

Deborah Richards
Byeong Ho Kang (Eds.)

LNAI 7457

Knowledge Management and Acquisition for Intelligent Systems

12th Pacific Rim Knowledge Acquisition Workshop, PKAW 2012
Kuching, Malaysia, September 2012
Proceedings

 Springer

Lecture Notes in Artificial Intelligence 7457

Subseries of Lecture Notes in Computer Science

LNAI Series Editors

Randy Goebel

University of Alberta, Edmonton, Canada

Yuzuru Tanaka

Hokkaido University, Sapporo, Japan

Wolfgang Wahlster

DFKI and Saarland University, Saarbrücken, Germany

LNAI Founding Series Editor

Joerg Siekmann

DFKI and Saarland University, Saarbrücken, Germany

Deborah Richards Byeong Ho Kang (Eds.)

Knowledge Management and Acquisition for Intelligent Systems

12th Pacific Rim
Knowledge Acquisition Workshop, PKAW 2012
Kuching, Malaysia, September 5-6, 2012
Proceedings



Springer

Series Editors

Randy Goebel, University of Alberta, Edmonton, Canada
Jörg Siekmann, University of Saarland, Saarbrücken, Germany
Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

Volume Editors

Deborah Richards
Macquarie University
Department of Computing
North Ryde, NSW, 2109, Australia
E-mail: deborah.richards@mq.edu.au

Byeong Ho Kang
University of Tasmania
School of Computing and Information Systems
Hobart, Tasmania, 7000, Australia
E-mail: bhkang@utas.edu.au

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-32540-3 e-ISBN 978-3-642-32541-0
DOI 10.1007/978-3-642-32541-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012943696

CR Subject Classification (1998): I.2.4, I.2.6-7, I.2.10-11, H.5, H.3, H.4, J.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

© Springer-Verlag Berlin Heidelberg 2012

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.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at PKAW2012: The 12th International Workshop on Knowledge Management and Acquisition for Intelligent Systems held during September 5–6, 2012, in Kuching, Malaysia, in conjunction with the 12th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2012).

The purpose of this workshop is to provide a forum for presentation and discussion of all aspects of knowledge acquisition from both the theoretician's and practitioner's points of view. While it is well accepted that knowledge is vital for our individual, organizational, and societal survival and growth, the nature of knowledge and how it can be captured, represented, reused, maintained, and shared are not fully understood. This workshop explores approaches that address these issues. PKAW includes knowledge acquisition research involving manual and automated methods and combinations of both.

A total of 141 papers were considered. Each paper was reviewed by at least two reviewers, of which 18% were accepted as Full Papers and 8% as Short Papers. Papers have been revised according to the reviewers' comments. As a result, this volume includes 21 Full Papers and 11 Short Papers.

PKAW2012 was the evolution of over two decades of knowledge acquisition (KA) research in the Pacific region that continues to draw together a community of researchers and practitioners working in the area of intelligent systems. As can be seen from the themes in the proceedings, this evolution reflects changes in what technology can do and how it is being used and the issues facing researchers. Following trends over the past decade, we continue to see a predominance of approaches and applications involving Web technologies, with a noticeable increase in research related to Web 2.0 and social networking. As a feature of PKAW, the workshop continues to include research on incremental knowledge acquisition methods as well as ontology and agent-based approaches. We also note a large number of papers seeking to address specific issues in KA and evaluation of KA methods.

The Workshop Co-chairs would like to thank all those who contributed to PKAW 2012, including the PKAW Program Committee and other reviewers for their support and timely review of papers and the PRICAI Organizing Committee for handling all of the administrative and local matters. We wish to thank the Air Force Office of Scientific Research, Asian Office of Aerospace Research and Development for their contribution to the success of this conference. Thanks to EasyChair for streamlining the whole process of producing this volume. Particular thanks goes to those who submitted papers, presented, and attended the workshop. We hope to see you again in 2014.

June 2012

Deborah Richards
Byeong Ho Kang

Organization

Program Committee

Ghassan Beydoun	University of Wollongong, Australia
Ivan Bindoff	University of Tasmania, Australia
Paul Compton	The University of New South Wales, Australia
Richard Dazeley	University of Ballarat, Australia
Nader Hanna	Macquarie University, Australia
Achim Hoffmann	University of New South Wales, Sydney, Australia
Byeong Ho Kang	University of Tasmania, Australia
Mihye Kim	Catholic University of Daegu, South Korea
Yang Sok Kim	University of New South Wales, Australia
Masahiro Kimura	Ryukoku University, Japan
Maria Lee	Shih Chien University, Taiwan
Kyongho Min	Independent Researcher
Toshiro Minami	Kyushu Institute of Information Sciences and Kyushu University Library, Japan
Hiroshi Motota	Osaka University and AFOSR/AOARD, Japan
Kozo Ohara	Aoyama Gakuin University, Japan
Frank Puppe	University of Würzburg, Germany
Ulrich Reimer	University of Applied Sciences St. Gallen, Switzerland
Deborah Richards	Macquarie University, Australia
Kazumi Saito	University of Shizuoka, Japan
Hendra Suryanto	UNSW, Australia
Takao Terano	Tokyo Institute of Technology, Japan
Shuxiang Xu	University of Tasmania, Australia
Seiji Yamada	National Institute of Informatics, Japan

Additional Reviewers

Busch, Peter
Halabi, Ana
Othman, Siti Hajar
Schwitter, Rolf

Sponsoring Organizations



Ministry of Science, Technology and
Innovation, Malaysia



MIMOS Berhad, Malaysia



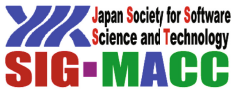
Sarawak Convention Bureau, Malaysia



Leibniz Information Centre for Economics,
Germany



Japanese Society for Artificial Intelligence,
Japan



SIG-MACC, Japan Society for Software Science
and Technology, Japan



AGRO-KNOW Technologies, Greece



agINFRA: A Data Infrastructure for
Agriculture



Franz Inc., USA



Air Force Office of Scientific Research, USA



Asian Office of Aerospace Research and Development, Japan



Quintiq Sdn. Bhd., Malaysia



Centre for Agricultural Bioscience International, United Kingdom



NOTA ASIA Sdn. Bhd., Malaysia



University of Tasmania, Australia

Supporting Institutions

Universiti Tenaga Nasional, Malaysia
 Swinburne University of Technology, Malaysia
 The University of Nottingham, Malaysia
 Monash University, Malaysia
 Multimedia University, Malaysia
 Sunway University, Malaysia
 University Tunku Abdul Rahman, Malaysia
 Universiti Sains Malaysia, Malaysia
 Universiti Malaya, Malaysia
 The British University in Dubai, United Arab Emirates
 The University of Tulsa, USA
 The University of New South Wales, Australia

Nanyang Technological University, Singapore

The University of Waikato, New Zealand

Macquarie University, Australia

Waseda University, Japan

National ICT Australia Ltd., Australia

Universiti Malaysia Sarawak, Malaysia

Universiti Kebangsaan Malaysia, Malaysia

Universiti Malaysia Sabah, Malaysia

Table of Contents

KA Issues and Evaluation

Constraints Dependent T-Way Test Suite Generation Using Harmony Search Strategy	1
<i>AbdulRahman A. Al-Sewari and Kamal Z. Zamli</i>	
Evaluating Disaster Management Knowledge Model by Using a Frequency-Based Selection Technique	12
<i>Siti Hajar Othman and Ghassan Beydoun</i>	
Perceptual Evaluation of Automatic 2.5D Cartoon Modelling	28
<i>Fengqi An, Xiongcai Cai, and Arcot Soumya</i>	
Automatic Acquisition of User Models of Interaction to Evaluate the Usability of Virtual Environments	43
<i>Nader Hanna, Deborah Richards, and Michael J. Jacobson</i>	
User-Centric Recommendation-Based Approximate Information Retrieval from Marine Sensor Data	58
<i>Zhao Chen, Md. Sumon Shahriar, and Byeong Ho Kang</i>	
Addressing Challenges for Knowledge Discovery from Data in the Domain of Seaport Integration	73
<i>Ana Ximena Halabi Echeverry and Deborah Richards</i>	
Data Envelopment Analysis for Evaluating Knowledge Acquisition and Creation	86
<i>Chuen Tse Kuah and Kuan Yew Wong</i>	

Language, Text and Image Processing

A High-Order Hidden Markov Model for Emotion Detection from Textual Data	94
<i>Dung T. Ho and Tru H. Cao</i>	
A Lazy Man's Way to Part-of-Speech Tagging	106
<i>Norshuhani Zamin, Alan Oxley, Zainab Abu Bakar, and Syed Ahmad Farhan</i>	
Knowledge Acquisition for Categorization of Legal Case Reports	118
<i>Filippo Galgani, Paul Compton, and Achim Hoffmann</i>	
Extraction of How-to Type Question-Answering Sentences Using Query Sets	133
<i>Kyohei Ishikawa and Hayato Ohwada</i>	

Image Indexing and Retrieval with Pachinko Allocation Model:
 Application on Local and Global Features 140
Ahmed Boulemden and Yamina Tlili

Incremental Knowledge Acquisition

Detection of CAN by Ensemble Classifiers Based on Ripple Down
 Rules 147
*Andrei Kelarev, Richard Dazeley, Andrew Stranieri,
 John Yearwood, and Herbert Jelinek*

Improving Open Information Extraction for Informal Web Documents
 with Ripple-Down Rules 160
Myung Hee Kim and Paul Compton

Ripple-Down Rules with Censored Production Rules 175
Yang Sok Kim, Paul Compton, and Byeong Ho Kang

RM and RDM, a Preliminary Evaluation of Two Prudent RDR
 Techniques 188
Omaru Maruatona, Peter Vamplew, and Richard Dazeley

Agent Based Knowledge Acquisition and Management

Planning Children’s Stories Using Agent Models 195
Karen Ang and Ethel Ong

A Framework for a Multi-agent Collaborative Virtual Learning
 Environment (MACVILLE) Based on Activity Theory 209
Nader Hanna and Deborah Richards

Emergence of Personal Knowledge Management Processes within
 Multi-agent Roles 221
Shahrinaz Ismail and Mohd Sharifuddin Ahmad

Ontology-Based Approaches

Towards an Ontology-Based Approach to Knowledge Management of
 Graduate Attributes in Higher Education 229
Amara Atif, Peter Busch, and Deborah Richards

Commonsense Knowledge Acquisition through Children’s Stories 244
Roland Christian Chua Jr. and Ethel Ong

Externalizing Senses of Worth in Medical Service Based on Ontological
 Engineering 251
Taisuke Ogawa, Mitsuru Ikeda, Muneou Suzuki, and Kenji Araki

Web 2.0 Methods and Applications

Crowd-Sourced Knowledge Bases	258
<i>Yang Sok Kim, Byeong Ho Kang, Seung Hwan Ryu, Paul Compton, Soyeon Caren Han, and Tim Menzies</i>	
Social Issue Gives You an Opportunity: Discovering the Personalised Relevance of Social Issues	272
<i>Soyeon Caren Han and Hyunsuk Chung</i>	
Identifying Important Factors for Future Contribution of Wikipedia Editors	285
<i>Yutaka Yoshida and Hayato Ohwada</i>	
Network Analysis of Three Twitter Functions: Favorite, Follow and Mention	298
<i>Shoko Kato, Akihiro Koide, Takayasu Fushimi, Kazumi Saito, and Hiroshi Motoda</i>	
User-Oriented Product Search Based on Consumer Values and Lifestyles	313
<i>Hesam Ziaei, Wayne Wobcke, and Anna Wong</i>	
Extracting Communities in Networks Based on Functional Properties of Nodes	328
<i>Takayasu Fushimi, Kazumi Saito, and Kazuhiro Kazama</i>	
Revealing and Trending the Social Knowledge Studies	335
<i>Maria R. Lee and Tsung Teng Chen</i>	
Workflow Knowledge Sharing through Social Networks	343
<i>Peter Busch and Amireh Amirmazaheri</i>	

Other Applications

Identifying Characteristics of Seaports for Environmental Benchmarks Based on Meta-learning	350
<i>Ana Ximena Halabi Echeverry, Deborah Richards, and Ayse Bilgin</i>	
A Situated Experiential Learning System Based on a Real-Time 3D Virtual Studio	364
<i>Mihye Kim, Ji-Seong Jeong, Chan Park, Rae-Hyun Jang, and Kwan-Hee Yoo</i>	
Author Index	373

Constraints Dependent T-Way Test Suite Generation Using Harmony Search Strategy

AbdulRahman A. Al-Sewari and Kamal Z. Zamli*

Software Engineering Research Group,
School of Electrical and Electronic Engineering, Universiti Sains Malaysia,
14300 Nibong Tebal, Penang, Malaysia

Abstract. Recently, many new researchers have considered the adoption of Artificial Intelligence-based Algorithm for the construction of t-way test suite generation strategies (where t indicates the interaction strengths). Although useful, most existing AI-based strategies have not sufficiently dealt or even experimented with the problem of constraints. Here, it is desirable for a particular AI-based strategy of interest to be able to automatically exclude the set of impossible or forbidden combinations from the final t-way generated suite. This paper describes our experience dealing with constraints from within a Harmony Search Algorithm based strategy, called HSS. Our experience with HSS is encouraging as we have obtained competitive test size as overall.

Keywords: T-Way Testing, Constraints Support, Software and Hardware Testing, Artificial Intelligent algorithms.

1 Introduction

In the last 50 years, many new and useful techniques have been developed in the field of software testing for preventing bugs and for facilitating bug detection. Even with all these useful techniques and good practices are in place, there is no guarantee that the developed software is bug free. Here, only testing can demonstrate that quality has been achieved and identify the problems and the risks that remain.

Although desirable, exhaustive testing is often infeasible due to resource and timing constraints. To systematically minimize the test cases into manageable one, a new sampling technique, called t-way strategies (where t indicates the interactions strength) has started to appear (e.g. AETG [1, 2], GTWay [3], IPOG families [4], PSTG [5-7], ITTDG [8], Aura [9], and Density [10-12]). In an effort to find the most efficient strategies capable of generating the most optimal test cases for every configuration (i.e., each combination is covered at most once), many new t-way strategies has been developed in the last 15 years.

Recently, many new researchers have considered the adoption of Artificial Intelligence-based Algorithm for the construction of t-way test suite generation strategies

* Corresponding author.

(where t indicates the interaction strengths). Although useful, most existing AI-based strategies have not sufficiently dealt or even experimented with the problem of constraints. Here, it is desirable for a particular AI-based strategy of interest to be able to automatically exclude the set of impossible or forbidden combinations from the final t -way generated suite. This paper describes our experience dealing with constraints from within a Harmony Search Algorithm based strategy, called HSS. Our experience with HSS is encouraging as we have obtained competitive test size overall.

This rest of the paper is organized as follows. Section 2 illustrates the problem domain model. Section 3 introduces the HSS strategy. Section 4 elaborates the benchmarking of HSS against TestCover. Finally, section 5 provides the conclusion.

2 Problem Domain Model

To illustrate the concept of constraints within t -way testing, consider an example of a pizza online ordering system as illustrated in Fig. 1 [3].



Fig. 1. Pizza Online Ordering System

In this system, there are 3 parameters for the user to choose from: the crust, flavour and toppings. For each of the parameters, there are 2 selections (or values) available (see Table 1).

Table 1. Pizza Online Ordering System Configuration

Configurations	Pizza Online Ordering System (parameters)		
	Crust	Flavor	Topping
	Classic Hand Tossed	Vegetarian	Pineapples
Crunchy Thin	Pepperoni Delight	Beef	

Table 2. Exhaustive Test Cases for Pizza Online Ordering System

Test Case ID	Crust	Flavor	Topping
1	Classic Hand Tossed	Vegetarian	Pineapples
2	Classic Hand Tossed	Vegetarian	Beef
3	Classic Hand Tossed	Pepperoni Delight	Pineapples

Table 2. (Continued)

4	Classic Hand Tossed	Pepperoni Delight	Beef
5	Crunchy Thin	Vegetarian	Pineapples
6	Crunchy Thin	Vegetarian	Beef
7	Crunchy Thin	Pepperoni Delight	Pineapples
8	Crunchy Thin	Pepperoni Delight	Beef

Table 3. Pairwise Test Cases for Pizza Online Ordering System

Test Case ID	Crust	Flavor	Topping
1	Classic Hand Tossed	Vegetarian	Pineapples
2	Classic Hand Tossed	Pepperoni Delight	Beef
3	Crunchy Thin	Pepperoni Delight	Pineapples
8	Crunchy Thin	Vegetarian	Beef

Based on the given parameters and values in Table 1, Table 2 depicts the all exhaustive test cases (with $2^3 = 8$ combinations). If we consider an interaction strength of 2 (i.e. $t=2$), we can get optimally reduce the test case to 4 (see Table 3). Analyzing Table 3, we note that all the 2-way interaction between Crust-Flavor, Crust-Topping, and Flavor-Topping has been covered by at least 1 test (refer to Table 4).

Table 4. Pair Coverage Analysis

Pair Interaction	Crust	No of Occurrences	Test ID
Crust-Flavor	Classic Hand Tossed, Vegetarian	1	1
	Classic Hand Tossed, Pepperoni Delight	1	2
	Crunchy Thin, Vegetarian	1	8
	Crunchy Thin, Pepperoni Delight	1	3
Crust-Topping	Classic Hand Tossed, Pineapple	1	1
	Classic Hand Tossed, Beef	1	2
	Crunchy Thin, Pineapple	1	3
	Crunchy Thin, Beef	1	8
Flavor-Topping	Vegetarian, Pineapple	1	1
	Vegetarian, Beef	1	8
	Pepperoni Delight, Pineapple	1	3
	Pepperoni Delight, Beef	1	2

At a glance, Table 4 usefully proves that all the pairwise interactions have been rightfully covered. Nonetheless, a closer look reveals some inconsistencies, that is, in terms of the existence of impossible pairing combinations or better known as constraints. Here, vegetarian flavor cannot appear with beef (see Table 5).

Table 5. Constraints Test Cases for Pizza Online Ordering System

Test Case ID	Crust	flavor	Topping
1	Classic Hand Tossed	Vegetarian	Beef
2	Crunchy Thin	Vegetarian	Beef

Table 6 shows the best possible test cases size which avoids the unwanted test cases {i.e., (Classic Hand Tossed: Vegetarian: Beef), and (Crunchy Thin: Vegetarian: Beef)} for the pizza software system configuration. Interaction element (or pair) analysis (see Table 7) confirms that no Vegetarian – Beef pair exists and all other pairings are covered at least once.

Table 6. Test Cases for Pizza Online Ordering System

Test Case ID	Crust	Flavor	Topping
1	Classic Hand Tossed	Vegetarian	Pineapples
2	Crunchy Thin	Pepperoni Delight	Beef
3	Classic Hand Tossed	Pepperoni Delight	Beef
4	Crunchy Thin	Vegetarian	Pineapples
5	Crunchy Thin	Pepperoni Delight	Pineapples

Table 7. Pair Coverage Analysis

Pair Interaction	Crust	No of Occurrences	Test ID
Crust-Flavor	Classic Hand Tossed, Vegetarian	1	1
	Classic Hand Tossed, Pepperoni Delight	1	3
	Crunchy Thin, Vegetarian	1	4
	Crunchy Thin, Pepperoni Delight	2	2,5
Crust-Topping	Classic Hand Tossed, Pineapple	1	1
	Classic Hand Tossed, Beef	1	3
	Crunchy Thin, Pineapple	2	4,5
	Crunchy Thin, Beef	1	2
Flavor-Topping	Vegetarian, Pineapple	2	1,4
	Vegetarian, Beef	0	0
	Pepperoni Delight, Pineapple	1	5
	Pepperoni Delight, Beef	2	2,3

To determine the best combinations to cover all combinations and avoid constraints within the context of applying Harmony Search Strategy is the focus of our paper.

