to relearn the ability to communicate effectively. The process is usually laborious and instead of the usual card-based prompts used by the therapist, this multimedia solution offers screen and audio prompts together with high quality digitised speech. The result is a reduced work load for the therapist. In parallel with this work, is a Mobility Aid Assistive Device Program for the visually impaired. The support clusters for all these groups, whether they be therapists, caregivers, manufacturers of hardware or software do not interact to the extent they might. The author proposes a collaborative digital ecosystem framework that would assist this challenge.

Keywords: Multimedia, computer-based therapy, speech therapy, cardiovascular accident, assistive technology, sound interface displays, laser, disabled, infrared, long cane, portable electronic device, sensory channels, visually impaired, ultrasonic pulse-echo, ambient sound cues.

1 Introduction

There is a worldwide shortage of speech training services to accommodate the clients who require speech therapy [1]. Cardiovascular accidents are the third largest killer in Australia after heart disease and cancer. Those that survive this trauma usually have need for rehabilitation. Motor impairment of the speech organ is a common occurrence. We have produced a Multimedia system which replaces the chart and paper-based cues found in many speech therapy units. The aim is to assist speech therapists by decreasing the time they have to spend in direct contact with patients. More importantly, there are also a number of benefits to clients and these will be discussed later.

As part of the Digital Ecosystem and Business Intelligence Institute (DEBII) research team at Curtin University, we worked closely with speech professionals from Royal Perth Hospital in order to tailor the interface exactly to their requirements. We were guided by practicing speech therapists in ensuring existing and well established interactive therapy strategies for certain classes of dysarthria, were followed as

closely as possible. The first version of the system has undergone beta testing in Western Australia. This is a straightforward software solution using standard personal computers already available to therapists within their environment.

The author has also been researching assistive devices for the visually impaired and portable speech aids for the speech impaired. These, on the other hand, require dedicated hardware and the design of *extreme demand* interfaces in order to offer the most appropriate solution. However, the support infrastructure clusters, for each of these above examples, do not collaborate or communicate to any great degree.

2 Cerebro-Vascular Accident

When the oxygen supply to the brain is blocked or when an artery in the brain is ruptured, a common stroke results. The former is the result of a clot traveling to the brain and occluding blood flow in a cerebral artery [2]. Once the oxygen supply has ceased, hydrogen ions propagate within the brain cells and damage the cells. Within four to eight minutes, the cells begin to die, and of course this destruction is irreversible. In the case of a reduced blood flow, cells may stop functioning but later recover when normal perfusion is restored [3].

Cardiovascular accidents are the third largest killer in Australia and in many other countries of the western world. The highest killer is heart disease followed by cancer [4]. Fifteen percent of people over the age of forty living in the western world die of cardiovascular accidents. A further fifteen percent are also stroke victims who survive and require some form of rehabilitation [5]. Of this group five percent make a full recovery whilst ten percent suffer some permanent disability [6].

As medical technologies have improved and the relative death rate has fallen, so the need to offer suitable rehabilitation services to the victims who have survived this trauma has increased. Patients are left with multiple disabilities and these often include loss of speech. The anterior portion of the brain is associated with speech production and damage to this area can result in motor speech disorders [7]. The severity of the motor speech disorder may vary from person to person but even where spontaneous recovery is evident, some speech rehabilitation will probably be carried out. In the case of more severe speech impairment, long term rehabilitation may result [8]. In most instances, the road to recovery is long and difficult, placing stress on both client and therapist. Any means that may assist this load for both the client and the therapist should be investigated.

3 Hardcopy Patient Cue Cards

During a standard speech therapy session, the client is presented with a number of cues which include symbols representing phonemes/vowels, drawings indicating air flow through the mouth and nose and the many combinations of lip, teeth and tongue positions which are essential in general articulate speech. Cards were identified as part of a major problem by speech professionals during the initial data analysis of the old system at Royal Perth Hospital, Shenton Park in Western Australia. The therapist is under pressure to find the correct flash card or phoneme chart as she sits with the

client. Together with these paper-based cues she uses her own speech and face to offer sound and visual cues to the client. In order to produce the consonant M for example, the client may be asked to place the lips together and allow the air to flow through the nasal cavity. The therapist will give an example by producing the sound herself and simultaneously accentuate the positioning of the lips.

The client is then asked to repeat this step. Consonant and consonant/vowel combinations such as this need constant repetition by the client and consequently the therapist before any improvement is expected. This face-to-face contact can be very tiring for both the client and the therapist. As therapy is done on a one to one basis, duration of contact per individual per week may be limited to only a few hours. Clients are therefore being deprived of the continuous training which would benefit their rehabilitation.

4 Multimedia Therapy Tool

The concept of a system to assist in speech therapy resulted from earlier work with speech therapists in the United Kingdom in developing a speech aid prototype [8]. The development of the therapy tool depended on a requirement to address some of the most significant problems associated with conventional therapy methods. The Department of Speech Pathology at the Royal Perth (Rehabilitation) Hospital, Australia, was involved in the development of the system and provided the old model of operations for analysis from which the computer-based system was developed. Assistant Professor G. Mann, now in the Department of Communication Disorders at the University of Florida and widely recognised as a key authority, worked closely with us over several years in developing the original version of the multimedia tool [10].

The therapy tool aims to relieve constant therapist/patient supervision, particularly where time consuming repetitive tasks are involved. Therapists can use their time more effectively in planning new goals whilst the computer provides visual and sound cues to the client. Therapists no longer have to organise cue cards or sort through hundreds of icons and drawings. These were seldom in colour whereas the computer-based system augments these traditional methods by using colour and animation. The latter was something that could not be achieved on loose pieces of paper or cardboard! Consequently the therapy process can run more smoothly and effectively as all cues are presented on the screen and/or produced by the high quality stored speech system.

There are a number of other benefits which particularly relate to patients and have been highlighted during beta testing. It has been found that the system could be used at home as well as in the conventional environment of the speech therapy unit. Where a client is left with the computer and removed from the clinical surroundings, the stress of embarrassment is removed. This is particularly evident when an older client has to "relearn how to speak" in the presence and under the direction of a young speech therapist. The stress placed on the client during these encounters should be seriously considered.

Another bonus for the system is that it is based on a standard IBM compatible PC and could even be run from a portable notebook mounted to a wheelchair. Most households now have several computers with a more than adequate built-in sound

systems [11]. This means that in most instances, only the software need be installed in the home to allow therapy to continue. Other members of the family could be involved in the rehabilitation and the clinical sessions at the hospital could be used for the monitoring of progress and not be associated with intrusion into the privacy of the patient [12]. Communication within the family is an important part of rehabilitation and the therapy tool could help in promoting this. The high cost of therapy is relieved but not replaced by this augmentative system [13].

Since the therapy tool saves clinician time, it may also help relieve the shortage of speech therapy services. If the amount of time spent with each client is reduced, the clinician could then take on a greater case load.

The nature of a computer-based system is that, "it doesn't have to see another patient in an hours' time". Therefore clients may train at their own pace under no stress. Sound and visual cues may be repeated over and over again without the pressure of supervision. Of course not all patients would be suitable candidates for this level of freedom but initial tests have indicated great success for certain motor impaired victims.

5 Multimedia Animation and Stored Speech

Asymetrix Multimedia Toolbook was used to develop this system. It operates in a high level, Windows-based object-oriented programming environment. Toolbook uses a book metaphor for multi screen production. Each screen is described as a page and all pages in a production are called a book. Making use of these properties pages were constructed to emulate the icons and visual cues used in conventional therapy.

Animated visual representations of the vocal system were added to the standard set of cues. High quality digitised speech was recorded from a practicing speech therapist so that an exact representation of a set of speech sounds could be achieved. As the normal therapy session involves a progressive set of prompts to initiate speech from the client, the therapy tool was set up with a progressive set of cue buttons which gradually give more and more assistance to the client in achieving a particular goal. For example, this might be help in pronouncing the consonant F. A set of six cue buttons mounted vertically on the right hand side of the screen allow for this positive reinforcement. These cues may take visual or spoken form. Another reason for a progressive interactive strategy is the problem of complexity which has plagued so many previous interface designs for the disabled [14].

The sound cues may be both instructive and exemplary. One cue button will produce an instruction such as, "place the lower lip against your upper front teeth and blow" whilst another produces the target as required, the actual sound "fff ...". Another button links the consonant to a word so as to place the target sound in context. Yet another offers the word in the context of a sentence.

All cue buttons and command buttons have explanatory logos which appear at the bottom of the screen when the cursor moves over their area. There are over 300 digitised sound recordings associated with the therapy tool.

The system is normally controlled by a mouse and test models incorporate this means of user control. Other remote switch devices are currently being investigated. This will make the system more versatile and therefore suitable for quadriplegics [15]. A system of this type should be flexible enough to adapt to suit the needs of each

individual client and varying physical disabilities [16]. The system being used by a speech therapy patient and therapist can be seen in Figure 1 below.



Fig. 1. Speech therapy session with stroke patient

Research by the author and therapists has established that phoneme synthesis is not suitable for the speech requirements of this system. A robotic sounding voice may be acceptable for certain games programs and simple communication tools but can never be adequate for the variety of accents, languages and key sounds that may be required by therapists. The Institution of Electrical and Electronics Engineers [17], published a recommended practice for speech quality measurements, but this has never been fully adopted as the basis for standardised testing [18]. Edwards showed that no existing rule-based synthesiser came close to passing the test of being indistinguishable from natural speech.

To date, the system has been tested on a small pilot sample of patients in Australia. Initial tests at the Royal Perth Hospital have indicated that the response to the system for middle age stroke victims was very positive. Other age groups including children are included in forward planning for the later version. The interface has proved easy to understand and selected stroke patients have managed to control the system themselves, i.e. choosing and activating cue buttons.

Future tests will be coordinated to fine tune the system before full trials take place. It must be emphasized that full and robust trials have yet to be carried out. This is a work in progress. However, feedback from the first pilot tests highlighted the need for fast adaptability in design refinement which could be partly addressed by the digital ecosystem paradigm [19]. Issues such as universal access are also relevant [20].

6 Other Support Clusters

Speech therapy tools form just part of a wider picture of related hardware or software based support systems, all falling roughly within the catchment area of assistive

technologies for the disabled. Most of these systems require special interface design and may cater for temporary or permanent loss of speech. The software tool described above is assisting in a hopefully temporary loss of speech. However, this loss may also be permanent. In which case alternative and augmentative devices such as portable speech aids (producing synthetically generated speech) are required. Even though these tools are closely related and familiar to most speech therapists, there is a huge gulf between the teams that produce and support both solutions.

This communication gulf turns into a chasm when more diverse assistive technologies are examined. These may include sensory loss such as hearing or visual impairment. The author has been working on both the above projects involving speech therapy and a substitution aid for natural speech. More recently we have been developing a navigation device for the blind. And this is where the infrastructure support clusters become very different.

There are numerous mobility aids and orientation mapping devices for the visually impaired on the market at present, some with significant drawbacks. Many assistive technology devices use ultrasonic pulse-echo techniques to gauge subject to object distance. Some use infrared light transceivers or laser technology to locate and warn of obstacles. These devices exhibit a number of problems, the most significant of which are related to the interface display that conveys navigation/obstacle warning information to the user. Other sensory channels should not be compromised by the device. This is exactly what can happen when, for example, audio signals are used in obstacle warning on/off displays.

The DEBII team has developed a prototype device, which it is hoped, will be the first step in addressing some of the above listed problems. Patent searches indicate this working prototype has a unique tactile interface design which, even in its basic form, should have distinct user advantages. As with some of the sonar systems listed above in the paper, this first prototype is best suited to outdoor use. Future models are not limited to sonar technology, however.

The design criteria has and will in the future, concentrate on intuitive interfaces that do not compromise certain other all-important sensory channels. These interfaces, now under development, will be configurable for users who are both deaf and visually impaired. There will also be an emphasis on ease of learning and use. It is unacceptable to expect someone, who may have multiple disabilities, to undertake a long and complex learning program in how to use a new device. The Innovation in the author's mobility aid design may be summarized as follows:

'A portable mobility aid incorporating warning obstacle ahead information with mapping capabilities as a stand-alone device.'

At this stage, no further technical specification can be given due to IP novelty protection. It is hoped that in future papers, we will be able to concentrate more freely on the detailed technical aspects of the design. However, the current system uses ultrasound for range-finding and is mounted on a cane for test purposes. Initial tests have proved the system to be at least as effective as many of the alternative commercial systems available.

7 Digital Ecosystem Models

Issues of complexity with respect to individual requirements must be seen within the context of a wider ecology of the particular user, with that person clearly at the centre, contributing to a team solution. An established and highly successful ecological approach to designing individualized education programs for the disabled student has been refined over twenty years into a highly recommended model and is now regarded as 'best practice' [21].

This ecological approach has not as yet permeated all areas of disability support. However, the power of the digital ecosystem framework is now accepted within many other disciplines, particularly with respect to small enterprise collaboration [22].

Within small business, the advent of the web has allowed sales penetration over vast distances. Accompanying these advances have come new modes of marketing and partnership possibilities that would have been impossible only a few years ago. With this connectivity has come a fertile and dynamic business theatre that cannot be avoided if small enterprises are to survive. This interaction has led to collaborative workflow models [23].

The logic behind collaborative workflows is to produce a sequence of activities that not only produce a meaningful result, but also to facilitate small groups working together to achieve common goals. The actual physical distance and associated limitations between these entities then becomes less important as web based tools are used to link enterprises and their common aspirations [24]. The entities themselves may be small companies competing against large predator corporations, or widely dispersed cottage industries (such as those associated with assistive devices) with a common interest [25].

Beyond the standard empowerment the digital ecosystem model has provided, are more specific areas that are pertinent to such groups operating in harmony. One of the most important of these is trust evaluation [26]. Other typical support areas are logistics and privacy [27, 28]. These would act as foundations for the framework I that is proposed.

Digital Ecosystems For Assistive Technology clusters (DEFAT) is a proposed collaborative cluster-based ecosystem model, neither limited by distance between clusters nor the particular disability types associated with each of the clusters. Individual clusters may include a range of specialist personnel associated with the support of a client requirement. The output of such an environment would not only be the efficient research and development of appropriate assistive devices, but also result in more streamlining for the teams in their everyday support of an individual, be that speech therapy for dysarthria patients or training in the use of a long cane or mobility aid for the visually impaired.

8 DEFAT Structure

With each client representing a nucleus at the centre of his or her support cluster, an individual's local ecological environment has been acknowledged (as discussed and cited in previous sections) as a worthwhile starting point, offering a framework from which specialist support action may be fleshed out.

Each support cluster would have a number of clients within its particular category of disability. Cluster membership would not be determined by distance or physical boundaries. The aim would be to maximize use of the digital ecosystem paradigm in order to break existing physical boundaries. By applying a DEFAT strategy, current digital technologies such as mobile, the internet and video conferencing can be coordinated and optimized to deliver the best outcome for all members of this ecosystem.

Open-ended but novel design solutions would be encouraged from both hardware and software developers. The sharing and exchange of common modular solutions at both a functional and user interface level would be part of the ecosystem membership requirement.