3 HSS Strategy

This paper extends our previous work in [13-15] on t-way interaction testing strategy, called HSS, using Harmony Search Algorithm (see Fig. 2).

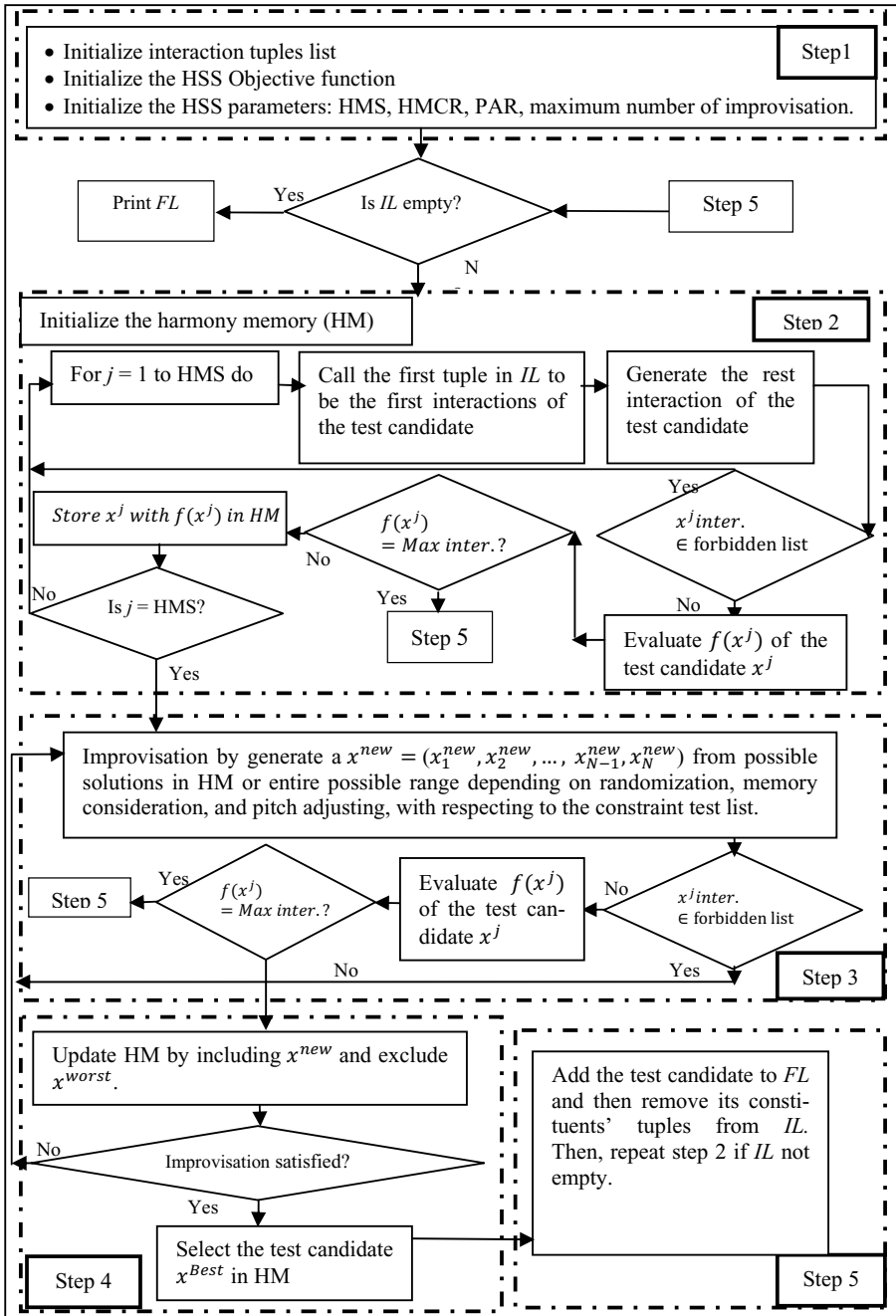


Fig. 2. HSS Strategy

In a nut shell, HSS mimics the musical performance in Harmony Search Algorithm searching for a good and stable tone from playing of a group of musical tools [16-20]. HSS is controlled by four parameters (Harmony memory size (HMS), Improvisations number, Harmony memory considering rate (HMCR), and Pitch adjusting rate (PAR)). HMS dictates the number of possible solutions. Typically, large HMS gives more optimal solution but in the expense of computational time. Improvisation number defines the possible number of iterations used to improve the existing solutions. HMCR determines the probability of diversification or the intensification to improve the solution either by generating new solution randomly or by improving the existing solution stored in harmony memory (HM). PAR controls the solution improvement by traveling around the best local solution stored in HM [19, 20].

HSS works by first initializing the HSS Objective function $f(x^j)$ (see Eq. 1) and the values of the (HMS, HMCR, and PAR) with favorable values. Then, HSS loads and initializes the forbidden list with the constraints test cases.

$$\text{Maximize } f(x^j) = \bigcup_{i=0}^N x_i^j \in \text{Interaction tuples list},$$

$$\text{Subject to } x_1^j, x_2^j, \dots, x_N^j \text{ in } k_1, k_2, \dots, k_N; j = 1, 2, \dots, \text{HMS}; i = 1, 2, \dots, N \quad (1)$$

After initializing the forbidden list, HSS loads and generates the interaction elements list (*IL*) which consists of all possible interactions[13-15].Noted here is the fact that the HSS Objective function $f(x^j)$ is to maximize the interaction elements upon the selection of any particular x^j (refer to Eq. 2).

$$x^j = (x_1^j, x_2^j, \dots, x_{N-1}^j, x_N^j) \quad (2)$$

When the generation of interaction elements completes, HSS loads the HM by a randomly generated test case candidate x^j (see Eq. 3). Here, x^j is neglected if it is listed in the forbidden list upon which a new random candidate will be re-generated. Based on the test candidate x^j , HSS evaluates its $f(x^j)$ in terms of the number of interaction elements that have been covered in *IL*. x^j will be added to the final test suite list (*FL*) and removes its corresponding interaction elements if the $f(x^j)$ covered the maximum interaction elements in *IL*. Otherwise, HSS stores this x^j to the HM. This step is repeated until either *IL* is empty or HM is full with all possible HMS test case candidates x^{HMS} . In this case, improvisation process is desired to improve the existing x^{HMS} in HM.

$$\begin{array}{l} x^1 \\ x^2 \\ \vdots \\ x^{\text{HMS}-1} \\ x^{\text{HMS}} \end{array} = \begin{array}{l} \left[\begin{array}{cccc} x_1^1 & x_2^1 & \dots & x_{N-1}^1 & x_N^1 \\ x_1^2 & x_2^2 & \dots & x_{N-1}^2 & x_N^2 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ x_1^{\text{HMS}-1} & x_2^{\text{HMS}-1} & \dots & x_{N-1}^{\text{HMS}-1} & x_N^{\text{HMS}-1} \\ x_1^{\text{HMS}} & x_2^{\text{HMS}} & \dots & x_{N-1}^{\text{HMS}} & x_N^{\text{HMS}} \end{array} \right] \\ \end{array} = \begin{array}{l} f(x^1) \\ f(x^2) \\ \vdots \\ f(x^{\text{HMS}-1}) \\ f(x^{\text{HMS}}) \end{array} \quad (3)$$

The improvisation process starts by constructing a new test case candidate x^{new} either randomly or improving the existing x^j stored in HM based on the probability value of the HMCR and PAR (see Eq. 4) (for more details see [13-15]).

$$x^{new} = (x_1^{new}, x_2^{new}, \dots, x_{N-1}^{new}, x_N^{new}) \tag{4}$$

Similar to the step for initializing HM, x^{new} will be neglected if it is listed in the forbidden list. Otherwise, $f(x^{new})$ will be evaluated accordingly. x^{new} will be stored in FL if $f(x^{new})$ covered the maximum interaction elements in IL . Otherwise, this x^{new} will be used to update the current contents of HM by replacing the worst test case candidate in HM x^{worst} . This process will continue until no further improvisation is possible (i.e. x^{new} is no better than x^{worst}). Then, HSS adds the x^{best} in HM to FL and removes its corresponding interaction elements from IL . The overall iteration will be repeated until all interaction elements in IL have been covered. Then, the FL will be printed.

4 Evaluation HSS Strategy

In this section, we benchmark the HSS support for constraints with TestCover as published in [21]. Here, the system under study involves a networked computer system with 3 4-valued parameters and 2 3 valued-parameters. Specifically, the parameters involved Operating System (4 values), Display Resolution (3 values), Connection, Browser (4 values), and Applications (3 values). Table 8 summarizes the complete parameters and values description for the system whilst Table 9 highlights the list of defined constraints (with defined ‘x’ as don’t care values).

Table 8. Networked Computer System

Operating System (OS)	Display Resolution (DR)	Connection (Con.)	Browser (Bro.)	Applications (App.)
XP	Low	Wi-fi	IE	App1
MacOS	Medium	Dsl	Firefox	App2
Linux	High	Cable	Opera	App3
Vista		Lan	Safari	

Table 9. Defined Constraints for Networked Computer System

Impossible Test Cases	Constraints on Valid Configurations				
	OS	DR	Con.	Bro.	App.
Linux	x	X	IE	x	
Linux	x	X	Safari	x	
MacOS	x	X	IE	x	

For this system under study, we adopt the improvisation=1000, HMS=100, HMCR=0.7 and PAR=0.2 based on our earlier work in [13-15]. We have reported the best test cases generated by TestCover and HSS strategy in Tables (9, and 10)

respectively. Our running environment consist of a desktop PC with Windows XP, 2.8 GHz Core 2 Duo CPU, 1 GB of RAM. The HSS strategy is coded and implemented in Java (JDK 1.6). Tables 10 and 11 highlight the constraints test cases produced by TestCover and HSS respectively for the interaction strength, $t=2$ (i.e. pairwise).

Table 10. Pairwise Constraints Test Cases for 3 4-Valued Parameters and 2 3-Valued Parameters Generated by TestCover

Test Case ID	OS	DR	Con.	Bro.	App.
	4 Values	3 Values	4 Values	4 Values	3 Values
1	MacOS	Low	Lan	Opera	App1
2	Vista	High	wi-fi	Opera	App2
3	Linux	Medium	Cable	Opera	App3
4	Vista	Medium	Dsl	Firefox	App1
5	XP	Low	Lan	Firefox	App3
6	MacOS	High	Cable	Firefox	App2
7	MacOS	Medium	wi-fi	Safari	App3
8	XP	Low	wi-fi	IE	App1
9	Linux	Low	Dsl	Opera	App2
10	Vista	Low	Cable	Safari	App3
11	XP	Medium	Dsl	IE	App3
12	XP	High	Cable	IE	App2
13	Vista	Medium	Lan	Safari	App2
14	MacOS	High	Dsl	Safari	App1
15	Linux	High	Lan	Firefox	App2
16	Linux	Medium	wi-fi	Firefox	App1
17	XP	High	Cable	Opera	App1
18	XP	High	wi-fi	Safari	App3
19	Vista	Low	Dsl	IE	App2
20	XP	Low	Lan	IE	App1

Table 11. Pairwise Constraints Test Cases for 3 4-Valued Parameters and 2 3-Valued Parameters Generated by HSS

Test Case ID	OS	DR	Con.	Bro.	App.
	4 Values	3 Values	4 Values	4 Values	3 Values
1	XP	Medium	wi-fi	Firefox	App2
2	Vista	Low	Lan	Safari	App2
3	MacOS	High	wi-fi	Safari	App1
4	Linux	Medium	Lan	Opera	App1
5	XP	High	Cable	IE	App3
6	MacOS	Low	Dsl	Firefox	App3
7	Linux	High	Dsl	Opera	App2

Table 11. (Continued)

8	Vista	Medium	Dsl	IE	App1
9	Linux	Low	wi-fi	Opera	App3
10	MacOS	Medium	Cable	Safari	App3
11	Vista	High	Cable	Firefox	App1
12	XP	Low	Cable	Opera	App1
13	Vista	Low	wi-fi	IE	App2
14	MacOS	High	Lan	Firefox	App2
15	XP	High	Lan	IE	App3
16	XP	Low	Dsl	Safari	App3
17	Linux	Medium	Cable	Firefox	App2
18	Vista	High	Cable	Opera	App3
19	MacOS	Medium	Cable	Opera	App3

Referring to Tables 10 and 11, we note that both TestCover and HSS are able to correctly generate the (pairwise) test case as required. Here, a close inspection reveals that all the interaction elements (i.e. pairwise interactions) do exist in the final test suite whilst the constraints ones are rightfully missing. Unlike TestCover which produces 20 test cases, HSS produces 19 test cases. As such, HSS appears to be more optimal as far as generation of test suite than that of TestCover for 3 4-valued parameters and 2 3 valued-parameters with defined constraints in Table 9.

Apart from TestCover, noted here is the fact that there exists other strategy that also addresses constraints including mAETG_SAT [22], PICT [23, 24], and SA_SAT [22] respectively. mAETG_SAT (i.e. a variant AETG strategy) generates one final test case for every cycle of iterations. For each cycle mAETG_SAT generates a number of candidate test cases, and from these candidates, one is greedily selected provided that it is not in the constraints list. In such a case, the other candidate is chosen even if it is only the second best. PICT generates all interaction elements and randomly selects their corresponding interaction combinations to form the complete test case. If the complete test case makes up the constraints test cases, a new random combination will be generated. SA_SAT (a variant of Simulated Annealing strategy) is perhaps the only AI-based strategy that addresses the problem of constraints. SA_SAT relies on large random space to generate t-way test suite. Using probability based transformation equation, SA adopts binary search algorithm to find the best test case per iteration by taking consideration of constraints test cases. Although mAETG_SAT, PICT, and SA_SAT usefully implements constraints, we are unable to compare with them here owing to the fact that there were limited published results and most of their implementations are not publicly available.

5 Conclusion

In this paper, we define the problem domain of incorporating constraints test cases for t-way testing. In doing so, we compare the support for constraints between HSS and

TestCover. We note that both HSS and TestCover give competitive results. As part of our future work, we hope to benchmark HSS with more strategies as well as with more configurations.

Acknowledgment. This research is partially funded by the generous grant (“Investigating T-Way Test Data Reduction Strategy Using Particle Swarm Optimization Technique”) from the Ministry of Higher Education (MOHE), the USM research university grants (“Development of Variable Strength Interaction Testing Strategy for T-Way Test Data Generation”), and the USM short term grants (“Development of a Pairwise Test Data Generation Tool with Seeding and Constraints Support”).

References

1. Cohen, D.M., Dalal, S.R., Fredman, M.L., Patton, G.C.: The AETG System: An Approach to Testing Based on Combinatorial Design. *IEEE Transactions on Software Engineering* 23(7), 437–444 (1997)
2. Cohen, D.M., Dalal, S.R., Kajla, A., Patton, G.C.: The Automatic Efficient Test Generator (AETG) System. In: *Proceedings of the 5th International Symposium on Software Reliability Engineering*, pp. 303–309. IEEE Computer Society, Monterey (1994)
3. Zamli, K.Z., Klaib, M.F.J., Younis, M.I., Isa, N.A.M., Abdullah, R.: Design And Implementation Of A T-Way Test Data Generation Strategy With Automated Execution Tool Support. *Information Sciences* 181(9), 1741–1758 (2011)
4. Younis, M.I., Zamli, K.Z., Isa, N.A.M.: MIPOG-Modification of the IPOG Strategy for T-Way Software Testing. In: *Proceeding of the Distributed Frameworks and Applications, DFmA* (2008)
5. Ahmed, B.S., Zamli, K.Z.: PSTG: A T-Way Strategy Adopting Particle Swarm Optimization. In: *Proceedings of 4th Asia International Conference on Mathematical /Analytical Modelling and Computer Simulation*. IEEE Computer Society (2010)
6. Ahmed, B.S., Zamli, K.Z., Lim, C.P.: Constructing a T-Way Interaction Test Suite Using the Particle Swarm Optimization Approach. *International Journal of Innovative Computing, Information and Control* 8(1), 431–452 (2012)
7. Ahmed, B.S., Zamli, K.Z.: A Variable-Strength Interaction Test Suites Generation Strategy Using Particle Swarm Optimization. *Journal of Systems and Software* 84(12), 2171–2185 (2011)
8. Othman, R.R., Zamli, K.Z.: ITTDG: Integrated T-way Test Data Generation Strategy for Interaction Testing. *Scientific Research and Essays* 6(17), 3638–3648 (2011)
9. Ong, H.Y., Zamli, K.Z.: Development of Interaction Test Suite Generation Strategy with Input-Output Mapping Supports. *Scientific Research and Essays* 6(16), 3418–3430 (2011)
10. Colbourn, C.J., Cohen, M.B., Turban, R.C.: A Deterministic Density Algorithm for Pairwise Interaction Coverage. In: *Proceedings of the IASTED International Conference on Software Engineering, Citeseer*, vol. 41, pp. 242–252 (2004)
11. Bryce, R.C., Colbourn, C.J.: A Density-Based Greedy Algorithm for Higher Strength Covering Arrays. *Software Testing, Verification & Reliability* 19(1), 37–53 (2009)
12. Bryce, R.C., Colbourn, C.J.: The Density Algorithm for Pairwise Interaction Testing: Research Articles. *Software Testing, Verification & Reliability* 17(3), 159–182 (2007)
13. Alsewari, A.R.A., Zamli, K.Z.: Interaction Test Data Generation Using Harmony Search Algorithm. In: *Proceeding of IEEE Symposium on Industrial Electronics & Applications*. IEEE Computer Society, Langkawi (2011)

14. Alsewari, A.R.A., Zamli, K.Z.: A Harmony Search Based Pairwise Sampling Strategy for Combinatorial Testing. *International Journal of the Physical Sciences* 7(7), 1062–1072 (2012)
15. Alsewari, A.R.A., Zamli, K.Z.: Design and Implementation of a Harmony-search-based Variable-strength T-Way Testing Strategy with Constraints Support. *Information and Software Technology* (in Press),
<http://dx.doi.org/10.1016/j.infsof.2012.01.002>
16. Geem, Z.W., Kim, J.H.: A New Heuristic Optimization Algorithm: Harmony Search. *Simulation* 76(2), 60–68 (2001)
17. Zou, D., Gao, L., Wu, J., Li, S., Li, Y.: A Novel Global Harmony Search Algorithm for Reliability Problems. *Computers & Industrial Engineering* 58(2), 307–316 (2009)
18. Pan, Q.-K., Suganthan, P.N., Tasgetiren, M.F., Liang, J.J.: A Self-Adaptive Global Best Harmony Search Algorithm for Continuous Optimization Problems. *Applied Mathematics and Computation* 216, 830–848 (2010)
19. Yang, X.-S.: Harmony Search as a Metaheuristic Algorithm. In: Geem, Z.W. (ed.) *Music-Inspired Harmony Search Algorithm*. SCI, vol. 191, pp. 1–14. Springer, Heidelberg (2009)
20. Zou, D., et al.: A Novel Global Harmony Search Algorithm for Reliability Problems. *Computers & Industrial Engineering* 58(2), 307–316 (2009)
21. Sherwood, G.: TestCover (2006), <http://testcover.com/pub/constex.php>
22. Cohen, D.M., et al.: The AETG System: An Approach to Testing Based on Combinatorial Design. *IEEE Transactions on Software Engineering* 23(7), 437–444 (1994)
23. Czerwonka, J.: Pairwise Testing in the Real World: Practical Extensions to Test-Case Scenarios. In: *Proceedings of 24th Pacific Northwest Software Quality Conference*, Citeseer (2006)
24. Keith, Doug, H.: PICT (2006),
<http://testmuse.wordpress.com/2006/04/05/pict-tool-available/2006>

Evaluating Disaster Management Knowledge Model by Using a Frequency-Based Selection Technique

Siti Hajar Othman^{1,2} and Ghassan Beydoun²

¹ Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
sho492@uowmail.edu.au

² School of Information Systems and Technology, Faculty of Informatics,
University of Wollongong, Wollongong NSW 2522, Australia
beydoun@uow.edu.au

Abstract. Disaster Management (DM) is a multidisciplinary endeavour and a very difficult knowledge domain to model. It is a diffused area of knowledge that is continuously evolving and informally represented. *Metamodel* is the output artefact of metamodeling, a software engineering approach, which makes statements about what can be expressed in the valid models of the knowledge domain. It is an appropriate high level knowledge structure to facilitate it being communicated among DM stakeholders. A Disaster Management Metamodel (DMM) is developed. To satisfy the expressiveness and the correctness of a DMM, in this paper we present a metamodel evaluation technique by using a *Frequency-based Selection*. The objective of this technique is to evaluate the importance of the individual concepts used in the DMM, thus, the quality of the metamodel can be measured quantitatively.