The difference would be in the focus and modular consideration of appropriate novel and relevant ideas, when first considering I.P. matters. This will not always be relevant to designs, but when it is, it should in fact enhance the potential for success within the DEFAT community itself, as well as in a wider context (external to the ecosystem).

Those academic cluster members who currently work within a limited research environment with a very small interest group would have the opportunity to share their research and ongoing projects on a wider stage within the digital ecosystem. Cross-disciplinary interaction would be nurtured by DEFAT.

A cluster of people with a vast range of interdisciplinary skills would focus on a user group of people all with a common disability. There would be many separate clusters, meeting the challenges of specific needs of different disability groups. As now, it may be assumed that special education specialists, therapists, medics, academics, engineers and particularly hardware and software experts would form part of each cluster, the main difference being a recognition of the greater ecosystem in which each cluster coexists and operates.

Users at the center of each cluster, the nucleus, would determine the nature of the environment. Clusters would communicate with each other for a common good and the ecosystem itself would be able to benefit from its size in terms of external links and its critical mass.

A starting point for such a structure may take into account the problem as defined by Liu et al when referring to building the right systems and the need for better tools and environments in their paper on component-based medical and assistive devices and systems [29]. They put forward a ten-year roadmap, which fits well as a start to implementing the DEFAT paradigm.

Clusters need to be client centered, taking into account breakthrough research such as that of Bach-Y-Rita into sensory substitution [30] and Merzenich into brain plasticity [31].

A global advantage and DEFAT's greater mass would benefit the ecology on many levels. There would be lower manufacturing costs than is now associated with small-run dedicated systems production. This advantage would result from greater demand for DEFAT modular units across clusters and existing boundaries. Relatively large production runs catering for a global DEFAT module demand would drive production costs down.

9 Conclusion

The Digital Ecosystem and Business Intelligence Institute at Curtin University in Western Australia is developing a multimedia-based training tool for use by speech therapists and their clients. A more advanced system is currently under development. This new version is specifically aimed at people with motor speech disorders. It is believed that the design may, in time, be developed into a wider range of solutions which may incorporate related therapy requirements.

The author compares the development of the above speech therapy and speech aid solutions with other assistive device development examples; specifically those with sensory impairments, such as blindness. As most of these software and dedicated hardware solutions are produced by small, unlisted companies, there is little in the way of publicly available, reliable sales figures, and as such the addressable market success is not well defined. However, interviews conducted with industry experts, in addition to the small size of the companies themselves, suggest that these competing devices have so far failed to achieve any significant market presence, and in many cases, have inherent and significant user interface design issues. The author's prototype programme and proposed DEFAT framework development will, hopefully, be a step in the right direction.

References

1. Editor, ASU Designs Program To Address Shortage of Speech Therapists, Phoenix Arizona News (May 22, 2008), doi=<http://evliving.com>
2. Hwer, R.L., Wade, D.T.: *Stroke - A Practical Guide Towards Recovery*. Methuen Australia, Australia (1986)
3. Kaplan, N.M., Stamler, M.D.: *Prevention of Coronary Heart Disease*. W.B. Saunders and Co., Philadelphia (1983)
4. Gross, P.F.: *The Total Economic Costs of Stroke in Australia*. Technology Assessment Corporation Pty Ltd., NSW, Australia (1991)
5. Hird, K.M.: *Communicative Competence, Self-care and Mobility in Left and Right Hemisphere Adult Stroke Patients*. Curtin University of Technology, Western Australia (1987)
6. Dyken, M.L., Woff, P.A., Barnett, J.H.: Risk Factors in Stroke: A Statement for Physicians by the subcommittee on Risk Factors and Strokes. *Stroke* (6), 1105–1111 (1984)
7. Darley, F.I., Aronson, A.E., Brown, J.R.: *Motor Speech Disorders*. W.B. Saunders Co., London (1975)
8. Gloag, D.: Rehabilitation After Stroke: What is the potential? *British Medical Journal* 290, 699–701 (1985)
9. Calder, D.J., Lister, P.F., Mack Watson, J.D.: Large Vocabulary Speech Aid Incorporating Softkey Switching. In: *Third Symposium International De Ingeniera Biomedica*, Madrid, pp. 755–781 (1987)
10. Mann G.: *The Mann Assessment of Swallowing Ability*. Cengage Delmar Learning, New York (2002)
11. Edwards, A., Pitt, I.: *Design of Speech-Based Devices*. Practitioner's Series. Springer Professional Computing, London (2002)
12. Alkalay, L., Asserman, M.W.: *Cerebrovascular Accidents: Care and Rehabilitation*. The Practitioner 227, 469–473 (1983)

13. Hodge, S.: Why is the potential of augmentative and alternative communication not being realized? Exploring the experiences of people who use communication aids. *Disability and Society* 22, 466 (2007)
14. Seidle, F.: Is Home Care Less Expensive? *Health Social Work* 2(5), 10 (1977)
15. Vanderheiden, G.C., Lloyd, L.L.: Communication Systems and Their Components. In: *Augmentative and Alternative Communication: American Speech and Language-Hearing Association, Rockville(Maryland)*, vol. 49, p. 162 (1986)
16. Reichle, J.J., York, J., Sigafoos, J.: *Implementing Augmentative and Alternative Communication*, pp. 239–256. Paul Brookes Publishing Co., London (1991)
17. IEEE: IEEE Recommended Practice For Speech Quality Measurements. *IEEE Transactions on Audio and Electroacoustics*, AU 17 (3), 225–246 (1969)
18. Edwards, A.D.: *Speech Synthesis: Technology for Disabled People*, pp. 12–40. Paul Chapman Publishing Ltd., London (1991)
19. Chang, E., West, M.: Digital Ecosystems and Comparison to Collaboration Environment. *WSEAS Transactions on Environment and development* 2, 1396–1404 (2006)
20. Zimmermann, G., Vanderheiden, G., Gilman, A.: Prototype Implementations for a Universal Remote Console Specification. In: *CHI 2002 Conference on Human Factors in Computing Systems, Minneapolis, MN*, pp. 510–511 (2002)
21. Rainforth, B., York, J., Macdonald, C.: *Collaborative Teams for Students With Severe Disabilities*, pp. 71–83. Paul Brookes, Baltimore (1993)
22. Chang, E., West, M.: Digital Ecosystems and Comparison to Collaboration Environment. *WSEAS Transactions on Environment and development* 2, 1396–1404 (2006)
23. Pudhota, L., Chang, E.: Modelling the Dynamic Relationships between Workflow Components. In: *ICEISI Porto, Portugal* (2004)
24. Neumann, D.: *An Introduction to Web Objects*, <http://mactech.com/articles/mactech/Vol.13/13.05/WebObjectsOverview.2004>
25. Ulieru, M., Brennan, M., Scott, W., Robert, W.: *The Holonic enterprise: a model for Internet-enabled Global Manufacturing Supply Chain and workflow Management Canada* (2000)
26. Chang, E., Dillon, T., Hussain, F.: *Trust and Reputation for service-oriented Environments: Technology for Building Business Intelligence and Consumer Confidence*. John Wiley and Sons, West Sussex (2006)
27. Clark, M., Fletcher, P., et al.: *Web Services Business Strategies and Architectures*. Expert press (2002)
28. Skinner, G., Chang, E.: A Projection of the Future Effects of Quantum Computation on Information Privacy and Information Security. *International Journal of Computer Science and Network Security* 6, 166–172 (2006)
29. Liu, J., Wang, B., Liao, H., Shih, C., Kuo, T., Pang, A., Huang, C.: Component-based Medical and Assistive Devices and Systems. In: *Proceedings of High Confidence Medical and Systems (HCMDSS) Workshop, Philadelphia, PA* (2005)
30. Bach-Y-Rita, P., Kercel, S.: Sensory Substitution and the Human-Machine Interface. *Trends in Cognitive Sciences* 7, 541–546 (2003)
31. Merzenich, M., Jenkins, W.: *Memory Concepts*, pp. 437–453. Elsevier, Amsterdam (1993)

Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment

Urmila Shrawankar^{1,3} and Vilas Thakare²

¹ IEEE Student Member & Research Scholar, (CSE), SGB Amravati University, India

² Professor & Head, PG Dept. of Computer Science, SGB Amravati University, India

³ G H Raison College of Engg., Nagpur, India

urmilas@rediffmail.com

Abstract. Noise is ubiquitous in almost all acoustic environments. The speech signal, that is recorded by a microphone is generally infected by noise originating from various sources. Such contamination can change the characteristics of the speech signals and degrade the speech quality and intelligibility, thereby causing significant harm to human-to-machine communication systems.

Noise detection and reduction for speech applications is often formulated as a digital filtering problem, where the clean speech estimation is obtained by passing the noisy speech through a linear filter. With such a formulation, the core issue of noise reduction becomes how to design an optimal filter that can significantly suppress noise without noticeable speech distortion.

This paper focuses on voice activity detection, noise estimation, removal techniques and an optimal filter.

Keywords: Additive Noise, Noise detection, Noise removal, Noise filters, Voice Activity Detector (VAD).

1 Introduction

Noise estimation and reduction [6] is a very challenging problem. In addition, noise characteristics may vary in time. It is therefore very difficult to develop a versatile algorithm that works in diversified environments.

Although many different transforms are available, noise reduction [1] have been focused only on the Fourier, Karhunen–Loeve, cosine, Hadamard transforms. The advantage of the generalized transform domain is the different transforms can be used to replace each other without change the algorithm formulation. The following steps will help to use generalized transform domain; i. Reformulate the noise reduction problem into a more generalized transform domain, where any unitary matrix can be used to serve as a transform and ii. Design different optimal and suboptimal filters in the generalized transform domain.

The points to be considered in signal de-noising applications that are i. Eliminating noise from signal to improve the SNR and ii. Preserving the shape and characteristics of the original signal. An approach is discussed in this paper, to remove the additive noise [2] from corrupted speech signal to make speech front-ends immune to additive noise. We address two problems, i.e., noise estimation and noise removal.

2 Voice Activity Detector (VAD)

VADs are widely evaluated in terms of the ability to discriminate between speech and pause periods at different SNR levels of 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. These noisy signals have been recorded at different places. Detection performance as a function of the SNR [7] was assessed in terms of the non-speech hit-rate (HR0) and the speech hit-rate (HR1). Most of the VAD algorithms [4] fail when the noise level increases and the noise completely mask the speech signal. A VAD module is used in the speech recognition systems within the feature extraction process.

The different approaches of VAD include: Full-band and sub-band energies (Woo 2000), Spectrum divergence measures between speech and background noise (Marzinzik & Kollmeier 2002), Pitch estimation (Tucker 1992), Zero crossing rate (Rabiner 1975), and higher-order statistics (Nemer 2001; Ramirez 2006a; Gorriz., 2006a; Ramirez 2007).

Most of the VAD methods are based on the current observations and do not consider contextual information. However, using long-term speech information (Ramirez2004a; Ramirez 2005a) has shown improvement for detecting speech presence in high noise environment. Some robust VAD algorithms that yield high Speech/non-speech discrimination in noisy environments include i. Long-term spectral divergence; the speech/non-speech detection algorithm (Ramírez 2004a) ii. Multiple observation likelihood ratio tests; An improvement over the LRT (Sohn 1999 and Ramírez 2005b) and iii. Order statistics filters.

3 Noise Estimation Algorithms

A noise-estimation algorithm [14] is proposed for highly non-stationary noise environments. The performance of speech-enhancement algorithms as it is needed to evaluate, i. The Wiener algorithms (Lim & Oppenheim 1978), ii. Estimate the a priori SNR in the MMSE algorithms (Ephraim & Malah 1984) iii. Estimate the noise covariance matrix in the subspace algorithms (Ephraim & Van Trees 1993).

The noise estimation can have a major impact on the quality of the enhanced signal i.e. i. If the noise estimate is too low, annoying residual noise will be audible and ii. If the noise estimate is too high, speech will be distorted resulting possibly in eligibility loss. The simplest approach is to estimate and update the noise spectrum during the silent (pauses) segments of the signal using a voice-activity detection (VAD) [4]. An approach might work satisfactorily in stationary noise, it will not work well in more realistic environments where the spectral characteristics of the noise might be changing constantly. Hence there is a need to update the noise spectrum continuously over time and this can be done using noise-estimation algorithms.

Several noise-estimation algorithms are available like, Doblinger 1995; Hirsch & Ehrlicher 1995; Kim 1998; Malah 1999; Stahl 2000; Martin 2001; Ris & Dupont 2001 Afify & Sioham 2001; Cohen 2002; Yao & Nakamura 2002; Cohen 2003; Lin 2003; Deng 2003; Rangachari, 2004;

Noise estimation algorithms consider the following aspects: i. Update of the noise estimate without explicit voice activity decision, and ii. Estimate of speech-presence

probability exploiting the correlation of power spectral components in neighboring frames.

Noise-Estimation algorithm follows four steps; i. Tracking the minimum of noisy speech methods, ii. Checking speech-presence probability iii. Computing frequency-dependent smoothing constants and iv. Update of noise spectrum estimate

4 Noise Reduction Techniques

The noise is classify into following category like, adaptive, additive, additive random, airport, background, car, Cross-Noise, exhibition hall, factory, multi-talker babble, musical, Natural, non-stationary babble, office, quantile-based, restaurant, street, suburban train, ambient, random, train-station, white Gaussian etc. Noise is mainly dividing into four categories: Additive noise, Interference, Reverberation and Echo. These four types of noise has led to the developments of four broad classes of acoustic signal processing techniques include, Noise reduction/Speech enhancement, Source separation, speech dereverberation and Echo cancellation/Suppression. The scope of this paper limited to noise reduction techniques only. Noise reduction techniques depending on the domain of analyses like Time, Frequency or Time-Frequency/Time-Scale.

4.1 Noise Reduction Algorithms

The Noise reduction methods [13, 16] are classified into four classes of algorithms: Spectral Subtractive, Subspace, Statistical-model based and Wiener-type. Some popular Noise reduction algorithms are, The log minimum mean square error logMMSE (Ephraim & Malah 1985), The traditional Wiener (Scalart & Filho 1996), The spectral subtraction based on reduced-delay convolution (Gustafsson 2001), The exception of the logMMSE-SPU (Cohen & Berdugo 2002), The logMMSE with speech-presence uncertainty (Cohen & Berdugo 2002), The multiband spectral-subtractive (Kamath & Loizou 2002), The generalized subspace approach (Hu & Loizou 2003), The perceptually-based subspace approach (Jabloun & Champagne 2003), The Wiener filtering based on wavelet-thresholded multitaper spectra (Hu & Loizou 2004), Least-Mean-Square (LMS), Adaptive noise cancellation (ANC) [3], Normalized(N) LMS, Modified(M)-NLMS, Error nonlinearity (EN)-LMS, Normalized data nonlinearity (NDN)-LMS adaptation etc.