Keywords: Frequency-based Selection, Metamodel, Disaster Management, Knowledge Model, Model Transformation.

1 Introduction

Knowledge is information presented within a particular context, yielding insight into actions taken in the context [5]. The effectiveness of a knowledge model depends on the abstraction effectiveness of individual concepts used to describe the domain [1]. The richer the meaning attached to the concepts, the less time a modeller requires to operationalise the model [2]. The meaning and definition of concept terminologies and their relationships are not only domain specific but may even differ from one observer to another [3, 4]. A challenge in creating a new model and identifying the domain concepts is resolving ambiguity and inconsistencies of domain terminologies. A model synthesis process adapts the software engineering practice, '*Metamodeling*' and provides means to reconcile the inconsistencies across observers. This is a modular and layered process typically used to endow a well-established methodology or a modelling language with an abstract notation, discerning the abstract syntax and semantics of the modelling elements. By focussing on the evaluation and the

metamodelling process on Disaster Management, this paper makes a significant contribution using metamodelling to unify key concepts into a metamodel that can be used as knowledge sharing platform. Later, this artefact can be reused by DM stakeholders to develop their DM customised models by retrieving parts and components of previous solutions to suit their current needs (disaster on hand). DM knowledge can be viewed from different lenses (e.g.: *Know What, Know Who, Know How, Know Where, Know Why...*) and understanding them is required to support its structuring. Structuring the Disaster Management (DM) knowledge requires understanding of its environment and elements (organisational, operations, processes or stakeholders). DM knowledge is also scattered in public resources such as the internet, books, online databases, libraries, newspapers or pamphlets. How this knowledge is applied in new situations is rarely explored [6]. Indeed, reusing and sharing knowledge is a form of knowledge creation and as pointedly stated in Beerli et. al [7 pp.3]: “*Knowledge can be regarded as the only unique resource that grows when shared, transferred and skilfully managed.*” By developing an appropriate high level knowledge structure for this domain through a metamodel, a DM modelling knowledge is identified.

A metamodel identifies domain features and related concepts (as any other model) and is created with the intent to formally describe the semantics underpinning a formal modelling language [8]. Without a metamodel, semantics of domain models can be ambiguous. In metamodel, *concept* and *relationships* are two important elements. A concept characterizes domain entities and relationships characterizes *links* between them [9]. Metamodel must form true or faithful representations so that queries of a model give reliable statements about reality, or manipulations of the model result in reliable adaptations of reality. A metamodel requires evaluation to satisfy the requirement of generality, expressiveness and completeness of the artefact. With respect to this, this paper presents how the *Frequency-Based Selection* (FBS) is used to evaluate the DM metamodel. The rest of this paper is structured as follows: Section 2 describes the related work on disaster management, metamodel evaluation in metamodelling and the DMM. Section 3 presents the actual evaluation of FBS against the DM metamodel. Section 4 presents result of the evaluation and Section 5 concludes the paper with a discussion on our findings and future work.

2 Related Works

In this section, the related works on disaster management knowledge, metamodel evaluation in a metamodelling environment and a DM metamodel are discussed before the actual implementation of FBS technique is presented.

2.1 Disaster Management Knowledge

Disaster Management (DM) aims to reduce or avoid the potential losses from hazards, assure prompt and appropriate assistance to victims of disaster and achieve rapid and effective recovery. The United Nation (UN) recognises at least 40 types of disasters

and classifies them into two types of disasters including: *natural* and *technological/man-made* disasters. Knowledge applied in this domain changes across various phases of a disaster. Standard DM phases include *mitigation*, *preparedness*, *response* and *recovery* [10]. Structuring the DM knowledge requires understanding of its environment and elements (organisational, operations, processes or stakeholders). There are varieties of DM models which have been developed by many stakeholders (researchers, government or non-government agencies, community and individuals). These models can broadly be grouped according to seven main perspectives: *disaster phase* oriented (e.g.: recovery or preparedness stage), *organisation* oriented (e.g.: Red-Cross coordination, State Police arrangement during emergency, *User/Role* oriented (e.g.: volunteers, hospitals, aid agencies), *disaster type* oriented (e.g.: earthquake, disease infection), *technology* oriented (e.g.: GIS, Satellite for disaster monitoring), *disaster activity* oriented (e.g.: evacuation, search and rescue) or *decision type* oriented (e.g.: reasoning technique for disaster decision making). In developing a metamodel specific to this challenging domain, typically, the first question that will be asked after any metamodel is successfully developed is *how the metamodel is relevant* to its real application domain. Therefore, evaluation to the artefact is crucial.

2.2 Metamodel Evaluation and Its transformation in a Metamodelling Approach

The quality of the metamodel is measured based on how the metamodel can fulfil the purpose of its development [11]. In other words, the created metamodel has to respond to the needs of the domain practitioners. This includes increasing the transparency to the knowledge encoded within the domain applications and be able to be validated by relevant experts in the domain. three motivations of why metamodel requires evaluation are: (i) initial domain literatures used to develop the metamodel is sometimes not complete, therefore it is necessary to fill in some blanks with hypothesis unsupported by the initial literature; (ii) domain literature is not always coherent, hence when creating a metamodel it might be inescapable to make controversial choices; (iii) metamodeler might be biased, thus when creating a metamodel, he or she might unwillingly create distortions [12].

In metamodelling, metamodels and models relate through *model transformation* [13]. During metamodel evaluation, model transformations are explored and evaluated. Model transformation is one of a process of converting one model to another model in a metamodelling framework. Also, the acceptance of a system of metamodels for practical use depends on the validity of the metamodels and the transformations on a given abstraction hierarchy [14 pp. 163]. Model-to-model transformation is a key technology for Object Management Group (OMG)'s Model Driven Architecture [15] and underpins realising the various functionalities of DMM. DM solutions need to be transformed to DMM during knowledge storage and DMM needs to be transformed back to various DM solution models by DM users later. This research follows the modelling abstraction offered by Meta Object Facility (MOF)

framework in performing a transformation of metamodel-to-model for DMM. The MOF framework provides a capability to support different types of metadata in its four meta-layers: *User Model* level (M0), *Model* level (M1), *Metamodel* level (M2), and *Meta-metamodel* (M3) and can be used to define different information models. Model transformation in MOF (presented in Figure 1) can be viewed in *vertical* and *horizontal* dimensions [16].

A *horizontal transformation* involves transforming a model into a target model at the *same level* of modelling abstraction. This is true no matter how high or low the artefact modelling abstraction level is [16]. Semantics of horizontal transformations is applied in this paper when DMM is horizontally transformed to produce its new updated version after performing the FBS technique against the metamodel. A *Vertical transformation* presents the transformation of model from one level to a *different level* of modelling abstraction. The transformation can either be from an upper to a lower level (e.g.: from metamodel (M2) level to model (M1) level), or conversely from a lower to an upper level (e.g.: from model level (M1) to metamodel level (M2)). The vertical transformation is performed when “*the DM model and DM User Model are being derived from its conformant DMM (metamodel)*”. The process of deriving individual concepts in the models is also vertical transformations.

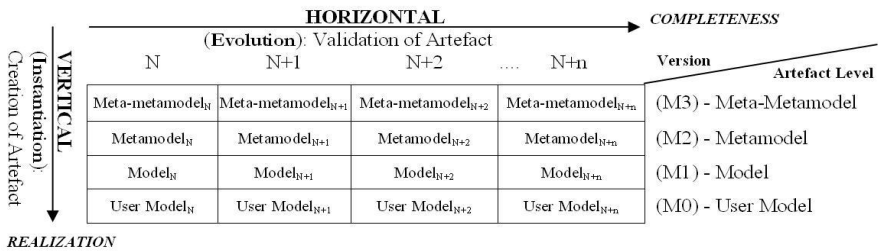


Fig. 1. Horizontal and vertical model transformation in MOF metamodeling

2.3 The Disaster Management Metamodel

The DMM is the output of the metamodeling approach applied in this paper. It will serve as a representational layer to enable appropriate domain modelling and knowledge storage relating to different DM activities and disaster scenarios. It is a DM specific language developed by using the *8 step Metamodeling Creation Process* adapted from Beydoun et al. [17, 18]. In [19], this initial DMM is developed and uses DMM1.0 as its version. The metamodel is presented in four sets of concept classes: the *Mitigation*, *Preparedness*, *Response* and *Recovery* class of concepts. Each set represent a corresponding DM phase and clearly describes the DM domain to its users. This initial metamodel has been first evaluated in [20] by using a ‘*Comparison against other models*’ technique. The aim of the first evaluation is to identify any missing concepts in the metamodel and to also ensure its broad coverage. Result from the first evaluation changes the DMM1.0 to its updated version, a DMM1.1. Normally

a metamodel requires iterative evaluation in its development because it needs to achieve different quality goal in each evaluation cycle. In this paper, with the aim to evaluate the importance of the individual concepts included in DMM, this time the DMM is validated for a second cycle by using the FBS technique. Result derived from the evaluation conducted in this paper creates the DMM1.2 version.

To visibly show the changes occurred before and after performing the FBS, this paper uses the Mitigation-phase and the Response-phase class of concepts as the metamodel samples, presented in the Figure 2 and Figure 3 respectively. In both classes, concepts and their relationships are depicted. The following shows the list of concepts used in each DMM classes:

- i) **DMM Mitigation concepts:** MitigationPlan, MitigationOrganisation, MitigationTask, NeedsPlanning, BuildingCodes, Land-UsePlanning, InformationUpdates, MitigationGoal, RiskReduction, People, Property, Lifeline, NaturalSite, HazardAssessment, RiskAnalysis, StructuralMitigation, Non-StructuralMitigation, Vulnerability, DisasterRisk, StrategicPlanningCommittee, Legislation, Insurance and Exposure;
- ii) **DMM Preparedness concepts:** PreparednessPlan, PreparednessOrganisation, PreparednessTask, SuppliesRegistry, Warning, PreparednessGoal, Evacuation, BeforeDisaster, Event, DecisionMaking, Administration, EmergencyPublicInformation, Pre-Position, DisasterFactor, Exposure, DisasterRisk, Training, PreparednessTeam, Media, MutualAidAgreement, PublicEducation, PublicAwareness, Resource, Monitoring, AidAgency;
- iii) **DMM Recovery concept:** RecoveryPlan, RecoveryOrganisation, RecoveryTask, Demobilization, LongTermPlanning, RecoveryGoal, Reconstruction, AfterDisaster, DamageAssessment, TaskReview, Resilience, Victim, EmergencyManagementTeam, Resource, DebrisRemoval, Effect, EconomicRestoration, Exposure, FinancialAssistance, MentalHealthRecovery, AidDistribution.

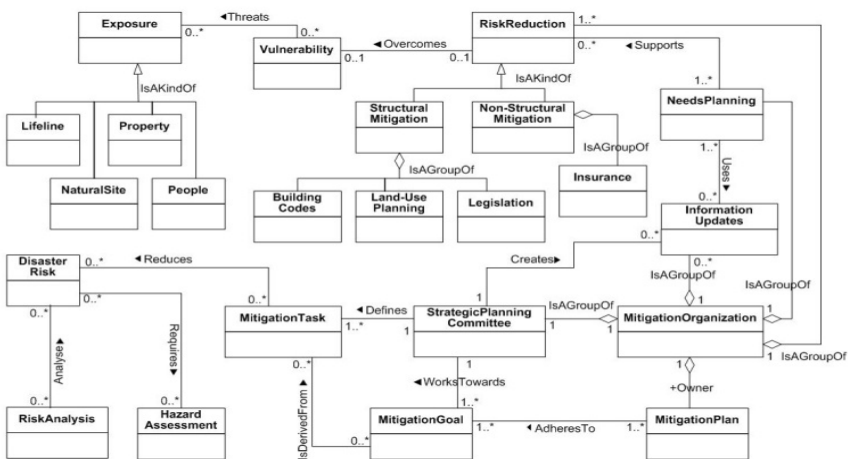


Fig. 2. The DMM1.1: The first validated version of Mitigation-phase class of concepts [20]

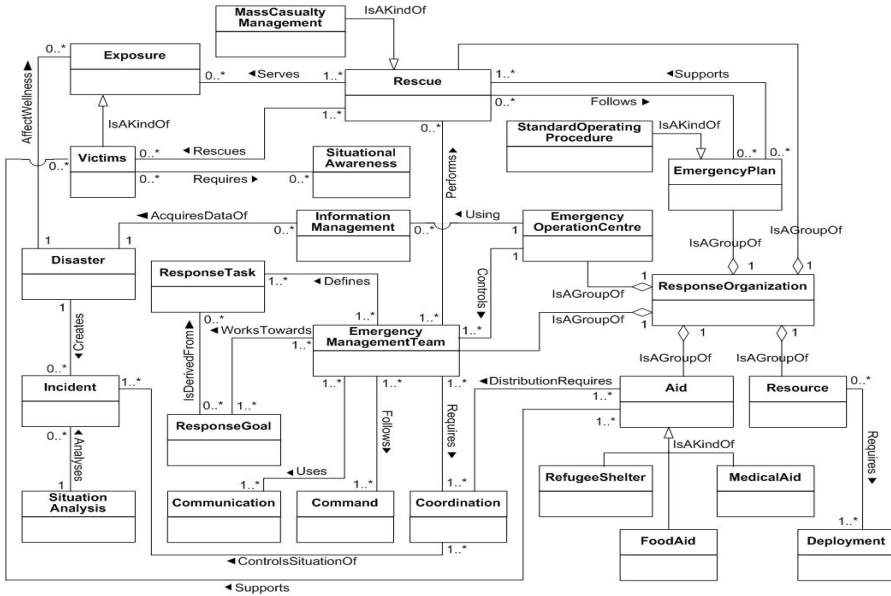


Fig. 3. The DMM1.1: The first validated version of Response-phase class of concepts [20]

3 Frequency-Based Selection Implementation

In this section, FBS and a special frequency parameter used to estimate the importance of the individual concepts in the DMM, a *Degree of Confidence* (DoC) are described. This is then followed by the representation of FBS actual implementation against the DMM. Result of this evaluation is presented in the next section.

3.1 Frequency-Based Selection

Frequency-Based Selection is a *Feature Selection* technique that evaluates the importance of individual concepts in the model developed [21]. It is based on the premise that the best model is formed using the most common features [22] and it is commonly used, for example, in data mining, software analysis, and medical retrieval systems. By performing FBS, *features (concepts)* that do not have correlations (or a need) to the classification are removed from DMM. The way FBS is adapted to validate the significance of DMM concepts acknowledges the five metamodel quality criteria described in [23]: a reasonable depiction (e.g. a statistical measure) of the relative ‘importance’ of candidate concepts; a predictive ability of the metamodel that is reasonably consistent with baseline models across the domain; the metamodel has independent meaningful variables; the metamodel highlights all input variables essential to describe critical components of a domain and the metamodel can provide a storyline to its users to tell how and why a derived model behaves as it does.

To perform FBS, 10 set of existing DM models of Set V2 is used (Table 1). The set is formed based on phase-specific and other perspectives (e.g.: role/user, operation, organisation, decision or technology-based DM models. For a selection, a model coverage values ($R_{coverage}$) are used: 0.3 is assigned to a model that has full coverage to all phases in DM (Mitigation, Preparedness, Response and Recovery phase). DM models with coverage less than 0.3 focuses on specific DM phases, activities or roles, as follows: 0.2 is assigned to coverage of models that can cover 2-3 DM phases in their models. 0.1 is set to a model that covers only one DM phase (any one of four DM phases) or a specific DM perspectives (e.g.: evacuation operation (*operation-based*), the roles of the disaster analyzer in disaster monitoring (*user/role-based*)). If a model does not cover any single DM phase fully, 0.0 is set to the model and will be excluded from any further investigation. This selection process ensures that all DMM concepts are tested against some concepts in the models selected. That is, each DMM concept is examined in a vertical model transformation. Where required, DMM is modified to ensure that it can represent all models in the validation sets (through a horizontal transformation).

Table 1. A set of 10 DM models (Set V2) for an evaluation of DMM

SET V2		$Y_{published}$	$R_{coverage}$	Model coverage: (Perspective)
1	Disaster Risk Management & Mitigation Management, [24]	2006	0.3	All Phases: (Activity-based)
2	Policies for Guiding Planning for Post-Disaster Recovery and Reconstruction, [25]	2005	0.2	Mitigation and Recovery: (Management-based)
3	Disaster Risk Management Working Concept, [26]	2002	0.3	All Phases: (Activity-based)
4	Disaster Information, Innovative Disaster Information Service, [27]	2008	0.3	All Phases: (Technological-based)
5	Situation-Aware Multi-Agent System for Disaster Relief Operations Management, [28]	2006	0.2	Preparedness and Response: (Technological-based)
6	An Approach to the Development of Commonsense Knowledge for Disaster Management, [29]	2007	0.3	All Phases: (Disaster and Activity-based)
7	Earthquake Protection, [30].	1992	0.3	All Phases: (Disaster and Organisation-based)
8	Disaster Stage and Management Model, [31]	2008	0.3	All Phases: (Disaster-based and Management-based)
9	Teaching Disaster Nursing by Utilizing the Jennings Disaster Nursing Management Model, [32].	2004	0.3	All Phases: (User/Role-based)
10	Disaster Management – a Theoretical Approach, [33]	2008	0.3	All Phases: (Disaster-based)

(Notes: $Y_{published}$ – The Year model is published, $R_{coverage}$ – The coverage of models)

3.2 The Degree of Confidence (DoC)

Using the concept frequency, an importance value for each concept in DMM is estimated and expressed as the ‘*Degree of Confidence* (DoC)’. This value designates the expected probability that a DMM concept is used in a randomly chosen disaster model. DoC is derived by dividing ‘the *frequency* of how many times a concept appears in all the investigated models (Set V2)’ with ‘the *total number* of Set V2 models’. For this purpose, DoC is based on the list of concepts that appeared in the DMM1.1 (our metamodel after its first evaluation) and is defined as follows:

$$\text{Degree of Confidence (DoC)} = \frac{\text{Frequency of Concept}}{\text{Total Model of Set V2}} \times 100\% \quad (1)$$

3.3 The FBS Evaluation against the DMM

To perform the FBS technique on DMM, concepts to be verified from models in the evaluation Set V2 are first collated. This is to ensure that these concepts can all be refined using DMM1.1. As described in Section 3.1, Set V2 is a selection of DM models that have a wider DM coverage. Specialised DM models will naturally focus on a specific DM phase and naturally omit the use of some concepts. Therefore using models with wider coverage will provide a better indication on the frequency of concepts across the models. Their use will enable a frequency count of the individual DMM concepts. Concepts used in the models of Set V2 that are found similar and that are a refinement of DMM concepts are scored in this evaluation. The higher their score, the more important the concepts are deemed to the DM domain. Concepts that have a low score are revisited and are liable for deletion.