4.2 Fusion Techniques for Noise Reduction

4.2.1 The Fusion of Independent Component Analysis (ICA) and Wiener Filter

The fusion uses following steps: i. ICA [10] is applied to a large ensemble of clean speech training frames to reveal their underlying statistically independent basis ii. The distribution of the ICA transformed data is also estimated in the training part. It is required for computing the covariance matrix of the ICA transformed speech data used in the Wiener filter iii. Then a Wiener filter is applied to estimate the clean speech from the received noisy speech iv. The Wiener filter minimizes the mean-square error between the estimated signal and the clean speech signal in ICA domain

v. An inverse transformation from ICA domain back to time domain reconstructs the enhanced signal. vi. The evaluation is performed with respect to four objective quality measure criteria. The properties of the two techniques will yield higher noise suppression capability and lower distortion by combining them.

4.2.2 Recursive Least Squares (RLS) Algorithm: Fusion of DTW and HMM

Recursive Least Squares (RLS) algorithm is used to improve the presence of speech in a background noise [11]. Fusion pattern recognition is used such as with Dynamic Time Warping (DTW) and Hidden Markov Model (HMM). There are a few types of fusion in speech recognition amongst them are HMM and Artificial Neural Network (ANN) [10] and HMM and Bayesian Network (BN) [11]. The fusion technique can be used to fuse the pattern recognition outputs of DTW and HMM.

5 Experimental Steps for Implementing RLS Algorithm

- Recording speech, WAV file was recorded from different speakers
- RLS : The RLS [8] was used in preprocessing for noise cancellation
- End point detecting: two basic parameters are used: Zero Crossing Rate (ZCR) and short time energy [11].
- Framing, Normalization, Filtering
- MFCC : Mel Frequency Cepstral Coefficient (MFCC) is chosen as the feature extraction method.
- Weighting signal, Time normalization, Vector Quantization (VQ) and labeling.
- Then HMM is used to calculate the reference patterns and DTW is used to normalize the training data with the reference patterns
- Fusion HMM and DTW:
 - DTW measures the distance between recorded speech and a template.
 - Distance of the signals is computed at each instant along the warping function.
 - HMM trains cluster and iteratively moves between clusters based on their likelihoods given by the various models.

As a result, this algorithm performs almost perfect segmentation for recoded voice, recoding is done at noisy places, segmentation problem happens because in some cases the algorithm produces different values caused by background noise. This causes the cut off for silence to be raised as it may not be quite zero due to noise being interpreted as speech. On the other hand for clean speech both zero crossing rate and short term energy should be zero for silent regions.

6 Comparative Study of Various Speech Enhancement Algorithms

Total thirteen methods encompassing four classes of algorithms [17], that are, three spectral subtractive, Two subspace, Three Wiener-type and Five statistical-model based. The noise, consider at two levels of SNR (0 dB, 5 dB, 10 dB and 15 dB).

6.1 Intelligibility Comparison among Algorithms [16]

At 5 dB SNR: KLT and Wiener-as algorithms performed equally well in all conditions, followed by the logMMSE and MB algorithms. pKLT, RDC, logMMSE-SPU and WavThr algorithms performed poorly.

At 0 dB SNR: Wiener-as and logMMSE algorithms performed equally well in most conditions, followed by the MB and WavThr algorithms. The KLT algorithm performed poorly except in the babble condition in which it performed the best among all algorithms. Considering all conditions, the Wiener-as algorithm performed consistently well for all conditions, followed by the logMMSE algorithms which performed well in six of the eight noise conditions, followed by the KLT and MB algorithms which performed well in five conditions.

6.2 Intelligibility Comparison against Noisy Speech

The Wiener-as algorithm maintained speech intelligibility in six of the eight noise conditions tested, and improved intelligibility in 5 dB car noise. Good performance was followed by the KLT, logMMSE and MB algorithms which maintained speech intelligibility in six conditions. All algorithms produced a decrement in intelligibility in train noise at 0 dB SNR. The pKLT and RDC algorithms significantly reduced the intelligibility of speech in most conditions.

6.3 Consonant Intelligibility Comparison among Algorithms

pKLT and RDC, most algorithms performed equally well. A similar pattern was also observed at 0 dB SNR. The KLT, logMMSE, MB and Wiener-as algorithms performed equally well in most conditions. The logMMSESPU performed well in most conditions except in car noise. Overall, the Wiener-type algorithms Wiener-as and WavThr and the KLT algorithm performed consistently well in all conditions, followed by the logMMSE and MB algorithms. The RDC and pKLT algorithms performed poorly relative to the other algorithms.

6.4 The Following Algorithms Performed Equally Well across All Conditions

MMSE-SPU, logMMSE, logMMSE-ne, pMMSE and MB. The Wiener-as method also performed well in five of the eight conditions.

6.5 The Following Algorithms Performed the Best, in Terms of Yielding the Lowest Speech Distortion, across All Conditions

MMSE-SPU, logMMSE, logMMSE-ne, pMMSE, MB and Wiener-as. The KLT, RDC and WT algorithms also performed well in a few isolated conditions. The pKLT method also performed well in five of the eight conditions. The KLT, RDC, RDC-ne, Wiener-as and AudSup algorithms performed well in a few isolated conditions.

6.6 Comparisons in Reference to Noisy Speech

The algorithms MMSE-SPU, log-MMSE, logMMSE-ne, and pMMSE improved significantly the overall speech quality but only in a few isolated conditions. The algorithms MMSE-SPU, log-MMSE, logMMSE-ne, pMMSE, MB and Wiener-as performed the best in all conditions. The algorithms WT, RDC and KLT also performed well in a few isolated conditions. The algorithms MMSE-SPU, log-MMSE, logMMSE-ne, log-MMSE-SPU and pMMSE lowered significantly noise distortion for most conditions. The MB, pKLT and Aud-Sup also lowered noise distortion in a few conditions.

6.7 In Terms of Overall Quality and Speech Distortion, the Following Algorithms Performed the Best

MMSESPU, logMMSE, logMMSE-ne, pMMSE and MB. The Wiener-as method also performed well in some conditions. The subspace algorithms performed poorly.

7 Conclusion

The optimal filters can be designed either in the time or in a transform domain. The advantage of working in a transform space is that, if the transform is selected properly, the speech and noise signals may be better separated in that space, thereby enabling better filter estimation and noise reduction performance. The suppress noise from the speech signals without speech distortion it is an art of the noise removal approach. All filters do not give equal performance in every condition. Fusion techniques give better performance in noise reduction than the single noise removal approach. The discussion given in this paper will help for developing improved speech recognition system for noisy environment.

References

1. Zehtabian, A., Hassanpour, H.: A Non-destructive Approach for Noise Reduction in Time Domain. *World Applied Sciences Journal* 6(1), 53–63 (2009)
2. Chen, J.: Subtraction of Additive Noise from Corrupted Speech for Robust Speech Recognition (1998)
3. Górriz, J.M.: A Novel LMS Algorithm Applied to Adaptive Noise Cancellation (2009)
4. Ramírez, J.: Voice Activity Detection. *Fundamentals and Speech Recognition System Robustness* (2007)
5. Husoy, J.H.: Unified approach to adaptive filters and their performance (2008)
6. Benesty, J.: Noise Reduction Algorithms in a Generalized Transform Domain (2009)
7. Droppo, J., Acero, A.: Noise Robust Speech Recognition with a Switching Linear Dynamic Model (2004)
8. Deng, L.: Large-Vocabulary Speech Recognition Under Adverse Acoustic Environments (2000)
9. Deng, L.: High-Performance Robust Speech Recognition Using Stereo Training Data (2001)

10. Hong, L.: Independent Component Analysis Based Single Channel Speech Enhancement Using Wiener Filter (2003)
11. Rahman, S.A.: Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering (2009)
12. Sharath Rao, K.: Improved Iterative Wiener Filtering For Non-Stationary Noise Speech Enhancement (2004)
13. Hasan, T.: Suppression of Residual Noise From Speech Signals Using Empirical Mode Decomposition (2009)
14. Sundarajan: A Noise-Estimation Algorithm for Highly Non-Stationary Environments (2005)
15. Zhu, W.: Using Noise Reduction And Spectral Emphasis Techniques To Improve Asr Performanc In Noisy Conditions (2003)
16. Hu, Y., Loizou, P.C.: A Comparative Intelligibility Study of Single-Microphone Noise Reduction Algorithms (2007)
17. Hu, Y.: Subjective Comparison and Evaluation of Speech Enhancement Algorithms (2007)

Proximity User Identification Using Correlogram

Shervin Shahidi¹, Parisa Mazrooei¹, Navid Nasr Esfahani², and Mohammad Saraei³

¹ Intelligent Databases, Data Mining and Bioinformatics Research Laboratory
Department of Electrical and Computer Engineering
Isfahan University of Technology, Isfahan, Iran 84156-83111

² Isfahan Mathematics House, Isfahan, Iran 81645-356

³ School of Computing, Science and Engineering, University of Salford,
Greater Manchester, UK

sh.shahidi@ec.iut.ac.ir, p.mazrooei@ec.iut.ac.ir,
navid@ec.iut.ac.ir, m.saraee@salford.ac.uk

Abstract. This paper represents a technique, applying user action patterns in order to distinguish between users and identify them. In this method, users' actions sequences are mapped to numerical sequences and each user's profile is generated using autocorrelation values. Next, cross-correlation is used to compare user profiles with a test data. To evaluate our proposed method, a dataset known as Greenberg's dataset is used. The presented approach is succeeded to detect the correct user with as high as 82.3% accuracy over a set of 52 users. In comparison to the existing methods based on Hidden Markov Model or Neural Networks, our method needs less computation time and space. In addition, it has the ability of getting updated iteratively which is a main factor to facilitate transferability.

Keywords: User Identification- Correlation- Classification- Pattern Recognition.

1 Introduction

The task of distinguishing the current system's user from other users, i.e. User Identification, is one of the issues which have been studied in many fields such as Human Computer Interaction (HCI) and Computer Security. As an instance, User Identification has tight resemblance with intrusion detection, which is the key to solve two important problems in computer security: Detecting the presence of an intruder masquerading as a valid user (*anomaly detection*) and detecting the perpetration of abusive actions on the part of an otherwise innocuous user (*misuse detection*). (Lane *et al.*, 1997).

On one point of view, different approaches to User Identification problem are classified as below:

- Detection based on supervised learning; in which labelled data is needed to train the agent. These approaches build a model over the normal data and then check to see how well new data fits into that model. (Warrender *et al.*, 1999) As an instance, in network security field, a technique developed at SRI in the EMERALD system

(Javitz *et al.*, 1993) compares the distribution of test data to a previously made distribution over a known user data to indicate an intrusion. The problem that causes supervised learning not to be always applicable is that preparing the labelled training data is sometimes difficult and expensive. (Eskin *et al.*, 2002).

- Detection based on unsupervised learning; which uses unlabeled data for anomaly detection. The overall idea of these approaches is based on two assumptions. First, the number of normal users enormously outnumbers the number of intruders in a large data set. Second, the set of user actions are different from each other, in a way that each user's profile can be classified as a dense region in a proper feature space. The feature space is an n -dimensional metric space, to which user actions are mapped. In intrusion detection systems, this approach is used as identifying sparse regions in a feature space and using proper metric, assuming that normal actions will tend to appear as dense regions in the same feature space, as in (Knorr *et al.*, 1998), (Knorr *et al.*, 1999), (Breunig *et al.*, 2000) and (Portnoy *et al.*, 2001). Although this assumption seems to be sensible, this is not always the case. For example, for a network being attacked too many times in the same way (e.g. a network under DOS attack), the anomaly patterns are so dense that cannot be distinguished as sparse regions. The same situation holds for user identification; the assumption of neatly classifiable regions for different users is not always correct.

The problem of user identification can be solved using different Machine Learning techniques; some learners try to generate a model consisting of a group of rules based on different attributes of user action history. Any new set of actions will be compared to the models to find out deviations from each user's profile. For instance, a rule learning program, named RIPPER is used to achieve such model in (Lee *et al.*, 1998). Some probabilistic methods have also been used to make a predictive model of user's next actions, and then, computing how well new data fits into the model (Eskin *et al.*, 2001). Markov Chains and Hidden Markov Models are also two widely used methods in this field. (Nong Ye, 2000)(Galassi *et al.*, 2005).

In this paper, we have considered the set of user actions as samples of signals and have used some signal processing techniques to identify the user. This approach is considered as a supervised learning approach, since we build user profiles over labelled training data and then, compare the profile patterns to test data patterns for getting results.

The training data is the stream of commands typed by each user in Unix *csh.*, saved as offline log files. We place our concentration on the importance of action sequences rather than the attributes calculated by the set of actions (e.g. user activity rate, mistake rate or word usage distribution). Based on the hypothesis that humans' set of actions are causal, autocorrelation and cross-correlation are used for a moving window on the sequence of actions to quantify the amount of serially performed actions. Finally, we compare the test sequence with previously built profiles using Mean Square Error (MSE) criterion to find the best profile matching with the test data.

The organization of the rest of this paper is as follows. In section 2 we introduce our general scheme along with some basic aspects of our model. Section 3 represents empirical results. Analysis of the results is provided in section 4. Finally, Conclusion and future works are presented in section 5.

2 General Scheme

2.1 Problem Model

The human-computer interaction is essentially a causal process. (Lane et. al., 1997) Therefore, some specific patterns can be found in a user's action sequence. Relying on this fact and assuming that these series of patterns vary between different users, we extract the repeated patterns in an action sequence and use these patterns for user identification.

For this purpose, after mapping the string of actions to a string of numerical codes, we define a moving window which takes a subsequence of a user action history each time, to be used as a unit for modelling and processing user behaviour. The window size determines the maximum distance of two related commands. Since user action patterns which follow a specific goal are not usually very long, it is reasonable not to consider the relevance of all of the commands in the sequence. To consider the dependence of sequential commands at the end of one window and the beginning of its successor, windows can overlap.

As a measure of self similarity, we calculate autocorrelation for each window. In other words, autocorrelation of a window and its cross-correlation with another window will be a scale to compare it with the other window. In extreme case, if the two windows contain the same sequence of commands, the autocorrelation of one of them is the same as their cross-correlation. Plotting the autocorrelation Coefficients against lag (τ) gives a Correlogram indicating different autocorrelation values for each lag value τ . Technically, Autocorrelation Coefficients represent the correlation of a sequence's elements with certain distance apart.

Correlation coefficient for lag value τ and time t can be computed as:

$$R_{XY}(\tau) = \frac{E[(X_t - \mu_X)(Y_{t+\tau} - \mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

Where E indicates the Expected value operator, μ is the mean value and σ is the standard deviation for X and Y random variables. The value R_{XX} or simply R_X is called auto-correlation coefficient. In signal processing, the above definition is often used without the normalization, that is, without subtracting the mean and dividing by the variance. In this work, we did the same, since comparisons are made between different Correlograms, and normalization decreases the distance between different user profiles which will lead to less classification accuracy. Therefore, the correlation for two arbitrary windows is calculated via the following formula:

$$R_{XY}(\tau) = E[(X_t)(Y_{t+\tau})] = \begin{cases} \sum_{t=0}^{N-\tau-1} X_t Y_{t+\tau} & \tau \geq 0 \\ R(-\tau) & \tau < 0 \end{cases} \tag{2}$$

Note that since the data is inherently non-numerical, only similarity of two elements should be taken into account. Hence, the multiplication between two elements of X and Y is defined as:

$$X_i Y_j = \begin{cases} 1 & \text{if } X_i = Y_j \\ 0 & \text{if } X_i \neq Y_j \end{cases} \quad (3)$$

Calculating the correlograms for each window in a user action sequence, we can store each user's profile, which consists of user action history and its correlograms. To be able to compare a user's profile with a test data -which is in the form of a user action sequence broken into overlapping windows-, we calculated the cross-correlation between each window in the test data sequence and every window in a user profile. These correlation values will be used for classification.