In applying FBS using the models in Set V2, DMM concepts that derive concepts of those models are identified. The frequency of usage of DMM concepts in those derivations is compiled and shown in Table 4 (for the Mitigation-phase concepts), Table 5 (for the Response-phase concepts). In what follows in this section, refinement of every model of Set V2 is overviewed. The outcome of FBS evaluation, leading to DMM1.2, is then presented in Section 4. Two models of Set V2 (Model 7 - The Organisation Model in Earthquake Disaster [30] and Model 9 – The Jennings Disaster Nursing Management Model [32] are used as the evaluation implementation samples of FBS.

3.3.1 Sample FBS 1: Against the Organisation Model in Earthquake Disaster (Model 7 of Set V2)

Reconstruction following an earthquake requires a renovation of the economy, jobs and income, daily life and social relations. Coburn [30] proposed that reconstruction tasks following an earthquake get organised sectorally (Figure 4). Coburn provides a few examples of how earthquake damage can be classified by sector and responsible organisations. Sectoral approach is advocated as different authorities have different responsibilities and reconstruction needs. As an example, for damages to schools, universities, and kindergartens including the number of lost classroom places and the loss of school equipment, become the responsibility of the Department of Education, Regional Education Authority, Private Education Institutions and the Department of Public Works of the country. As another example, any damage that may occur to agricultural building stock, loss of livestock, damage to equipment, vehicles, market gardening, greenhouses, food processing plants, food and produce storage becomes the responsibility of the Department of Agriculture and Food, Farming Organisations, Private Owners and Consumer Organisations. This model can be generated from the concepts *RecoveryTask*, *RecoveryOrganisation* and *RecoveryGoal* in DMM.

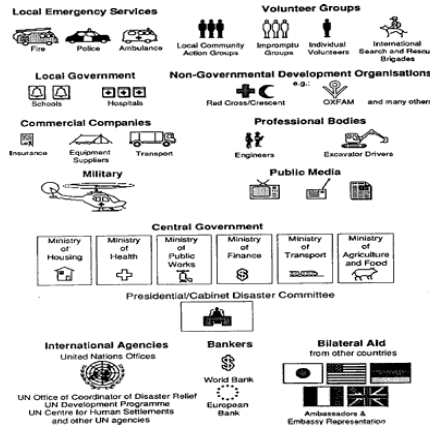


Fig. 4. The Organisation in Earthquake Disaster Model [30]

Evaluation against this model brings us to identify different ways to classify the *Aid* concept of DMM. A *Bilateral Aid* concept is found to not be covered in DMM. Thus, the *Aid* concept of DMM is split into three concepts: *HumanitarianAid*, *DevelopmentAid* and *BilateralAid*. The previous concepts that are used in previous DMM (*FoodAid*, *MedicalAid* and *RefugeeShelter*) were earlier grouped under *HumanitarianAid*. These changes are in the *Response*-phase of DMM. Following this, Table 2 shows the full list of Coburn model’s concepts as derived from concepts in DMM.

Table 2. Derive concepts in Coburn model by concepts in DMM

DMM	Coburn model
EmergencyManagementTeam	- Local emergency services: Fire, Police, Ambulance - International search and rescue brigade - Professional Bodies: Engineers, Excavator drivers, Military
Property	Local Government: School, Hospital
People	Volunteer Groups: Local community action groups, Impromptu groups
PreparednessTeam	Individual Volunteers
AidAgency	- Non-governmental development organisations: Red Cross/crescent - International Agencies: United Nations Office, Bankers: World Bank
Media	Public media
Insurance	Insurance
Resource	Commercial companies, Equipment Suppliers, Transport
Aid	Bilateral Aid from other countries: Ambassadors, Embassy representation
RecoveryOrganisation, RecoveryTask, RecoveryGoal	Department of Agriculture and Food, Farming Organisations, Private Owners and Consumers Organisations

3.3.2 Sample FBS 2: Against the Jennings Disaster Nursing Management Model (Model 9 of Set V2)

The Jennings Disaster Nursing Management model [32] presented in Figure 5 defines nursing during DM as “the systematic and flexible utilisation of knowledge and skills specific to disaster-related nursing, and the promotion of a wide range of activities to minimise the health hazards and life threatening damage caused by disasters in collaboration with other specialised fields”. The model aims to help community nurses plan for and manage disasters in hospitals. There are four phases incorporated in the model: Phase I (Pre-Disaster), Phase II (Disaster), Phase III (Post-Disaster), and Phase IV (Positive Client/Population Outcomes). This model is taken to validate DMM concepts with the activities presented by the Jennings model. DMM can successfully derive all concepts in the Jennings model. The pre-disaster stage which is the first phase Jennings used in her model is identified clearly and represents the mitigation and preparedness-phase of the DMM. However the Jennings model disaster phase represents DMM’s Response-phase and her post disaster represents DMM’s Recovery-phase. The DMM concepts used to generate the Jennings model are shown in Table 3.

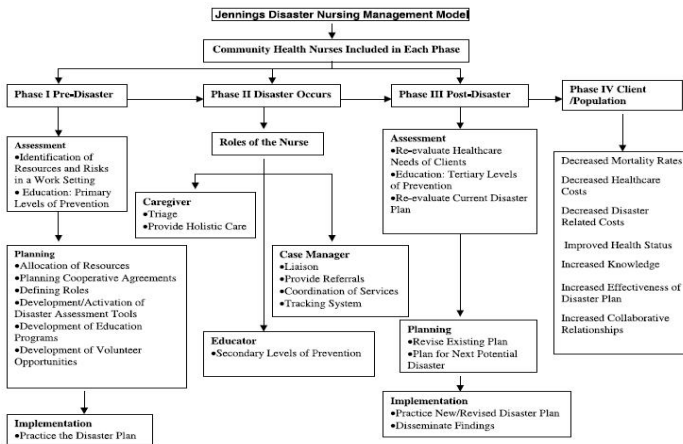


Fig. 5. Jennings Disaster Nursing Management Model [32]

Table 3. Jennings Concepts Support for DMM Concepts

JENNINGS Concepts		DMM Concepts	
Phase	Activity	Phase	Concept
Pre	Identification of Resources and Risks	Mitigation	RiskAnalysis, HazardAssessment,
	Education: Primary Level of Prevention	Preparedness	PublicEducation
	Allocation of Resource	Mitigation Preparedness Response Recovery	NeedsPlanning Pre-Position, SuppliesRegistry, Deployment, Demobilization
	Planning Cooperative Agreement	Mitigation	StrategicPlanningCommittee
	Defining Roles	Mitigation	StrategicPlanningCommittee
	Development/Activation of Disaster Assessment Tools	Preparedness	Monitoring
	Development of Education Programs	Preparedness	Public Education
	Development of Volunteer Opportunities	Preparedness	PreparednessPlan
	Practice the Disaster plan	Preparedness	Training

Table 3. (Continued)

Occur	Triage	Preparedness	Warning
	Provide Holistic Care	Response	ResponseTask
	Liaison	Response	EmergencyOperationCentre
	Provide Referrals	Response	StandardOperatingProcedure
	Coordination of Services	Response	Coordination, Command
	Tracking System	Response	EmergencyOperationCentre
	Secondary Level of Prevention	Response	Rescue
Post	Re-evaluate Health Care	Recovery	TaskReview
	Education: Tertiary Level of Prevention	Recovery	TaskReview
	Re-evaluate Current Disaster Plan	Recovery	LongTermPlanning
	Revise Existing Plan	Recovery	TaskReview
	Plan for Next Potential Disaster	Recovery	LongTermPlanning
	Disseminate Findings	Recovery	TaskReview

(Notes: MIT=Mitigation, PRE=Preparedness, RES=Response, REC=Recovery)

4 The FBS Evaluation Result

Table 4 and Table 5 respectively show the DoC values for all DMM concepts evaluated in the Mitigation and Response-phase class. The following five categories of concepts based on their DoC are defined: *Very Strong* (DoC result: 100 – 70 %), *Strong* (69 – 50 %), *Moderate* (49 – 30 %), *Mild* (29 – 11 %), and *Very Mild* (10 – 0 %). Very Strong DoC is assigned to concepts that appear frequently in Set V2 models, whereas Very Mild is at the other end of the scale. For example, the DMM concept, *MitigationPlan*, has a DoC value of 90%. It is expected that 90% of DM models with a mitigation phase will include it. It is also expected that 10% of DM models with a mitigation phase will not include it. For example, few models suggest forming a Strategic Planning Committee instead. Metamodel development is not about achieving perfection [34 pp. 23]. Aiming for a complete metamodel can lower its generalisability and has been cited as a common bad practice in metamodel development [34]. These views suggest that if a DMM concept is rarely used or needed, it may be better to delete it in some cases. As a result of this evaluation, concepts with zero DoC values are revisited and liable for deletion. For example, another DMM concept, *BuildingCode*, has a DoC value of 0 and is later revisited.

The DoC categorisation of all DMM concepts (for all four DMM classes including the Preparedness and the Recovery) is shown in Table 6: 19 concepts in DMM1.1 are categorised as ‘Very Strong’, 23 are ‘Strong’, 25 are ‘Moderate’, 13 are ‘Mild’ and 4 concepts are ‘Very Mild’. The four very mild concepts are *Property*, *NaturalSite*, *BuildingCodes* and *Land-UsePlanning*. Including them in DMM requires a reassessment. *BuildingCodes* and *Land-UsePlanning* are deleted as they are deemed as too specific to one kind of disaster (Bushfires). By revisiting DMM, it is found that the *StructuralMitigation* is in fact more generic to represent the *BuildingCodes* and *Land-UsePlanning*. As for the other two (*Property* and *NaturalSite*), they are opted to be kept as they are common across varying disasters.

The changes made to DMM1.1 here are affecting only the Mitigation-phase and Response-phase classes of concepts. Preparedness and Recovery-phase classes of concepts of DMM1.1 do not change here. New extension to the terminology is used to define three new concepts in the Response-phase class:

- 1) *HumanitarianAid* - Material or logistical assistance provided for humanitarian purposes, typically in response to an event or series of events which represents a critical threat to the health, safety, security or wellbeing of a community or other large groups of people, usually over a wide area.

Table 4. Frequency result of *Mitigation*-phase concepts

DMML1 Mitigation Concepts		Model Set V2										Concept Frequency
		1	2	3	4	5	6	7	8	9	10	
1	MitigationPlan	*	*	*	*	*	*	*	*	*	*	9
2	MitigationOrganisation	*	*	*	*	*	*	*	*	*	*	10
3	MitigationTask	*	*	*	*	*				*	*	7
4	NeedsPlanning	*		*		*				*		4
5	InformationUpdates				*	*			*	*		4
6	MitigationGoal			*		*	*		*	*		5
7	RiskReduction	*	*	*	*	*	*		*	*	*	9
8	People					*		*	*	*		4
9	Property											0
10	Lifeline		*		*	*						3
11	NaturalSite			*								1
12	HazardAssessment		*	*	*		*	*	*	*	*	8
13	RiskAnalysis	*	*	*	*	*	*		*	*	*	9
14	StructuralMitigation		*							*		2
15	Non-Structural Mitigation		*							*		2
16	Vulnerability		*			*			*	*		4
17	DisasterRisk	*	*	*			*		*	*		6
18	StrategicPlanningOrganisation	*		*	*		*		*	*	*	6
19	BuildingCodes											0
20	Legislation		*			*						2
21	Land-UsePlanning											0
22	Insurance							*				1

Table 5. Frequency result of *Response*-phase concepts

DMML1 Response Concepts		Model Set V2										Concept Frequency
		1	2	3	4	5	6	7	8	9	10	
1	EmergencyPlan	*	*	*	*	*	*	*	*	*	*	10
2	ResponseOrganisation	*	*	*	*	*	*	*	*	*	*	10
3	ResponseTask	*	*	*		*				*	*	6
4	Deployment		*			*		*		*		4
5	SituationalAwareness		*			*						2
6	ResponseGoal			*		*	*		*	*		5
7	Rescue	*				*		*		*	*	4
8	Disaster	*	*		*				*			4
9	SituationAnalysis		*		*	*			*			4
10	Incident								*			1
11	Coordination		*		*	*		*		*		5
12	Command		*			*						2
13	Communication		*			*		*		*		4
14	StandardOperatingProcedure				*					*		2
15	Victim									*	*	0
16	EmergencyManagementTeam	*	*		*	*		*		*	*	7
17	EmergencyOperationCentre		*					*				2
19	Aid				*	*		*		*	*	4
20	InformationManagement		*		*	*		*		*		5
22	RefugeeShelter				*					*		2
23	MassCasualtyManagement				*	*			*	*		2
24	FoodAid				*			*	*	*	*	4
25	MedicalAid		*		*	*		*	*	*	*	6

Table 6. Degree of Confidence of DMM Concepts after FBS

DoC Classification	DMM Concepts
100 – 70 % (Very Strong)	MitigationPlan, MitigationOrganisation, MitigationTask, RiskReduction, Resilience, HazardAssessment, RiskAnalysis, PreparednessPlan, PreparednessOrganisation, EmergencyPublicInformation, ResponseOrganisation, RecoveryPlan, Reconstruction, EmergencyManagementTeam, EmergencyPlan, RecoveryOrganisation, RecoveryTask, DamageAssessment, MentalHealthRecovery
69 – 50 % (Strong)	MitigationGoal, DisasterRisk, StrategicPlanningOrganisation, PreparednessTask, Warning, PreparednessGoal, Evacuation, BeforeDisaster, DisasterFactor, Training, Media, PublicAwareness, Resource, Monitoring, ResponseTask, ResponseGoal, Coordination, InformationManagement, MedicalAid (modify) , RecoveryGoal, After-Disaster, EconomicRestoration, FinancialAssistance
49 – 30 % (Moderate)	NeedsPlanning, InformationUpdates, People, Lifeline, Vulnerability, Event, Effect, SuppliesRegistry, DecisionMaking, Administration, Pre-Position, PublicEducation, AidAgency, Deployment, Rescue, Disaster, SituationAnalysis, Communication, Aid, FoodAid (modify) , Demobilization, LongTermPlanning, TaskReview, Exposure AidDistribution,
29 – 11 % (Mild)	StructuralMitigation, Non-Structural Mitigation, Legislation, Insurance, Victim, MutualAidAgreement, SituationAwareness, Command, MassCasualtyManagement StandardOperatingProcedure, EmergencyOperationCentre, Incident, RefugeeShelter (modify) ,
10 - 0% (Very Mild)	Property (•), NaturalSite (•), BuildingCodes (x), Land-UsePlanning (x)

(Legend: (modify) = modification is made to the concept, (•) = Keep the concept, (x) = Delete the concept)

- 2) *DevelopmentAid* - Aid to support the economic, environmental, social and political development of developing countries.
- 3) *BilateralAid* - Aid or funds that are given to one country from another.

Since two concepts (*BuildingCode* and *Land-UsePlanning*) have been deleted in the second evaluation (Figure 6), the association relationships of '*isAGroupOf*' owned by these concepts (in DMM1.1) are also deleted. The new version, DMM1.2, incorporates these changes as shown in Figures 6 (Mitigation-phase class) and 7 (Response-phase class).

5 Conclusion

In this paper, the evaluation of the Disaster Management Metamodel (DMM) is undertaken using the Frequency-based Selection (FBS) technique. To perform the FBS evaluation, a set of 10 DM models is formed as a validation set (based on wider coverage to provide overlaps and to enable a frequency count of the individual DMM concepts). As a result from this evaluation, 3 concepts (*HumanitarianAid*, *BilateralAid* and *DevelopmentAid*) are added and 2 concepts are deleted (*BuildingCode* and *Land-UsePlanning*) from DMM. These changes are realised in DMM1.2. In addition, two concept relationships (aggregation - '*isAGroupOf*') are also been deleted. After performing the evaluation, the objective to evaluate the

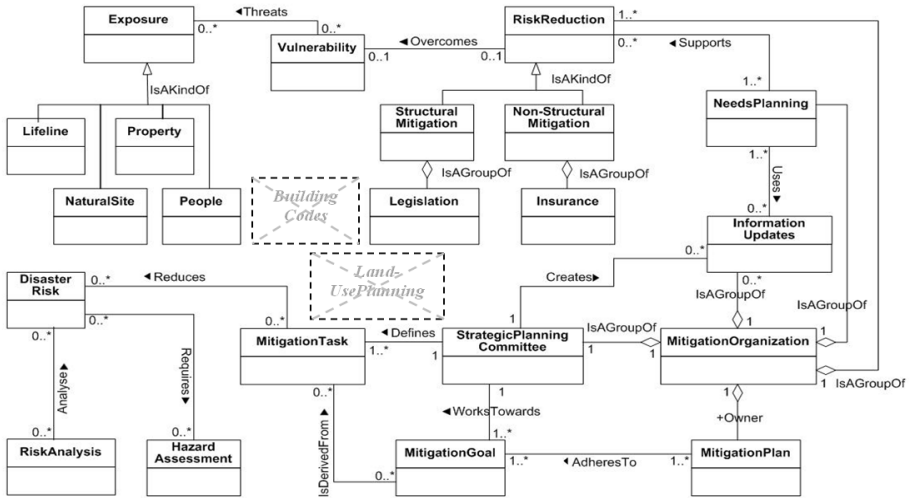


Fig. 6. The DMM1.2: A validated version of *Mitigation*-phase class of concepts

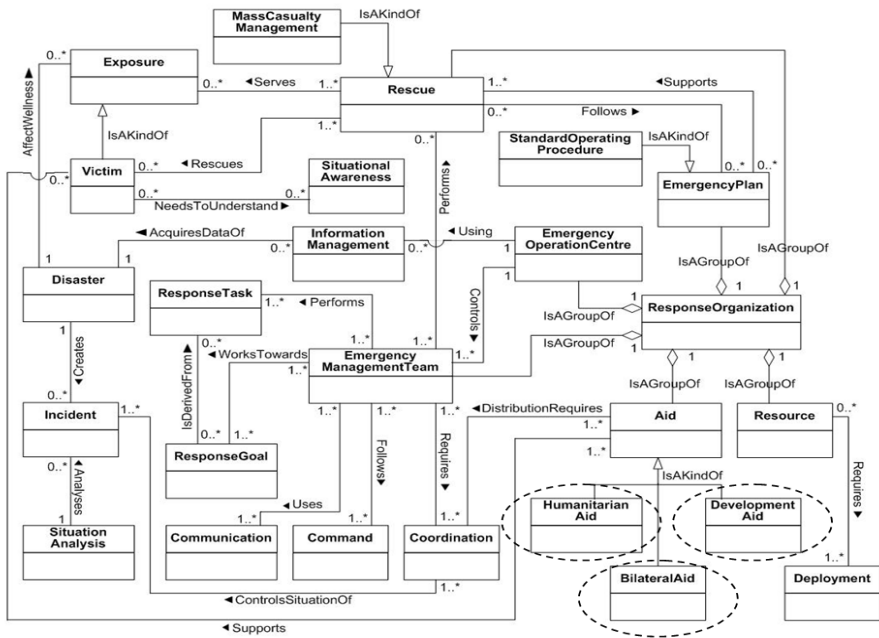


Fig. 7. The DMM1.2: A validated version of *Response*-phase class of concepts

importance of the individual concepts used in each phase class of DMM is achieved. DMM is further improved. Its expressiveness and completeness of its concepts are enhanced. By deploying a proven evaluation method from the knowledge based community to metamodeling as used by software engineers, the paper makes an

original contribution to both the KB and the SE communities. The use of metamodelling has hitherto being characterised by lack of systematic evaluation. By focussing on the evaluation and the metamodelling process on disaster management, this paper makes a significant contribution to this important domain unifying key concepts into a metamodel that can be used as knowledge sharing platform. Future works will develop a system prototype to demonstrate the applicability of the metamodel (DMM) in real world scenarios of disaster management.