2.2 Classification

To compare cross-correlations obtained from a test data with users' profiles autocorrelations, Mean Squared Error (MSE) is used as a measure of distortion:

$$D = E[\|R_X - R_{XY}\|^2] = \frac{1}{2N-1} \sum_{\tau=-N}^{\tau=N} (R_X(\tau) - R_{XY}(\tau))^2 \quad (4)$$

where N is the length of the windows.

The window in profile which has minimum distortion D with the window in test data is chosen as the matching window and the distortion is saved as a penalizing factor. Average distortion of the test data from user profile is calculated over all of the matching windows' distortions. This job is done for all of the user profiles and the average distortion is considered as a discriminant function for classification; That is, the profile with the minimum average distortion is considered as the resulting profile.

Instead of taking average over all distortions, another method was used for classification by means of confusion matrix. Confusion matrix is a well-known visualization method in supervised learning, in which each row of the matrix represents the correct class and the columns reveal the predicted classes according to the classifier. Note that confusion matrix is used when there are several test data's available, but here we have modified it to be used for classification over one test data. In this method, each window of the test data is classified independently similar to the above method and the resulting class for each window will be rewarded one point. Finally, the class with maximum number of points will be considered as the resulting class. Figure 1. reveals a confusion matrix for some test data.

3 Experimental Results

Method represented in this paper has been tested by a data set of user actions collected by Greenberg (Greenberg S. , 1988). This data set includes actions of 168 people in four groups of Computer Scientists, Expert Programmers, Novice Programmers and Non Programmers, typing in UNIX `csh`. These logs contain quite complete information about user actions such as commands, start and end times login sessions, working directory, etc.. For our approach, only user commands have been

used. In order to form each user action history, all commands used by each group were coded into integer values separately. Table 1. gives information about each group's number of users and number of different commands used. In order to consider occurrence of wrong commands, they are coded as another occurrence of the first subsequent correct command; in this way, typing wrong commands will be shown as multiple occurrences of a command.

Table 1. Data set information

	Number of users	Different commands
Non-programmers	25	196
Computer scientist	52	851
Experienced programmers	36	588
Novice programmers	55	264
Overall	168	1307

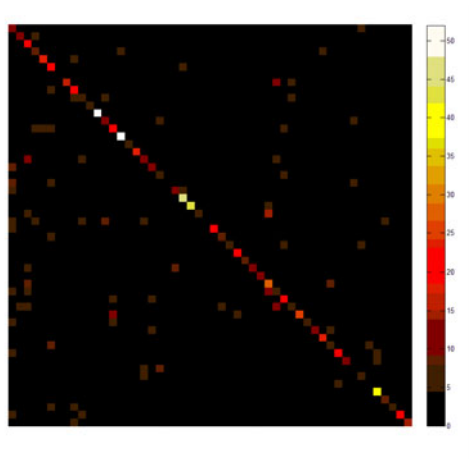


Fig. 1. Confusion matrix for Computer Scientists

With regards to the fact that human actions are not stationary, using the whole user log would decrease the accuracy of identification. In other words, since users' behavior change through the time, their actions should be compared with more recent logs. To examine this ability in our method only last part of user logs have been considered. This last part has been defined to be of the size Maximum Log Length (MLL), so that the preceding actions in the log will be neglected. Using last *MLL* commands of the sequence in the algorithm has also the advantage of less computation and storage requirements.

In order to check the algorithm's efficiency we examined it for different values of variables, Window Size (W), Window Shift (W_s)-which indicates the amount of window overlaps- and Maximum Log Length (MLL). The abovementioned method has been run for all possible combinations of $W= 10, 20, 40$, $W_s=W/2, W/4$ and $MML= 1000, 2000, 4000$. To avoid accidental results, 5-fold cross-validation is used, in which user logs are partitioned to five subsets and each subset is considered as test data, while the remaining is considered to be training data. The results of using average distortion and confusion matrix for classification are shown in Table 2. and Table 3. respectively.

Table 2. Classification using Average distortion

	W	W_s	MLL	Accuracy
Exp	10	3	2000	73.3%
Non	10	3	2000	59.2%
Novice	10	3	2000	45.8%
Scientist	10	3	1000	82.3%
Overall	10	3	2000	61.7%

Table 3. Classification using confusion matrix

	W	W_s	MLL	Accuracy
Exp	10	3	2000	71.6%
Non	10	3	2000	56%
Novice	20	5	2000	33.1%
Scientist	10	3	2000	80%
Overall	10	3	2000	54.3%

To test the method, we used each one of the four groups as a separate data set, so that the algorithm should recognize a member of each group in its own group. This gives us the opportunity to test it four times. Since these groups noticeably vary, we assumed that if the agent is able to find a computer scientist in a group of computer scientists (for instance), it would obviously be able to find it among all 168 users. To justify our assumption, we have also tried our method over all users to classify them to either correct user and correct group (Table 2., 3. last rows and table 4.). As can be seen, the method can classify the user in his right group with more than 90% average accuracy, in spite of no specific training for group classification and without any information about the groups of users.

Table 4. Classification of users to their corresponding groups

Predicted class \ Target class	Exp%	Non%	Novice%	Scientist%
Exp	91.1	0.5	1.7	6.7
Non	4.8	86.4	1.6	7.2
Novice	1.8	0	97.8	0.4
Scientist	2.3	0.8	1.1	95.8

4 Analysis

To gain a deeper understanding of used methods, a brief discussion will be provided in this section.

4.1 Analysis of the Classification Results

According to Table 2. and Table 3. some interesting results can be deduced. As can be seen, there is a huge gap between the classification accuracy of the groups "experienced programmers" and "computer scientists" with the groups "non programmers" and "novice programmers". We believe this difference is due to the fact that the two former mentioned groups use a wider variety of commands, use the system for a bigger set of goals and do more complicated tasks; hence, there exists a bigger implicit difference between the users' actions in these groups. On the other hand, the two latter mentioned groups use a smaller set of commands to achieve simpler and fewer tasks, which can be very common with their group-mates.

4.2 Maximum Log Length of Command Sequences

As previously mentioned, using MLL limit on commands sequence in our method, provides a better basis for the algorithm to distinguish between users. As an evidence, the optimum size for MLL never exceeded 2000, while the method was also tested for size 4000. Another advantage of putting this limit is to decrease the required time and space for generating and storing profiles. Furthermore, this method allows us to update the profiles easily by calculating only the autocorrelations for the updated part of log files, while by removing the oldest autocorrelations we keep the complexity of classification invariant. This is an advantage over a lot of proposed algorithms which need to generate the whole profile over, each time the log files are updated. Since calculating an autocorrelation is not a process of high time complexity, the aforementioned procedure can be done as an online process. Enhancing the system by this attribute would promote the system to an online supervised model, which can be used in real-time applications.