References

1. Beydoun, G., Lopez-Lorca, A.A., Sanchez, F.G., Martinez-Bejar, R.: How do we measure and improve the quality of a hierarchical ontology? *J. Syst. Softw.* 84, 2363–2373 (2011)
2. Sprinkle, J.M.: *Metamodel Driven Model Migration*. Doctor of Philosophy, 176 pages. Vanderbilt University, Tennessee (2003)
3. Gharehdaghi, A.: *Design of a Generic Metamodel for Fieldwork Data Management*. International Institute for Geo-Information, Science and Earth Observation. Master Theses, Netherlands, Enschede (2003)
4. Beydoun, G.: Formal concept analysis for an e-learning semantic web. *Expert Syst. Appl.* 36, 10952–10961 (2009)
5. Norris, D.M., Jon Mason, R.R., Lefrere, P., Collier, G.: *A Revolution In Knowledge Sharing*. EDUCAUSE Review (September, October 2003)
6. Smith, W., Dowell, J.: A case study of co-ordinative decision-making in disaster management. *Ergonomics* 43, 1153–1166 (2000)
7. Beerli, A.J., Falk, S., Diemers, D.: *Knowledge Management and Networked Environments: Leveraging Intellectual Capital in Virtual Business Communities*. AMACOM Books, New York (2003)
8. OMG: *Unified Modelling Language Infrastructure Specification, Version 2.0*, OMG document ptc/03-09-15.8,10,12, 13 (2004)
9. Trabelsi, C., Atitallah, R.B., Meftali, S., Dekeyser, J.-L., Jemai, A.: *A Model-Driven Approach for Hybrid Power Estimation in Embedded Systems Design*. *EURASIP Journal on Embedded Systems*, 15 (2011)
10. Gary Berg-Cross, E.S.T., Rebecca Curzon, I., Chamindra de Silva, L.S.F., Paola Di Maio, U.o.S., Cutter Consortium, Renato Iannella, N., Mandana Sotoodeh, U.o.B.C., Olle Olsson, S., Guido Vetere, I.I.: *W3C Incubator Group* (2008), <http://www.w3.org/2005/Incubator/eiif/XGR-Framework-20090806/#ack>
11. Garcia, P.B.: *A Metamodel To Annotate Knowledge Based Engineering Codes As Enterprise Knowledge Resources*. PhD, pp. 489. Cranfield University (2007)
12. Lagerstrom, R., Johnson, P., Hook, D.: *Architecture analysis of enterprise systems modifiability - Models, analysis, and validation*. *Journal of Systems and Software* 83, 1387–1403 (2010)
13. Vytautas Stuiikys, R.D., Aleksandras, T.: *A Model-Driven View To Meta-Program Development Process*. *Information Technology and Control* 39 (2010)
14. Falkenberg, E.D., Hesse, W., Lindgreen, P., Nilsson, B.E., Oei, J.L.H., Rolland, C., Stamper, R.K., Assche, F.J.M.V., Verrijn-Stuart, A.A., Voss, K.: *A Framework of Information System Concepts*, The FRISCO Report. University of Leiden (1998)

15. Gardner, T., Griffin, C., Koehler, J., Hauser, R.: A review of OMG MOF 2.0 Query / Views / Transformations submissions and recommendations towards the final standard. In: Workshop on Metamodeling for MDA, pp. 179–197 (2003)
16. Bieman, R.F.a.J.M.: Multi-View Software Evolution: A UML-based Framework for Evolving Object-Oriented Software. In: Proceedings International Conference on Software Maintenance, ICSM 2001 (2001)
17. Beydoun, G., Low, G., Henderson-Sellers, B., Mouraditis, H., Sanz, J.J.G., Pavon, J., Gonzales-Perez, C.: FAML: A Generic Metamodel for MAS Development. *IEEE Transactions on Software Engineering* 35, 841–863 (2009)
18. Beydoun, G., Gonzalez-Perez, C., Henderson-Sellers, B., Low, G.: Developing and Evaluating a Generic Metamodel for MAS Work Products. In: Garcia, A., Choren, R., Lucena, C., Giorgini, P., Holvoet, T., Romanovsky, A. (eds.) SELMAS 2005. LNCS, vol. 3914, pp. 126–142. Springer, Heidelberg (2006)
19. Othman, S.H., Beydoun, G.: Metamodelling Approach To Support Disaster Management Knowledge Sharing. In: Proceeding Australasian Conference on Information Systems (ACIS 2010), Brisbane, Australia, Paper 97 (2010)
20. Othman, S.H., Beydoun, G.: A Disaster Management Metamodel (DMM) Validated. In: Kang, B.-H., Richards, D. (eds.) PKAW 2010. LNCS (LNAI), vol. 6232, pp. 111–125. Springer, Heidelberg (2010)
21. Kok, D.d.: Feature Selection for Fluency Ranking. In: Proceedings of the Sixth International Natural Language Generation Conference, INLG 2010 (2010)
22. Zhang, Z., Ye, N.: Locality preserving multimodal discriminative learning for supervised feature selection. *Knowledge and Information Systems* 27, 473–490 (2011)
23. Davis, P.K., Bigelow, J.H.: Motivated Metamodels. In: Proceedings of the 2002 Performance Metrics for Intelligent Systems Workshop, PerMIS 2002 (2002)
24. Ahmed, I.: Disaster Risk Management Framework. In: International Training Workshop on Disaster Risk & Environmental Management (2008)
25. Rosenberg, C.(FEMA): Chapter 3: Policies for Guiding Planning for Post-Disaster Recovery and Reconstruction. Appear in PAS Report No. 483/484 by American Planning Association (2005)
26. Garatwa, W., Bollin, C.: Disaster Risk Management: Working Concept (2002)
27. Ulrich Boes, U.L.: Disaster Information, Innovative Disaster Information Services (2008)
28. Buford, J.F., Jakobson, G., Lewis, L.: Multi-Agent Situation Management for Supporting Large-Scale Disaster Relief Operations. *International Journal of Intelligent Control and Systems* 11, 284–295 (2006)
29. Mendis, D.S.K., Karunananda, A.S., Samaratunga, U., Ratnayake, U.: An Approach to the Development of Commonsense Knowledge Modeling Systems for Disaster Management. *Artificial Intelligence Review* 28, 179–196 (2007)
30. Coburn, A., Spencer, R.: Earthquake Protection. Wiley Inc., Chichester-New York (1992)
31. Shaluf, I.M.: Technological Disaster Stages and Management. *Journal of Disaster Prevention and Management* 17, 114–126 (2008)
32. Jennings-Sanders, A.: Teaching disaster nursing by utilizing the Jennings Disaster Nursing Management Model. *Nurse Education in Practice* 4, 69–76 (2004)
33. Khan, H., Vasilescu, L.G., Khan, A.: Disaster Management Cycle - A Theoretical Approach. *Journal of Management and Marketing* 6, 43–50 (2008)
34. Kelly, S., Pohjonen, R.: Worst Practices for Domain-Specific Modeling. *IEEE Software* 26, 22–29 (2009)

Perceptual Evaluation of Automatic 2.5D Cartoon Modelling

Fengqi An, Xiongcai Cai, and Arcot Sowmya

School of Computer Science & Engineering,
The University of New South Wales
Kensington, NSW 2052, Australia
{fan,xcai,sowmya}@cse.unsw.edu.au
<http://www.cse.unsw.edu.au>

Abstract. 2.5D cartoon modelling is a recently proposed technique for modelling 2D cartoons in 3D, and enables 2D cartoons to be rotated and viewed in 3D. Automatic modelling is essential to efficiently create 2.5D cartoon models. Previous approaches to 2.5D modelling are based on manual 2D drawings by artists, which are inefficient and labour intensive. We recently proposed an automatic framework, known as Automatic 2.5D Cartoon Modelling (*Auto-2CM*). When building 2.5D models using Auto-2CM, the performance of different algorithm configurations on different kinds of objects may vary in different applications. The aim of perceptual evaluation is to investigate algorithm selection, i.e. selecting algorithm components for specific objects to improve the performance of Auto-2CM. This paper presents experimental results on different algorithms and recommends best practice for Auto-2CM.

Keywords: 2.5D, cartoon, modelling.

1 Introduction

The argument about 2D versus 3D cartoons has lasted for decades. Those who prefer 3D argue that “2D is dead”, and Hollywood has abandoned 2D feature animation with the success of Pixar’s 3D animations [6]. On the other hand, those who prefer 2D believe that 3D techniques restrict the imagination of artists. Nowadays many animations use both 2D and 3D graphics in order to balance their advantages and disadvantages. Some recent work on cartoons provides 3D rotation for 2D elements, called *2.5D cartoons*, which mix the rotation ability and reusability of 3D, and the unreal 3D impossible shapes of 2D. Di Fiore et al. [8] proposed an approach for animation production that generates key frames by interpolating hand-drawn views. Bourguignon et al. [4] presented another approach using 2D strokes manually drawn on 3D planes. However these two methods give the artists less freedom than the 2.5D Cartoon Models recently presented by Rivers et al. [11]. The 2.5D cartoon model is a novel approach to render 2D cartoons in 3D. However, Rivers’ 2.5D model is created purely by manual means, until the *Automatic 2.5D cartoon modelling (Auto-2CM)* [2] was

introduced. Auto-2CM aims to build 2.5D cartoon models automatically from real world objects, so that human artists can be released from labour intensive and repetitive work, and are able to focus on more creative aspects, see Fig. 1.

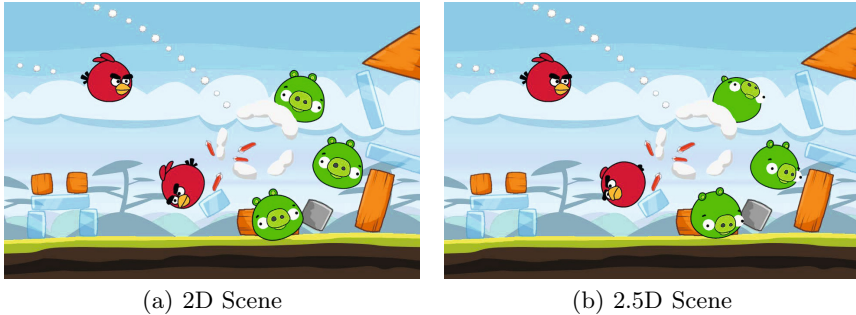


Fig. 1. Contrast of 2D and 2.5D scene of Angry Birds, where 2.5D characters can do 3D rotation. 2.5D cartoon models built using Auto-2CM [2].

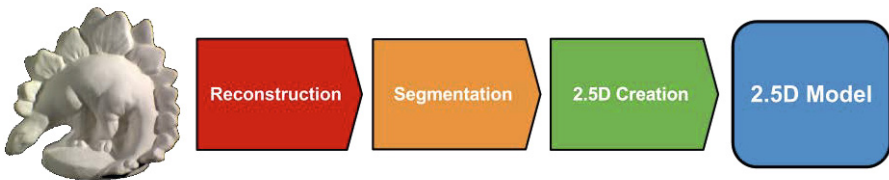


Fig. 2. Auto-2CM approach, different algorithms are used at each step

The 2.5D models by Rivers et al. are able to rotate in 3D space while maintaining 2D stylistic elements. However, their technique is not suitable for representing objects with long thin shapes such as tables and chairs, and these kind of shapes may lead to wrong shapes when rendering. This limitation can not be overcome unless the simple interpolation method used by Rivers et al. is improved, which is a problem beyond the scope of this paper. Automatic 2.5D Cartoon Modelling (Auto-2CM) [2] is a recent novel approach for creating 2.5D Cartoon Models [11] automatically. It provides artists with the option to build 2.5D sketches faster and more easily. The framework consists of several steps as shown in Fig. 2. The reconstruction step may be omitted when a 3D model of the candidate object exists. Different 3D segmentation methods can be used in the segmentation step. However, when building 2.5D models using the Automatic 2.5D Cartoon Modelling approach, the performance of different algorithm configurations and for different objects may vary for different applications. The aim of perceptual evaluation is to investigate algorithm selection, i.e. selecting appropriate algorithm components for specific kinds of objects in order to improve the performance of Auto-2CM. The goal of this evaluation is to find the best practice for Auto-2CM, and the motivations are:

- (i) evaluation of different component combinations of Auto-2CM,
- (ii) assessment of advantages and limitations of different combinations,
- (iii) deeper understanding of the examined field.

Currently, no previous experimental study of this area exists. Finding a method to quantitatively evaluate 2.5D cartoons, which is a novel type of visual art, is a very difficult task. Perceptual evaluation is widely used to evaluate results that are not suitable for quantitative evaluation, such as those related to visual art or audio [5, 9, 10]. In this paper, a perceptual evaluation of the Auto-2CM approach is presented via a series of experiments. The hypotheses of this evaluation are: (i) that certain component combinations perform differently on different kinds of objects; (ii) the approach can produce better results by selecting different combinations for specific kinds of objects. This is facilitated because the Auto-2CM approach does not require specific methods for segmentation, and the system still works with different algorithms at each stage.

Different configurations of the approach on different types of objects are tested, and their advantages and disadvantages discussed. Recommendations on suitable configurations for specific kinds of objects are also provided.

This paper is organized as follows. The data used in this experiment, and the reasons they are selected are discussed in Section 2. The design of the experiments is discussed in Section 3. Results and analysis are in Section 4. Recommendations and suggestions for best practice of Auto-2CM are in Section 5. Finally the conclusion is in Section 6.

2 Data Sets

Assuming 3D models already exist that Auto-2CM can start from, the data sets used in this evaluation are from several different sources, and fall into two main categories. The first category contains models from different 3D mesh segmentation benchmarks currently in use, called *scientific models*. These could provide information on how 3D segmentation affects 2.5D modelling. This is discussed in more detail in Section 2.1. The second category includes handmade models suitable for 3D video games, called *industry models*. Some of them are created using Rivers’ models as reference, others are from games such as *Angry Birds*¹ and *Ruby Run*². These models are most suitable as 2.5D models, and the best models for Auto-2CM evaluation, and are discussed in more detail in Section 2.2.

2.1 Scientific Models

The reasons to include these models are to test the performance of different segmentation approaches in experiments. The relationship between the results of general 3D segmentation and 2.5D modelling may be tested, to examine the influence of general 3D segmentation on the performance of 2.5D modeling.

¹ <http://angrybirds.com>

² http://zhanstudio.net/games/ruby_run

Secondly, models selected for this experiment are those that are better candidates for 2.5D modelling than others in the benchmarks. Those that may cause error shapes according to Rivers’ limitations, no matter which algorithm is used, are avoided.

This category contains public shape benchmarks. These benchmarks were designed for the purpose of evaluating different segmentation methods, and include: (i) SHREC [14], (ii) McGill 3D Shape Benchmark [15], (iii) Princeton Shape Benchmark [13] and (iv) 3D Shape Segmentation Benchmark [7].

2.2 Industry Models

These models came from three different sources. The first two models, Simple-Head (Fig 3(a)) and Alien (Fig 3(b)), were built manually based on Rivers’ work. These two models are included in order to make it easy to compare with original manually created 2.5D cartoon models. These 2.5D models were built to demonstrate the usability of the 2.5D Cartoon Models for industrial animations. The next two characters, Pig (Fig 3(c)) and Bird (Fig 3(d)), are from the popular video game *Angry Birds*, and the last model Ruby (Fig 3(e)) is from another game *Ruby Run*. These models are good to test the performance of different processes aimed at real industry. Current 3D video game models are normally ‘low-poly’ models, which means there is a limit to the number of polygons in the model.

Models in this category are designed to evaluate the usability of the Auto-2CM approach for industry. They are created such that they do not violate Rivers’ limitations, thus guaranteeing that any error in the final 2.5D Cartoon Models is not caused by these 3D models.

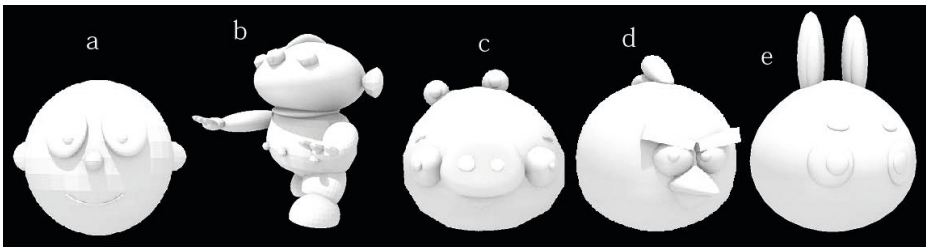


Fig. 3. Industry models used for 2.5D Cartoon Modelling

3 Design

The experiments were designed to test the performance of different algorithm combinations of Auto-2CM on different kinds of objects. In order to achieve this goal, three segmentation algorithms (PBS, SDF and FP) and two categories of models are used.

The three segmentation algorithms deployed are:

- (i) ProtBasedSeg (PBS) [1],
- (ii) Shape Diameter Function (SDF) [12],
- (iii) Fitting Primitive (FP) [3].

Results and analysis of these experiments are in Section 4.

The 2.5D model building system Auto-2CM was used in the experiments. This system takes segmented 3D meshes as inputs. As it just captures shapes of segmented parts from views, no elements that violate Rivers’ limitations will be created at this step.

Following the perceptual evaluation methods of previous works, especially the cartoon related research by Garcia et al. [9], the evaluation of this research examines the following two aspects:

- (i) **errors:** shapes of final 2.5D model that lead to incorrect interpolation results;
- (ii) **appearance:** perceptual judgment of appearance.

Some models from public shape benchmarks contain some parts that almost always fail Rivers’ limitations. This is unavoidable and can only be fixed by improving the 2.5D Cartoon Model structure itself, which is a task beyond the topic of this paper. Such errors will be ignored in this experiment since they are caused by a rendering system outside the modelling system.

4 Results and Analysis

The final results of the experiments are the 2.5D Cartoon models built by different algorithm combinations. Before evaluating the quality of 2.5D Modelling, it is necessary to evaluate the intermediate product, namely the 3D segmented meshes, to determine how segmentation influences the whole process. The first half of this section is analysis of the 3D segmentation results of different segmentation methods. The second half is analysis of the 2.5D results.

4.1 Segmentation Results and Analysis

Three segmentation approaches, PBS, SDF and FP were tested on all models, as listed in Section 3.

Sunglasses: The three methods provide similar results. SDF unnecessarily separated the tips of frames. See Fig. 4(a)(b)(c).

Table: PBS and FP provide the best results, as in Fig. 4(d)(f). SDF focused on details, however this will not cause errors in the final modelling. See Fig. 4(e).

Bear: PBS gives a fair result, but it ignores the details on the head (ears) which will lead to an faulty 2.5D shape, shown in Fig. 7(e). PBS also produced a meaningless part at the bottom of the bear. SDF is the best. FP divided the body into two parts, and the border is uneven, which will lead to meaningless 2.5D shapes.

Dolphin: SDF is better than PBS in this case, as it successfully separated the fins. FP failed. See Fig. 4(j)(k)(l).

Gull: Both SDF and PBS are acceptable. FP successfully segmented the wings, but failed at the body part. See Fig. 5(a)(b)(c).

Teapot: The result of SDF is the only one that can produce a good 2.5D model, as in Fig. 7(l). Result for PBS in Fig. 5(d) will lead to unrecognisable shapes, as shown in Fig. 7(k).

Cow: SDF is the best. PBS is acceptable except for an unsegmented leg. See Fig. 5(g)(h).

Camel: The problem of PBS in this case is the same as for the Cow model, namely failure to segment a main part of the model that will lead to errors in a later step. In this case, the head of Camel will be missing in final 2.5D models, as shown in Fig. 7(o). SDF provides acceptable results, as in Fig. 5(k).

To conclude, of the three segmentation algorithms on public shape benchmark models shown in Figs. 4 and 5, FP and PBS are good at simpler and more obvious cases, such as the first two models. But FP can hardly provide any useful results beyond that. PBS generally produced acceptable segments except for more challenging models, such as the teapot, cow and camel. See Table 1 for a summary.

The results of industry models are shown in Fig. 6. These models share common features, namely, they have neither long stick shapes nor complex concave parts. Neither of PBS and FP can provide acceptable results for these models, while SDF achieves almost perfect results.

Table 1. Acceptable results(✓) and those may lead to 2.5D errors(✗) of segmentation

	Glasses	Table	Bear	Dolphin	Gull	Teapot	Cow	Camel	Head	Alien	Bird	Pig	Ruby
PBS	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
SDF	✓	✗	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓
FP	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗

4.2 2.5D Models and Analysis

The final 2.5D models are the most intuitive results for evaluation. FP was not tested for 2.5D models building, because the segmentation results of FP is worse than the other two methods, and is unlikely to lead to meaningful 2.5D models. Only the other two segmentation algorithms are tested in this step, namely PBS and SDF.

As in the previous step, the results are organized in two parts: results and analysis for scientific models, and for industry models.

Scientific Models. 2.5D models built from shape segmentation benchmark datasets are shown in Fig. 7.

For the first two models, Sunglasses and Table, PBS and SDF give almost similar results. They both look good, but wrong shapes at certain angles are

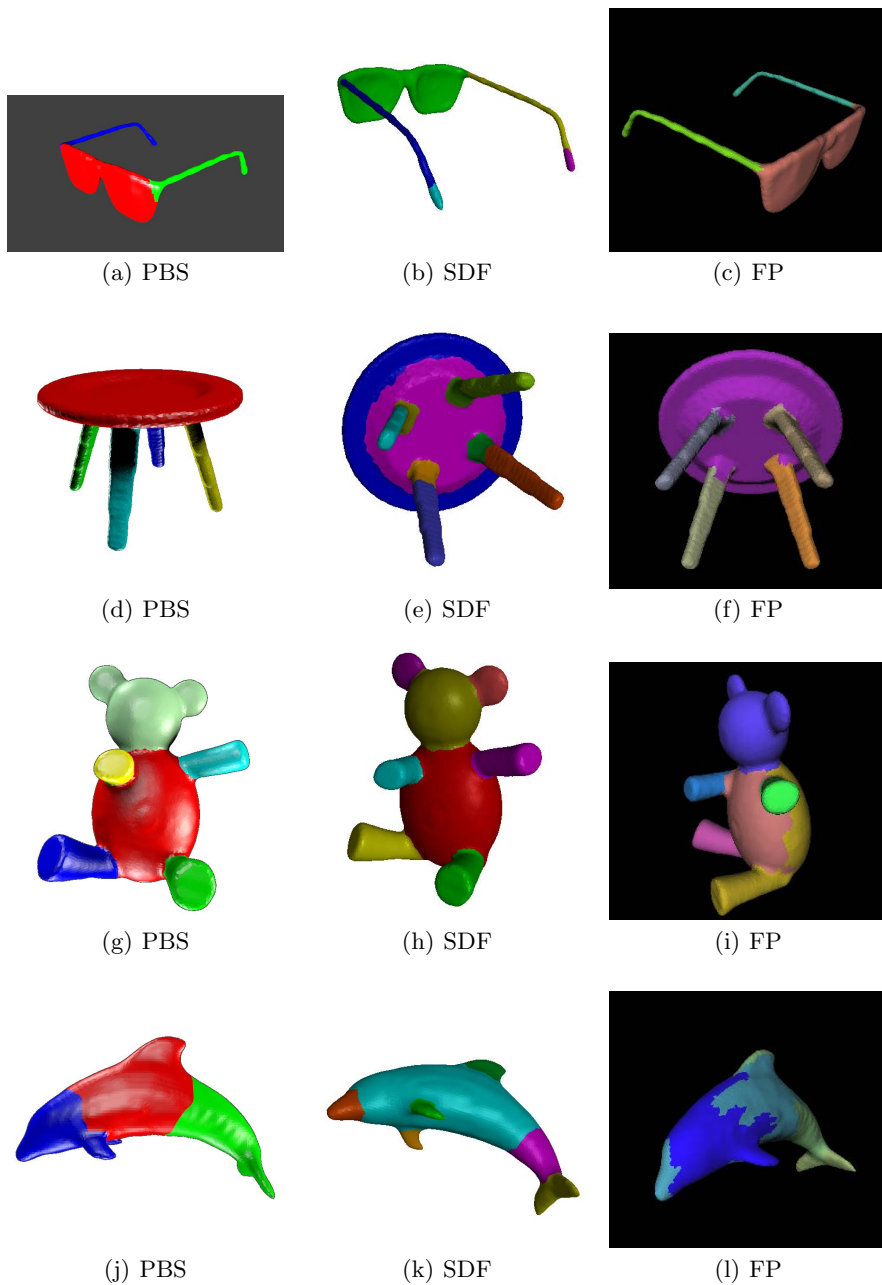


Fig. 4. Segmentation of models from shape segmentation benchmarks

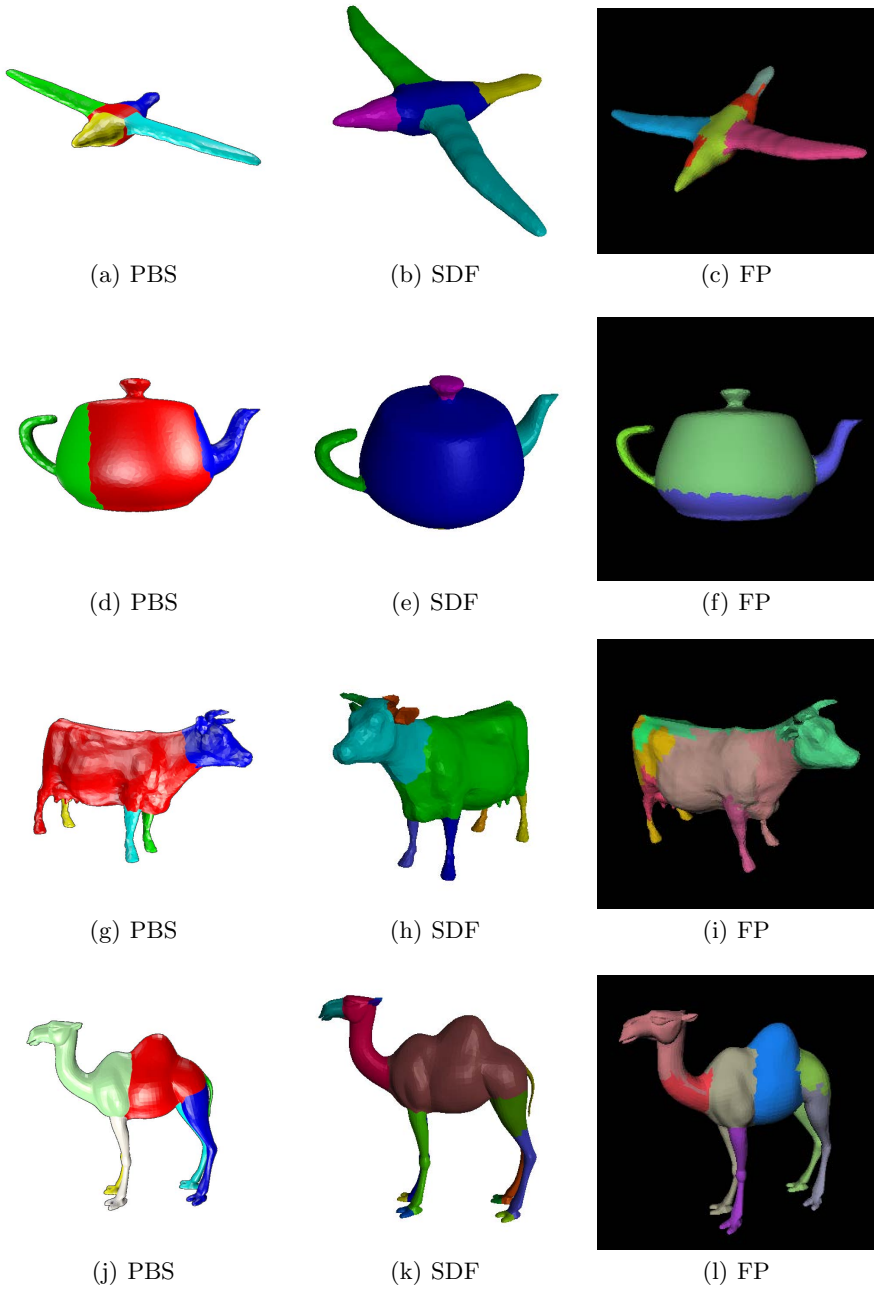


Fig. 5. Segmentation of models from shape segmentation benchmarks

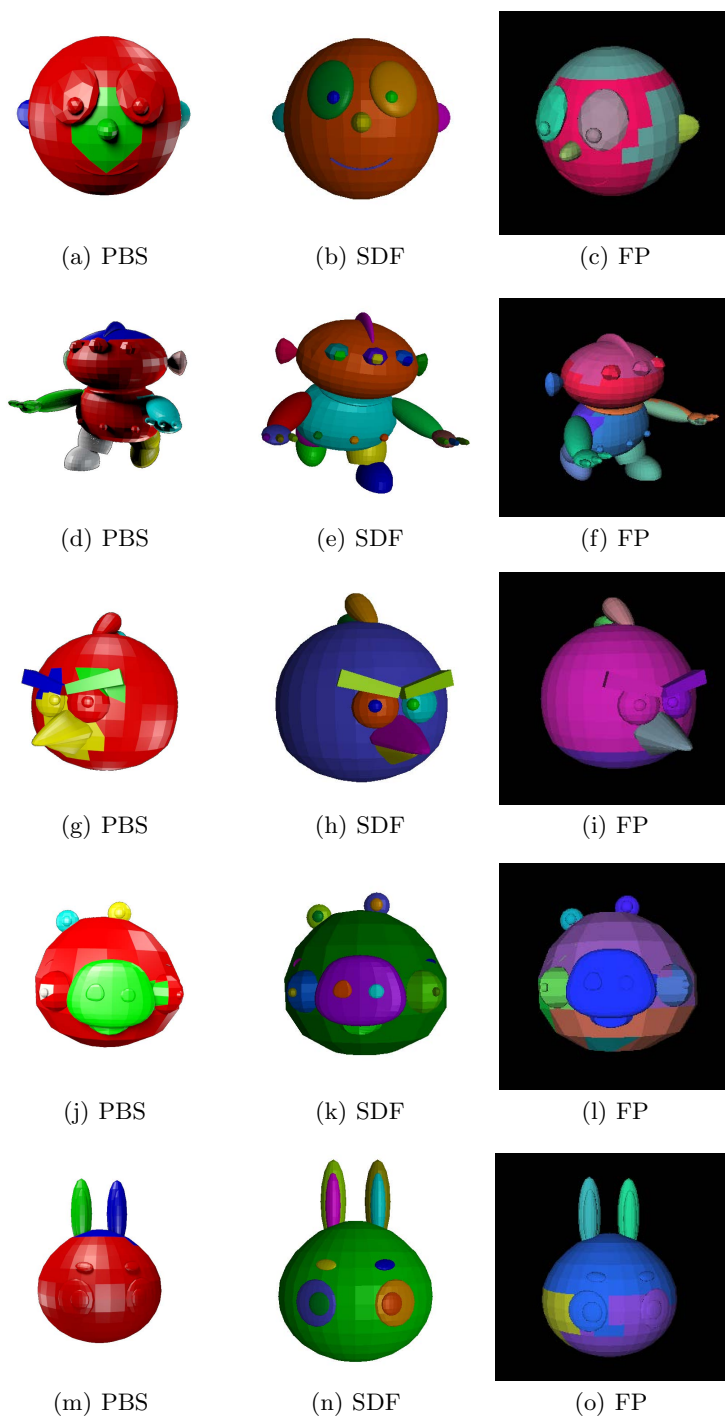


Fig. 6. Segmentation results of Head, Alien, Bird, Pig and Ruby

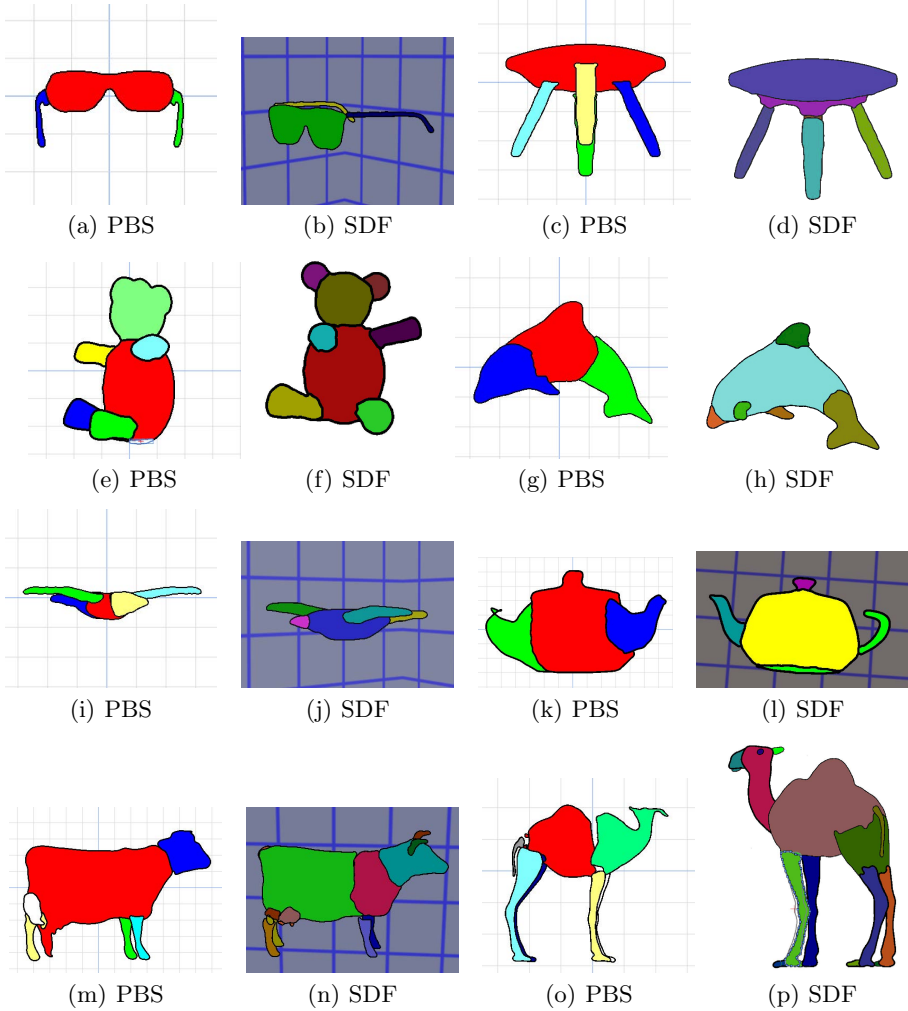


Fig. 7. 2.5D models from shape segmentation benchmark 3D models

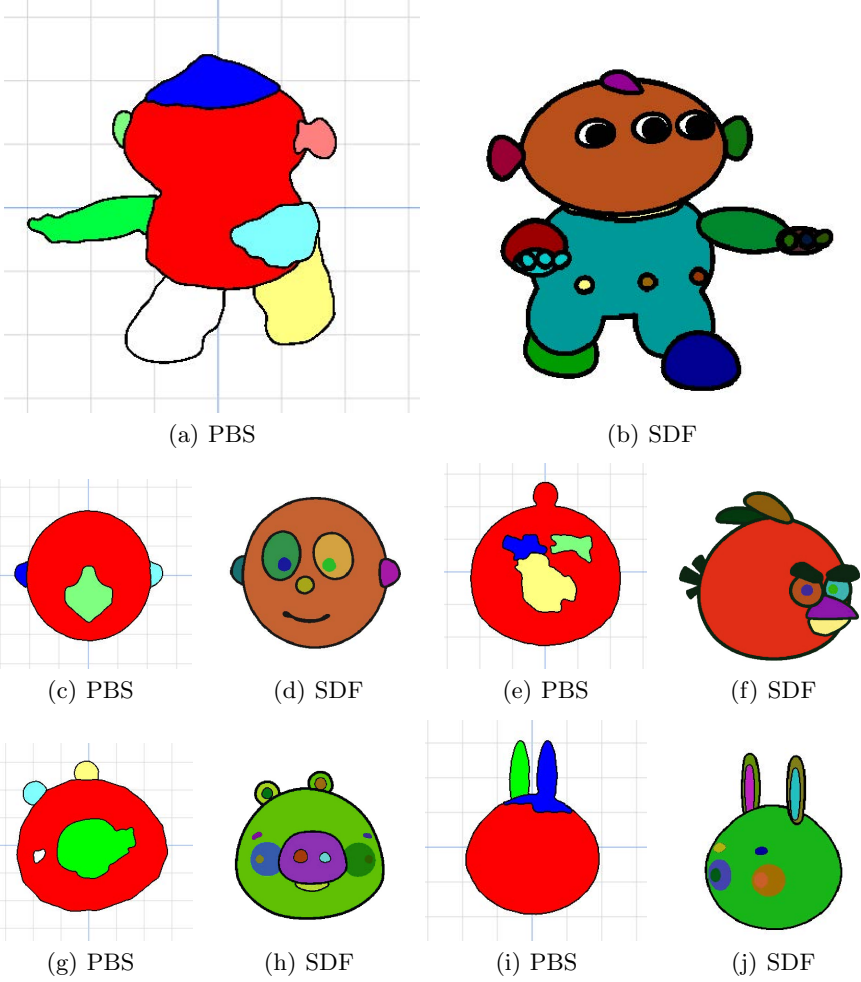


Fig. 8. 2.5D models built from industry models, based on the models by Rivers et al. as reference, and other games characters

caused by the long-thin stick part of the original 3D model. This is not caused by 3D segmentation but by the simple interpolation method of Rivers' approach.

Bear, Dolphin and Gull show differences caused by 3D segmentation methods. Bear built from PBS did not separate the ears, giving rise to an error at in-between angles. However, the extra part at the bottom of the Bear did not lead to serious faults. The Dolphin from SDF has its fins separated as shapes, and can be recognised when rotating. Dolphin from PBS, on the other hand, leaves the fins on the body, which will cause the fins to disappear at certain view angles and lead to wrong interpolation of the body part. It is still recognizable at most angles though. The two segmentation methods do not have much difference when dealing with the Gull, and both give good results.

The last group contains three more difficult models for automatic segmentation, Teapot, Cow and Camel. All 2.5D models built from PBS segmentations have at least one serious error. The handle of Teapot, right-rear leg of Cow and head of Camel are not recognizable. These errors are caused by the segmentation results not being suitable for 2.5D modelling, and it is also not the way that manual segmentation will cut the mesh. For example, manual segmentation will not cut the teapot handle from the middle of the body, but from the end of the handle like SDF. Manual segmentation will cut the leg as a different part of the body of the Cow, similar to SDF, and cutting the head of the Camel from the neck as shown in SDF is more reasonable than from the shoulder as in PBS.

Thus to conclude, a segmentation method that has better performance on a general 3D segmentation benchmark, when compared to manual segmentation, will also have better performance in the 2.5D Cartoon Modelling process. Because manually created 2.5D Cartoon Shapes rely on human artists to separate the object, it is reasonable that an automatic segmentation method that is most similar to human segmentation will give the best result.

Industry Models. The results of industry models are shown in Fig. 8. Referring to the segmentation results shown in Fig. 6, it is clear that PBS did not perform well on industry models, on the other hand SDF is almost perfect for these models. The good performance of SDF leaves the artist less manual work in modifying the model and makes it practically useful. The reason that PBS and SDF perform significantly differently on these models might be the difference precisely between industry models and scientific benchmark models.