4.3 Confusion Matrix

As mentioned briefly before, a confusion matrix consists of some rows indicating the target classes and the same number of columns (usually) for the predicted classes. Putting the corresponding rows and columns of target and predicted classes in the same order, each correct classification will appear on the diagonal of the matrix. Therefore in general, a perfect classification should result a diagonal matrix. Here, we used a modified version of confusion matrix, in which instead of classifying over each test data, we have done the classification on the consisting windows of the test data to form the confusion matrix. As a result of this modification, we have to take all of the classification results into account at the same time. Hence, for a perfect classifier for this modified matrix it is only needed to have the maximum value of each row on the diagonal. As it is shown in Figure 1. the confusion matrix for computer scientists has this property in most of the rows, although the proposed method did not work as good as the average distortion method for classification. (Tables 2. and 3.).

5 Conclusion

We introduced a new supervised learning algorithm based on using correlograms. Mapping users' history of commands to numerical values, autocorrelation is used to generate each user's profile. Afterwards, cross-correlation is used to compare a test data with user profiles. Based on Mean Squared Error, two measures of similarity were applied to find the matching profile.

We have presented experimental results supporting the applicability of our method on Greenberg data set. While this method preserves satisfactory accuracy (up to 82.3% accuracy within 52 different classes), it needs less computation time and the ability of being updated easily in comparison with other methods. In this paper, we focused only on taking advantage of the command patterns used by a user for identification. Extracting other information out of user action history can be the subject of the future works.

Acknowledgement

The authors would like to thank Mr. Ali Bakhshali for his kind helps. Also to thanks Dr. S. Greenberg who shared his dataset.

References

- Skin, E., Lee, W., Stolfo, S.: Modeling system calls for intrusion detection with dynamic window sizes. In: DARPA Information Survivability Conference and Exposition II (DISCEX II), Anaheim, vol. 1, pp. 165–175 (2001)
- Ye, N.: A markov chain model of temporal behavior for anomaly detection. In: The 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop, pp. 171–174 (2000)

- Lane, T., Brodley, C.: Sequence matching and learning in Anomaly Detection for computer security. In: The Fourth National Conference on Artificial Intelligence, pp. 43–49 (1997)
- Warrender, C., Forrest, S., Pearlmutter, B.: Detecting intrusions using system calls: alternative data models. In: 1999 IEEE Symposium on Security and Privacy, pp. 133–145 (1999)
- Javitz, H., Valdes, A.: The NIDES statistical component: description and justification. Computer Science Laboratory, SRI International, Tech. Report (1993)
- Knorr, E., Ng, R.: Algorithms for mining distance-based outliers in large data sets. In: 24th Int. Conf. Very Large Data Bases, VLDB, Technique et Science Informatiques, pp. 392–403 (1998)
- Knorr, E., Ng, R.: Finding international knowledge of distance-based outliers. The VLDB Journal, 211–222 (1999)
- Breunig, V., Kriegel, H., Ng, R., Sander, J.: LOF: identifying density-based local outliers. In: ACM SIGMOD Int. Conf. on Management of Data, pp. 93–104 (2000)
- Portnoy, L., Eskin, E., Stolfo, S.: Intrusion detection with unlabeled data using clustering. In: ACM CSS Workshop on Data Mining Applied to Security (DMSA 2001), Philadelphia, PA (2001)
- Galassi, U., Giordana, A., Saitta, L., Botta, M.: Learning Profiles Based on Hierarchical Hidden Markov Model. In: Hacid, M.-S., Murray, N.V., Raš, Z.W., Tsumoto, S. (eds.) ISMIS 2005. LNCS (LNAI), vol. 3488, pp. 47–55. Springer, Heidelberg (2005)
- Link, H., Lane, T., Magliano, J.: Models and Model Biases for Automatically Learning Task Switching Behavior. In: Proceedings of the 2005 HCI International (HCII) Conference on Augmented Cognition, HCI International, HCII (2005)
- Lee, W., Stolfo, S.: Data mining approaches for intrusion detection. In: 1998 USENIX Security Symposium (1998)
- Greenberg, S.: Using Unix: collected traces of 168 users. In: Research Report 88/333/45, Includes Tar-Format Cartridge Tape. Department of Computer Science, University of Calgary, Alberta (1998)

Erratum: An Efficient Data Indexing Approach on Hadoop Using Java Persistence API

Yang Lai^{1,2} and Shi Zhong Zhi¹

¹ The Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

² Graduate University of Chinese Academy of Sciences, Beijing 100039, China
{yanglai, shizz}@ics.ict.ac.cn

Z. Shi et al. (Eds.): IIP 2010, IFIP AICT 340, pp. 213–224, 2010.
© IFIP International Federation for Information Processing 2010

DOI 10.1007/978-3-642-16327-2_42

In the original version, the first and last names of the authors are interchanged. They should read “Lai Yang, and ZhongZhi Shi”.

The original online version for this chapter can be found at
http://dx.doi.org/10.1007/978-3-642-16327-2_27

Author Index

- Aamodt, Agnar 82, 102
Abe, Hidenao 130, 139

Bruland, Tore 82

Cai, Guang-Jun 16, 35
Calder, David 326
Cannas, Laura Maria 297
Cao, Cungen 55, 64, 154,
271, 291
Cao, Yanan 154
Castellani, Stefania 92
Chang, Liang 45
Chen, Limin 6
Christley, Rob 163
Chua, Stephanie 193
Chung, Paul W.H. 245
Coenen, Frans 163, 193

Dai, Wei 245
Dessì, Nicoletta 297
Dou, Quansheng 26, 112, 122
Duan, Zhenhua 72

Esfahani, Navid Nasr 343

Fatima, Shaheen 245
Fu, Chuan 203
Fu, Kailei 112

Gao, Hualing 234
Gavrilova, Tatiana 225
Gu, Tianlong 45
Guang, Jiang 271
Guo, LiJun 308

Han, Lu 271
Han, Xu 26
Hendam, Ashraf 281
Hirano, Shoji 139

Jiang, Ping 112, 122
Jin, Zhi 35

Lagos, Nikolaos 92
Langseth, Helge 82

Leake, David 1
Li, Peng 183
Li, Zhixin 259
Liao, Zhuhua 203
Lin, Shouxun 316
Liu, Shasha 122
Liu, Yizhi 316

Ma, Huifang 259
Ma, Yue 55, 64
Mach, Maria A. 251
Malcolm, Grant 193
Mazrooei, Parisa 343
Muggleton, Stephen 2

Niu, Wenjia 26
Nohuddin, Puteri N.E. 163

O'Neill, Jacki 92
Owoc, Mieczyslaw L. 251

Pes, Barbara 297

Qin, Ping 234

Rafea, Ahmed 281

Saleem, Mohammad 245
Saraee, Mohammad 343
Segond, Frederique 92
Setzkorn, Christian 163
Shaalán, Khaled 281
Shahidi, Shervin 343
Shawe-Taylor, John 3
Shen, Yuming 55, 64
Shi, Zhiwei 183
Shi, Zhongzhi 4, 6, 26, 45, 112,
122, 173, 183, 213, 259, E1
Shrawankar, Urmila 336
Skalle, Pål 102
Song, Xiaoyu 72
Sui, Yuefei 55, 64, 271

- Tang, Sheng 316
Techataweewan, Wawta 145
Thakare, Vilas 336
Tian, Cong 72
Tsumoto, Shusaku 130, 139
- Valipour Shokouhi, Samad 102
- Wang, Bin 183
Wang, Dongsheng 154, 291
Wang, Ju 55, 234
Wang, Shi 154, 271, 291
Wu, Yuming 291
- Yang, Jing 203
Yang, Lai 213, E1
Yang, Xinghua 26
- Zang, Liangjun 154
Zhang, Guoqing 203
Zhang, Rong 308
Zhang, Yongdong 316
Zhang, Zhongxiang 234
Zhao, Bin 35
Zhao, JieYu 308
Zhou, Jing 173
Zhou, Xiuhua 122
Zhu, Haiyan 112