One obvious difference between these industry models and the more research-oriented benchmark models is that the industry models should be able to perform real-time rendering with limited computing power. Industry models are often 'low-poly' models, having fewer triangles than scientific models. Low-poly industry models will affect the performance of different segmentation methods. For example, research benchmark meshes may have hundreds of triangles for a wrist, while the wrist of a game character may only have tens of triangles, which is sufficient for animation deformation, but leaves less information of the shape for the segmentation algorithms. However, scientific models often have redundant information caused by automatic 3D reconstruction, while industry

models are often handmade and therefore smoother. Both these factors affect the segmentation results, further affecting the 2.5D modelling performance.

A summary of results is shown in Table 2.

Table 2. Acceptable(✓) and error(✗) 2.5D results

	Glasses	Table	Bear	Dolphin	Gull	Teapot	Cow	Camel	Head	Alien	Bird	Pig	Ruby
PBS	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
SDF	✓	✗	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓

5 Recommendations

This section discusses the best practice for 2.5D cartoon modelling, which is the purpose of this evaluation. Based on the experiments, different kinds of models should be segmented using different methods to get the best results.

5.1 Scientific Models

For most simple shapes, such as the sunglasses and the table in Fig. 4 used in the experiments, where the boundary of segmentation are clear shape edges, Fitting Primitives (FP) gives the best results, followed by PBS, which gives similar performance to FP. SDF tends to segment simple models into more parts, and some segments are unnecessary and meaningless. However, models which have long thin shapes and sharp edges are not suitable to be presented as 2.5D cartoon models.

Other scientific models used in this experiment are not image-based reconstructed but scanned models. Both methods provide similar mesh models, which have high polygon count (usually more than 10k triangles per model). Moreover, automatically reconstructed meshes have irregular vertex positions, but their distance to each other is often even. For example, a rectangle may consist of not just two triangles, but many triangles that have similar areas.

When building from these reconstructed 3D models, SDF will provide the best quality results. In some cases PBS may have similar results, but overall SDF performs the best.

5.2 Industry Models

3D models made for games are low-poly models, for example Ruby in Fig. 6 (around 1k triangles if targeting mobile devices, 5-7k triangles if targeting next-gen game consoles). They are often handmade and normally no triangle is wasted, i.e. a rectangular plane will consist of only two triangles. Another difference between handmade models and reconstructed models is that the latter always use one mesh per object, but handmade models could have multiple meshes per object.

Based on the experiments, when building from low-poly game models, SDF performs the best for segmentation. Segments of SDF are almost perfect, while in contrast, results from the other two methods are not acceptable.

5.3 Summary

To build simple models with sharp edges, though which are not suitable for 2.5D cartoon models, FP should be used in the segmentation step. In the case of building reconstructed models, both PBS and SDF could be considered, and the user can run both and select the better one for the specific model under construction. For low-poly game models, SDF should be selected as it has the best performance on such models.

6 Conclusion

This research shows that choosing appropriate algorithms for the type of objects can improve the performance when building 2.5D cartoon models.

Based on the experiments, FP is good at segmenting simpler models that have long stick shapes, but does not give acceptable results on round shapes. Such models are not suited to 2.5D in any case. SDF is slightly worse than FP and PBS at simpler shapes, but good for almost all shapes that follow Rivers' conditions. Therefore, (i) for the simplest models with sharp edges, FP is the best choice; for other reconstructed (scientific) models, both PBS and SDF may be used; (ii) SDF is currently the best segmentation method for industry models, and it can lead to 2.5D models with least error and best appearance.

References

- [1] Perantonis, S., Agathos, A., Pratikakis, I., Sapidis, N.S.: Protrusion-oriented 3d mesh segmentation. *The Visual Computer* 26(1), 63–81 (2010)
- [2] An, F., Cai, X., Sowmya, A.: Automatic 2.5d cartoon modelling. *Image and Vision Computing New Zealand* 20, 149–154 (2011)
- [3] Attene, M., Falcidieno, B., Spagnuolo, M.: Hierarchical mesh segmentation based on fitting primitives. *The Visual Computer* 22(3), 181–193 (2006)
- [4] Bourguignon, D., Cani, M.P., Drettakis, G.: Drawing for illustration and annotation in 3d. *Computer Graphics Forum* 20, 114–123 (2001)
- [5] Čadík, M.: Perceptual evaluation of color-to-grayscale image conversions. *Computer Graphics Forum* 27(7), 1745–1754 (2008)
- [6] Cavallaro, D.: *The animé art of Hayao Miyazaki*. McFarland & Co. (2006)
- [7] Chen, X., Golovinskiy, A., Funkhouser, T.: A benchmark for 3d mesh segmentation. In: *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 28(3) (August 2009)
- [8] Di Fiore, F., Schaeken, P., Elens, K., Van Reeth, F.: Automatic in-betweening in computer assisted animation by exploiting 2.5d modelling techniques. In: *Computer Animation*, pp. 192–200 (2001)

- [9] Garcia, M., Dingliana, J., O’Sullivan, C.: Perceptual evaluation of cartoon physics: accuracy, attention, appeal. In: Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization, pp. 107–114 (2008)
- [10] Pražák, M., Hoyet, L., O’Sullivan, C.: Perceptual evaluation of footskate cleanup. In: Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 287–294 (2011)
- [11] Rivers, A., Igarashi, T., Durand, F.: 2.5d cartoon models. *ACM Transactions on Graphics* 29(4) (2010)
- [12] Shapira, L., Shamir, A., Cohen-Or, D.: Consistent mesh partitioning and skeletonisation using the shape diameter function. *The Visual Computer* 24(4), 249–259 (2008)
- [13] Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.: The princeton shape benchmark. In: Proceedings of Shape Modeling Applications, pp. 167–178 (2004)
- [14] ter Haarf, F., Veltkamp, R., Whitmarsh, T.: Shrec 2007 (2007), www.aimatshape.net/event/SHREC/
- [15] Zhang, J., Siddiqi, K., Macrini, D., Shokoufandeh, A., Dickinson, S.J.: Retrieving Articulated 3-D Models Using Medial Surfaces and Their Graph Spectra. In: Rangarajan, A., Vemuri, B.C., Yuille, A.L. (eds.) EMMCVPR 2005. LNCS, vol. 3757, pp. 285–300. Springer, Heidelberg (2005)

Automatic Acquisition of User Models of Interaction to Evaluate the Usability of Virtual Environments

Nader Hanna¹, Deborah Richards¹, and Michael J. Jacobson²

¹ Department of Computing
Macquarie University
NSW 2109, Australia

{nader.hanna,deborah.richards}@mq.edu.au

² Centre for Computer Supported Learning and Cognition
The University of Sydney
NSW 2006, Australia

michael.jacobson@sydney.edu.au

Abstract. Evaluation is an essential step in the research and development of software, particularly for new technologies such as Virtual Environments. One of the challenges to such evaluation is to collect data needed for analyzing the behavior of the users of the virtual environment. Conventional acquisition of evaluation data is time-consuming and prone to bias. In this paper, we present a taxonomy to assist identification and collection of appropriate variables for automatic data collection. We further show how these variables, such as navigation paths and characters met, can be used to capture the behavioral interaction of learners in a virtual ecosystem and to produce a user-model to evaluate the usability of the world under development.

Keywords: Virtual Environment, Automatic Data Acquisition, Usability Evaluation, User-model Data Collection.

1 Introduction

A 3D virtual world or Virtual Environment (VE) can only be considered successful if it is both usable and useful for its intended purpose. Usability measurement of VE could include ease of use and the usefulness of the system, in addition to quantifiable characteristics such as learnability, speed and accuracy of user performance in achieving tasks, low user error rate, and user satisfaction [1]. Due to the differences which exist in the ways that users interact and experience virtual worlds, and the increased complexity of these interactions compared to most desktop applications, like word processors or accounting packages, we propose that the data to be captured, the method of data capture and use of the data for evaluation purposes of VE may differ.

There is some research work that is concerned with usability in virtual environments (e.g. [2]). Typically these studies employ methods used for conventional systems to VE to collect data from which meaningful information about the usability of the system can be extracted. A common method to evaluate VE usability includes

direct evaluation by on-site evaluators who observe subjects using the VE in the laboratory. However, this procedure requires suitable test spaces and access to an adequate number of evaluators to suit the number of subjects. Additionally, this approach can be financially expensive and prohibitively costly in terms of the experts' time involved in conducting the evaluation and their transportation to the location of the experiment. Usability testing for VEs could be done remotely where human subjects and evaluators could be separated by time and space either manually through video-conferencing or through automatic data collection [3, 4]. Furthermore, it is well accepted that usability evaluation for any software should be done as early as possible and begin with developing a demonstration of the virtual system. Bowman et al. [5] note that keeping in mind usability from the very beginning stage in development process of VE, will make developers more likely to avoid creating unintuitive VE that do not match task requirements.

Another traditional method to collect data about VE usability issues is to ask the VE users to report if they have experienced any problems with usability or if some other system aspect did not fulfill their needs. While using self-reporting questionnaires is one of the most common methods to collect data, it has been shown to produce unreliable measures [6]. Self-reporting data can also be collected via interviews. However, these can be time consuming to conduct and, depending on the structure of the interview and the nature of the questions, capture of the data can be difficult. A particular problem with the use of observation and interview methods of data collection is the extensive effort required to later analyse the data. Analysis often involves costly transcription and coding schemes that are prone to bias due to reliance on interpretation of the original data by the coder/s.

To conduct studies of VEs it is critical to collect relevant participant data quickly and accurately [7]. De Leeuw and Nicholls [8] affirm the advantages of computer-assisted data collection over traditional methods. The capture of data using a virtual reality system which supports playback and simulation has great advantages compared to classical paper and pencil methods of data collection [9]. Some research handles automatic collection and analysis of data for standard GUI applications, e.g. [10], but very few efforts have been directed to VE automatic data collection.

To overcome the barriers of evaluating VE usability on-site and to support early formative evaluation of the system, an automated and remote usability evaluation method that uncovers a user model of interaction in the VE is introduced in this paper. After a review of relevant literature in this area and presentation of our data collection taxonomy (section 2), we present the VE in which we have developed and trialed our method (section 3) followed by the method itself (section 4) and results (section 5). Our conclusions and future work appear in section 6.

2 Literature Review and Taxonomy of Data Collection

Collecting data for the events and actions that take place in a VE may involve different approaches according to the purpose of the VE. Andreas et al. [11] divide their collaborative learning system into three different phases, and in each phase they use

various data collection methods to acquire data before, during and/or after a session. The data collection methods include initial and final questionnaires, text-chat log files, video recordings and interviews. To address the problem of the rapidly growing amount of user-generated social media data, Zhang et al. [7] developed a technical framework to demonstrate how to collect avatar-related data using a combination of bot- and spider-based approaches. The authors used the Second Life virtual social system to examine the differences in physical activity between male and female avatars and between young and old avatars. Similarly, Yee and Bailenson [12] presented a method that relied on bot-based avatars implemented in PHP and MYSQL to collect and store the longitudinal behavioral profiles of users involved in a Second Life social virtual reality.

Teixeira et al. [4] followed a user-centered design methodology to develop a virtual system called ErgoVR. One of the most important features of ErgoVR, of relevance to this paper, was the automatic data collection of variables of behavioral interaction such as: dislocation paths, trajectories, collisions with objects, orientation of the field of view and the occurrence of events activated by actions done by the user. Using ErgoVR software to automatically register the paths taken by the participants, Duarte et al. [13] showed how some features such as colors, windows, furniture, signage and corridor width may affect the way users select paths within a VE. The study aimed to determine if these factors could be considered as predictors for route selection.

While the majority of research focuses on collecting data from the virtual world, research has been conducted that collects data from the physical world via mobile devices. The data is then analyzed and used for augmenting the virtual worlds with real-life data. For example, Laaki et al. [14] first track the walking path of a human participant while listening to music and later use the data to simulate the trajectory on a map of a walking avatar listening to music.

From our review of the literature, we can classify data which has been collected in virtual worlds according to the following taxonomy:

- *What data is collected.* The type of data that can be collected is limitless. Data could be navigation information within the VE, actions achieved/attempted, text and chat archives, audio and video responses. Further examples of different types of data collected are mentioned in the other categories below.
- *Source/nature of the data (Physical/Virtual).* Data collected could include the behavior of the user in the physical world while using the virtual world, or the behavior of the user in the virtual world. For example, Grillon and Thalmann [15] present a method to track the physical movement of a human's eye while interacting with a virtual character. Eye-tracking data is used in order to determine whether a virtual character is being looked at so that it can adapt to appear more attentive during a public speaking exercise. White [16] presents a technique for classifying human motion through a virtual environment using Support Vector Machines and Kernel learning. The author makes use of the ease of use and flexibility of VE data collection to classify the human's motion while exploring a modified version of an open source video game (*Quake II*). The authors claim the classified virtual motion

could be scaled to multi-agent team behaviors, demonstrating the myriad of ways in which the data captured can be utilized (next point).

- *How the collected data can be used.* Collected data might be used to evaluate the usability of the virtual system, or to understand the behavior of the user. Sebok and Nystad [17] implement a design process to evaluate the usability of a virtual environment. Different types of data are collected. These include task completion measures, understanding of the radiation field, sense of presence, workload and usability ratings. Holm et al. [18] present a method to evaluate the usability of a VR-based safety training system called SAVE. The usability of the system including certain incident or usability problems encountered during the session was evaluated using data collected while testers used the virtual environment. Interaction and simulation data is acquired automatically and joined with external physiological data to be visualized inside a complete 3D representation of the training scenario. Chernova et al. [19] record human-human interaction collected in a virtual world and use this record to generate natural and robust human-robot interactive behavior in similar real-world environments. Bonebright et al. [20] present a very different usage of data captured in a VE. They offer a general methodological framework for evaluating the perceptual properties of auditory stimuli. The framework provides analysis techniques to measure the effective use of sound for a variety of applications including virtual reality. Börner and Lin [21] present work about the analysis and visualization of chat log data collected in 3-D virtual worlds. The log files contain chat utterances from different people that attended a demonstration of different learning environments. From the chat log files, the authors aim at answering questions such as: How many users participated in the discussion surrounding the demo as logged in the chat files? How much overlap exists among the log files? How much do users chat and who chatted the most? How many utterances are devoted to greeting, explanation, commands, questions, or other topics? How long is the average utterance length (number of words in an utterance) for different users? How often do users whisper? In later work [22], the authors present an analysis and visualization of user interaction data that reveals the spatial and temporal distribution of different user interactions in a 3D VE. We note from our examples that *what* data is collected is directly related to *why* it was collected and *how* it will be used.
- *What module collects the data.* The data may be collected automatically from the virtual system itself, or there could be another external module that exports/pulls the data from the running virtual environment. As an example of the former, Teixeira et al. [23] automatically collect, via the ErgoVR system, the following data: Time spent, Distance travelled and Behavioral compliance with exit signs. As an example of the use of external modules, Börner et al [22] introduce two tools which read VE data files and visualize it. The first tool, called *WorldMapper*, creates a 2D clickable map showing the layout of the world as well as interaction possibilities. The second tool visualizes user interaction data such as navigation.
- *Purpose of the VE used to collect data.* The VE could be the targeted VE or a demo VE used as a practice trial to collect data before the user can work with the targeted VE. As an example to testing VE, Griffiths et al. [24] present a tool called the

Nottingham Tool for Assessment for Interaction in Virtual Environments (NAIVE) for screening experimental participants experiencing difficulties. NAIVE comprises a set of VE tasks and related tests, with appropriate performance criteria levels, covering the main aspects of navigation (viewpoint) control and object manipulation and operation.

- *How the log files are analyzed.* The type of analysis performed on the log file/s collected from a VE can vary. Bruckman [25] differentiate between two types of log file data analysis: qualitative or quantitative. Usually, qualitative log file analysis needs a manual interpretation, while quantitative analysis could be performed totally automatically or with some manual translation to data meaning. The author provides two examples for both qualitative and quantitative log files data analysis.
- *Location where the log file is stored.* Log files used to monitor participants' behaviors while using the virtual system may be classified into two categories 1) log files stored on the user's computer, or 2) log files piped to an external database on a central server [26].
- *Time period and trigger for registering data.* Data may be continuously collected at regular time intervals (persistent registration) or data collection may be triggered to begin and end based on certain events (action-based registration)
- *Time when collected data is visualized.* Visualizations can be performed in real-time while acquiring the data during a session or from recorded data after the session has been completed. Either could be achieved remotely where the subject of the VE and human evaluators are separated by time and space [27].

3 The Virtual Environment to Be Evaluated

Before presenting the data collected and use of that data to draw meaningful conclusions about the usability of our VE, it is important to review the goals of our VE [28]. Our VE is an educational virtual world that consists of a simulated ecosystem for an imaginary island called Omosa in which school students can learn scientific knowledge and science inquiry skills. The Omosa Virtual World has been implemented using Unity3D. The learning goal is for students to use science inquiry skills to determine why the simulated animals, known as Yernt, are dying out. Our particular focus is on creating a world that encourages collaboration between the agents and the human and between the humans. Our future need to measure the extent and nature of collaboration has driven our interest in finding an automated method to measure and visualize the users' interactions.

The island of Omosa consists of four different locations the learner can visit. In each location there will be a virtual agent waiting for the visit of the group of companion learners. The learners can ask each agent a set of questions (between 7 and 9 questions). The group members will collaborate to explore the island and visit several different locations. Currently we have developed four locations: the village, the research lab, the hunting ground, and the weather station. In the village the student will meet both the *fire stick agent* and the *hunter agent*. In the research lab the students can meet the *ecologist agent*, and in the weather station the students can meet the

climatologist agent. Each agent has a list of questions that the user can ask about the agent and each agent will present an alternative view on why the Yernt are dying out.

In addition to interacting with various agents about the possible causes for the Yernt's increased death rates, the students collect information and data notes to compare the current and past states of Omosa and to generate hypotheses about the possible causes of the problem. There are four sets of notes the students can pick up. First, the *rainfall notes* are located in the weather station and contain information about temperature and rain level readings in different periods. Second, the *village field notes* are located in the village and contain information about the activities of the people in Omosa during an earlier period in time. Third, *tree ring notes* are located in the research lab and contain information about the internal structure of the stems of the trees on the island. Fourth, *ecologist notes* located in the research lab contain notes about the changes in the predator-prey ecosystem of Omosa Island. After exploring the virtual world and collecting notes, data, and other observational evidence from the simulated island, the group members will be asked to write a report that summarizes their conclusion about what is the cause of the changes in the ecosystem of Omosa and what is the reason the Yernt are dying out.

4 Data Acquisition and Visualization

A user model is introduced by Mikovec and Curin [29] that can represent user activities in the virtual environment. The proposed model has three levels of detail concerning activities: a) the motion of the user in the environment, b) the detailed behavior of the user in communicating with the other agents/users in the system and c) the users interactive activities with the program such as selecting from a menu or clicking on a button.

In Omosa Virtual World, we used the above three levels of user model as a guide to determine which data should be collected about the user activities. In order to visualize the users' activities and analyze their behavior, data acquisition occurs every one second, and the acquired data is stored in 3 different log files. The first log file allows us to simulate the movement of the user by storing the Cartesian coordinates of locomotion of the learner and contains the time of registration and the x-y-z coordinate. The second log file allows us to measure communication within the environment by storing which agent the learner meets, which question the learner asks and when the question is asked. The third log file allows us to measure the user's interaction with the program/environment by storing which item the learner collects and when it was collected. The amount of data collected during the user activity tracking is very large. For proper understanding of the meaning of these data, visualization methods must be used.

4.1 Cartesian Coordinates of Locomotion

Recording the trajectory of the learner while navigating within the system can be helpful in different ways. This piece of information could illustrate several important issues such as the length of time in seconds spent on a specific task and the distance

travelled with a certain locomotion. The following questions may be answered using the data in the Cartesian coordinate log file: which location of the virtual world does the user go to; which places does he see; which place/s the user does not visit and so does not get the information contained in it? Since the log file registers the time along with the Cartesian coordinate, we may recognize the speed of navigation and in which location the user stays idle for a long time. When the user stays idle for a long time in an important place which contains data to be collected that is reasonable. However, if the user stays idle in an unimportant place this might identify a difficulty in using the system and this is a sign to potentially revise the usability of the current virtual system. Also, the time spent in a specific task could be a measure of its difficulty.

4.2 Detailed Behavior of the User

Another level to help analyze and understand the behavior of the learner while using the VE is to determine which character (agent) the user meets and when and which object the user picked up and when. This information can illustrate the interaction the user has with the objects in the virtual world. The information could be used to show what agents and items the user is interested in while in the virtual system, and which items or agents the user does not pick up or see which may indicate problems in the human-computer interface design.

4.3 The Interactive Activities with the Program

The third level of user activity that is collected concerns the users' interactions with the virtual world. This can involve clicking on certain symbols or choosing a function key. These actions can give information about the usability of the system, for example, repeated exiting of the system (i.e., repeatedly taking the exit button and then going back in to the world) may be a sign that the user faces a problem about how to achieve something or is unsure what they should be doing.

4.4 User Identification and User Models

By collecting the above data we are able to capture a model of the user's activities. These different types of data are stored in multiple Log files. Each of the three levels of user-model presented above has its own file to store its unique data. For each unique user and for each session there are three types of files—inventory Items, positions, and questions—and each type of log file contains data of each element of the user model. Log files in the Omosa Virtual World have a unique structure. The name of each log file contains a unique ID which is a concatenation of the user ID plus the date when the virtual world was used to distinguish each separate user/session.

5 Results

We have conducted two classroom studies involving Omosa Virtual World involving Year 9 students in Australia at the end of 2011. The data that we include in this paper

is taken from the second study that involved 54 students in four online sessions over a period of two weeks. For technical and logistical reasons, we did not collect data from all students in every session. The focus of our data collection was on traditional methods such as video capture, pre and post knowledge tests, interviews and focus groups. The students used a printed “guidebook” that had learning activities for Omosa for each class period and written problems for the student to answer as assessments.

In terms of retracing and understanding what the students did in the world, we have focused on using the log files captured using functions in Unity3D. Using this approach, we have been able to gain an understanding of what the students did in Omosa and to be able to visualize their activities at the press of a button in a manner that is much faster than human coding of screen capture videos. Indeed, it is in response to the challenges of how to collect and analyze data in an efficient and meaningful way that we have developed the approach presented in this paper. Given that our interest is in the complex human behavior of collaboration, it was imperative that we developed techniques that would not require intensive qualitative analysis or the involvement of usability or collaboration/communication experts.

In order to evaluate the usability of the Omosa Virtual World and to create user-models of interaction we have written Boo scripts to automatically collect data into log files while students used the Unity3D game. Data analysis was done using both Microsoft Excel 2010 and Matlab 2012a. In accordance with the user-model presented above, three log files for each of the participants were collected, namely one log file for the user’s position, one log file for the virtual agents the user met, and another log file that captured which questions were asked and which notes were picked up. We next present the results for each of these levels.

5.1 Cartesian Coordinates of Locomotion

The user can click a button in Omosa Virtual World to see an aerial picture/map of the island. By clicking on any part on the map the user is taken directly to this location. The position log file is used to register the path the user takes while navigating in the VE. This log file is able to show how many times the user transferred from one place to another and whether there is a certain place the user repeatedly visits. This file can be used to identify if the participant has experienced any navigation obstacles and also helps to identify how many times the user quit exploring the VE and went back to the main menu of the application. We used both Microsoft Excel 2010 for visualizing the first level of user-model of testing usability due to its ease of use and flexibility in creating charts, we also used Matlab 2012a for analyzing the sudden changes in user location coordination which is a reference of quitting the VE and getting back to application menu. The chart/graph produced for a single user is shown in Fig 1. We can include multiple users so that we can compare different navigation patterns, which will be particularly useful when seeking to compare different cohorts or different experimental treatments.



Fig. 1. Visualizing the motion path of one of the learners (dark trajectory), and the transition between different locations (light trajectory)

5.2 Detailed Behavior of the User

In Omosa, the user can interact with the system by meeting various agents and asking them questions regarding the problems facing the imaginary island, namely why the Yernt are dying out. Fig. 2 depicts which agent the student has met, where they met them and what notes/evidence they collected in that location. If we add arrowheads to the edges, the diagram will show the order in which these activities were performed.

The detailed behavior log file provides a lot of data which can be used to evaluate the usability of the VE. Questions which were answered by the data in this log file include the percentage of encounters that have been made with each agent and the percentage of questions asked of each agent. The results for the percentage of questions asked to each agent by all the participants are shown in Fig. 3. It is clear that the Firestick and Hunter agents were asked 39.29% and 31.11%, respectively, of the questions the user should ask, the Climatologist agent is asked only 21.25%. This result reveals that participants did not ask the Climatologist agent as much as other agents. Table 1 provides the actual numbers of questions available and asked for each agent. The results suggest further investigation why the Climatologist was least popular. The results, in general, suggest possible modification of the guidebooks to ensure that students engage more with all agents in order to achieve the learning outcomes.

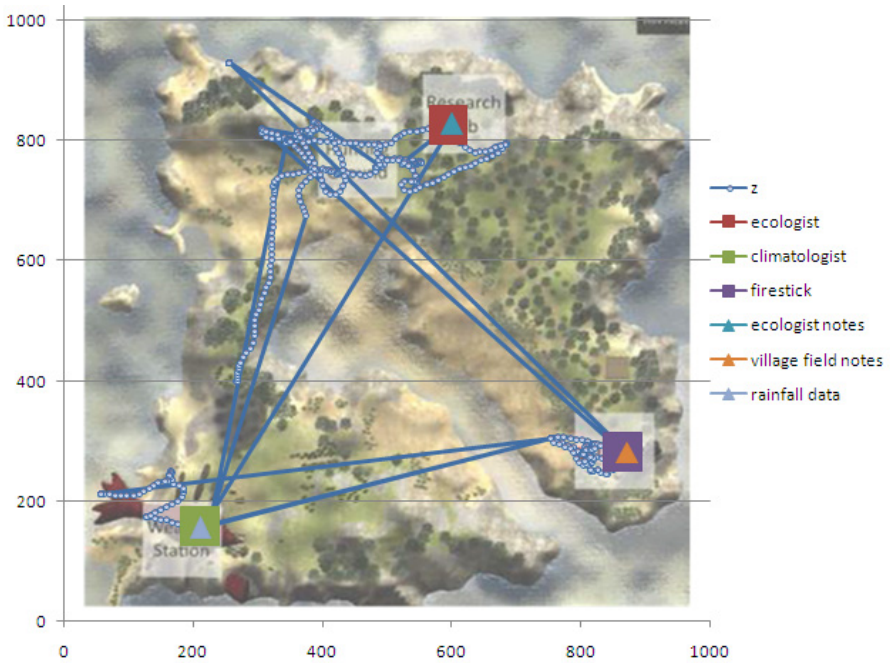


Fig. 2. Visualizing the agents the user meets and the items collected

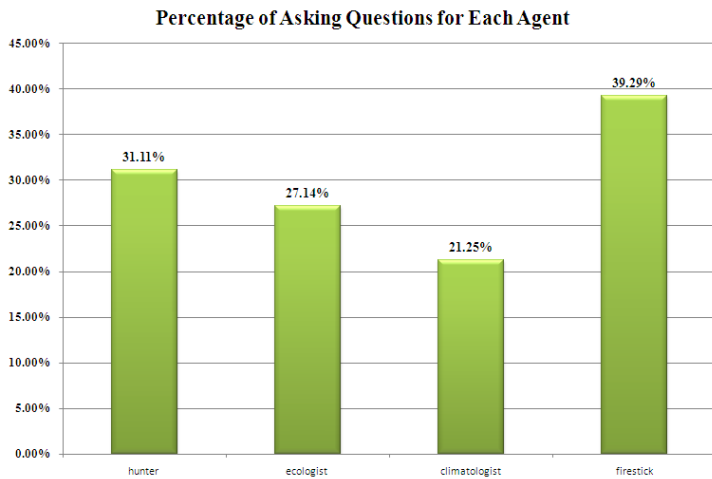


Fig. 3. Percentage of asking questions to each virtual agent

To drill down more deeply into the users’ interactions with each agent, the data in the log files were used to show the percentage that each individual question was asked in order to determine which question/s were asked less by the participants. This could be used to change Omosa or the guidebook. In Fig. 4, we see that the least asked

questions to the Hunter agent are questions 2 and 3 which inquiries about “How long have you been here?” and “Where am I?” The questions least asked to the Ecologist agent are 3, 1, 2 and 5 “Where am I?”, “How are you?”, “How long have you been here?” and “What instruments do you use to study Omosa?”. The least questions asked to climatologist agent are 3, 6, 4 and 5 which inquiries about “Where am I? “, “Where should I go next? “, “What do you eat?” and “What do you hunt?” The Firestick agent is the agent with the most asked questions, the question number 2 and 4 are the least asked questions that are both asked only 30% of the time.

Table 1. The average of asking questions to each agent

	Hunter	Ecologist	Climatologist	Firestick
No. of Questions	9	7	8	7
Average of asking questions	2.8	1.9	1.79	2.75

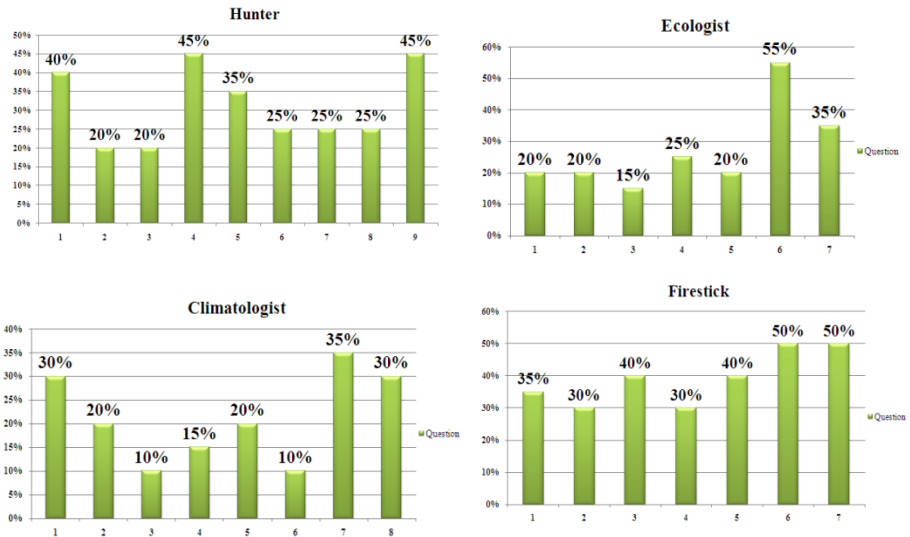


Fig. 4. Percentage of asking each question to each virtual agent

Another aspect of the interaction within the Omosa VE is the ability to collect items and notes. The log files have been used to determine which note is collected and when. Fig. 5. shows the percentage of participants who collected each item. According to the result of analysis, the notes about rainfall data which exist near the Climatologist agent is the least collected item, picked up by only 40 % of the users.

5.3 The Interactive Activities with the Program

The third level in the user-model of interaction is the activities/actions the user performs with the application; these activities could be clicking on a button or returning

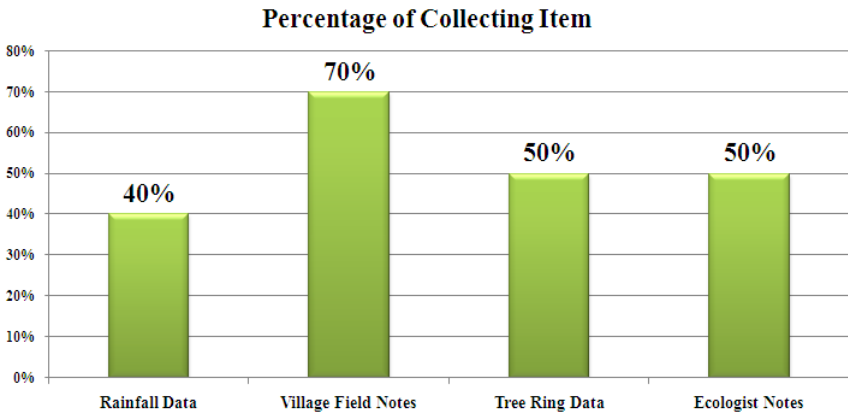


Fig. 5. Percentage of collecting each item

to the main menu. An option in Omosa VE is to show a map for the whole island and the user can be located directly in and place he selects. The user usually clicks to show the map for one of three reasons: first, when s/he finishes exploring certain location and wants to move to another location, second, when s/he gets lost or has a problem in a specific area and wants to move to another location, third, when s/he gets lost or has a problem in a specific area and wants to move to get back to the same location again. In Fig. 6 the frequency of using the program map for each participant is shown along with the average of using the system application. It is clear that 75% of the users (15 of 20 users) are around the average of using the application map while 25% (5 of 20 users) are over the average of using the program. Overuse could indicate confusion over what they should be doing in each place. Underuse could indicate lack of coverage of the four locations. Some further discussion appears in the conclusion.

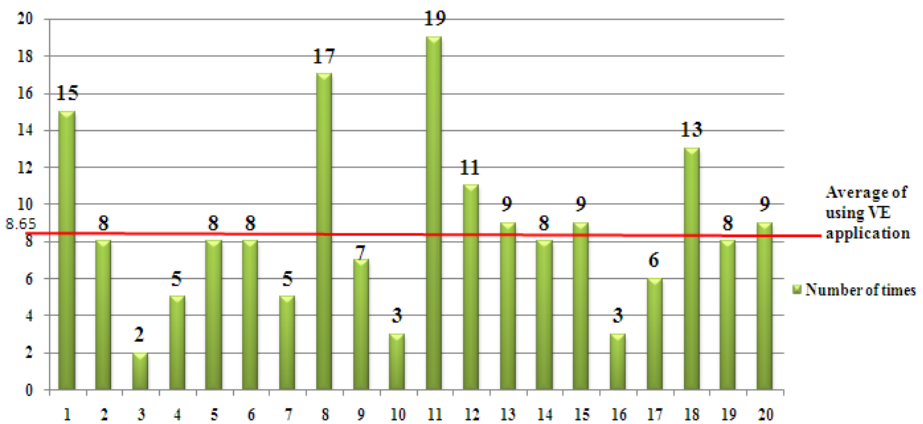


Fig. 6. The frequency of using Omosa VE application Map

6 Conclusion and Future Work

The aim of this paper is to present automated methods that can be used to shed light on the users' interactions in a VE to better understand how the world was utilized and its usability. Concerning the first level of the user model, the results show smooth exploration of Omosa, the only exception is when the user keeps navigating into the water around the island. In this situation the user will have to use the island map to get back to land.

Concerning the second level of the user-model, the result of analyzing log file data shows that all four agents in Omosa are interrogated by some participants but that interaction is as low as 21.25% in the case of the Climatologist agent. Also we found that while all four notes were collected by one or more people, the percentage varies and the least collected evidence is the rainfall notes (with 40% of participant users). This apparent underutilization of the information available in Omosa has prompted a review of our design of Omosa and the associated learning activities and guidebooks.

Concerning the third level of the user-model, the results show that the majority of the participants (75%) use the application map around 8-9 times to go from one location to another to achieve the task they have to complete, while around 25% of the users used the application map more than this average. By reviewing the cases that registered the highest usage for the map, namely user numbers 1, 8, and 11 we find out that they are curious to explore the remote or isolated locations of Omosa Virtual World such as surfing on the surface of water or going deeper into the wilderness, and as a result they have to click on the application map to get back on land again. Contrary to our initial assumptions, the small percentage of over usage of the application map did not mean a problem in the flow of navigating Omosa. Rather it represented the curiosity of some students to explore different and new virtual places.

In future work, we intend to make Omosa a more collaborative VE through adding the ability of online communication between the groups and introducing authentic collaborative tasks that require the human and agent to plan and work together to perform a task in each of the areas within Omosa. Automatic data collection will be upgraded to include collaboration awareness and collaborative interaction [30] to log if the individuals have faced difficulties in finding other groups online and clarifying other participants' thoughts. Information in the upgraded log files about social interaction will shed light on how teams collaborate, or do not collaborate, and that will help in improving the collaborative ability of Omosa. For these purposes, the log files may include text-chat logging, audio, and video logging. Overall, we hope that this research will contribute to the field by demonstrating the viability of using automated techniques for log file analysis in order to understand the dynamics of students learning trajectories in an educational virtual environment.

References

1. Hix, D., Hartson, H.R.: *Developing User Interfaces: Ensuring Usability through Product & Process*. John Wiley and Sons, New York (1993)
2. Johnson, C.R.: *Evaluating the Contribution of DesktopVR for Safety-Critical Applications*. In: Felici, M., Kanoun, K., Pasquini, A. (eds.) *SAFECOMP 1999*. LNCS, vol. 1698, pp. 67–78. Springer, Heidelberg (1999)

3. Jacko, J.A., Sears, A.: *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. Psychology Press (2003)
4. Teixeira, L., Rebelo, F., Filgueiras, E.: *Human Interaction Data Acquisition Software for Virtual Reality: A User-Centered Design Approach*, pp. 793–801. CRC Press/Taylor & Francis, Ltd. (2010)
5. Bowman, D.A., Gabbard, J.L., Hix, D.: *A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods*. *PRESENCE: Teleoperators and Virtual Environments* 11, 404–424 (2002)
6. Slater, M.: *How Colorful was Your Day? Why Questionnaires Cannot Assess Presence in Virtual Environments*. *Presence: Teleoperators and Virtual Environments* 13, 484–493 (2004)
7. Zhang, Y., Yu, X., Dang, Y., Chen, H.: *An Integrated Framework for Avatar Data Collection from the Virtual World*. *IEEE Intelligent Systems* 25, 17–23 (2010)
8. de Leeuw, E., Nicholls II, W.: *Technological Innovations in Data Collection: Acceptance, Data Quality and Costs*. *Sociological Research Online* 1 (1996)
9. Lampton, D., Bliss, J., Morris, C.: *Human Performance Measurement in Virtual Environments*. In: Stanney, K.M. (ed.) *Handbook of Virtual Environments: Design, Implementation, and Applications*, pp. 701–720. Lawrence Erlbaum Associates, Mahwah (2002)
10. Siochi, A.C., Hix, D.: *A study of computer-supported user interface evaluation using maximal repeating pattern analysis*. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching Through Technology*, pp. 301–305. ACM, 108926 (1991)
11. Andreas, K., Tsiatsos, T., Terzidou, T., Pomportsis, A.: *Fostering Collaborative Learning in Second Life: Metaphors and Affordances*. *Computers & Education* 55, 603–615 (2010)
12. Yee, N., Bailenson, J.: *A Method for Longitudinal Behavioral Data Collection in Second Life*. *PRESENCE: Teleoperators and Virtual Environments* 17, 594–596 (2008)
13. Duarte, E., Vilar, E., Rebelo, F., Teles, J., Almeida, A.: *Some Evidences of the Impact of Environment's Design Features in Routes Selection in Virtual Environments*. In: Shumaker, R. (ed.) *Virtual and Mixed Reality 2011, Part I*. LNCS, vol. 6773, pp. 154–163. Springer, Heidelberg (2011)
14. Laaki, H., Kaurila, K., Ots, K., Nuckchady, V., Belimpasakis, P.: *Augmenting virtual worlds with real-life data from mobile devices*. In: *2010 IEEE Virtual Reality Conference (VR)*, pp. 281–282 (2010)
15. Grillon, H., Thalmann, D.: *Eye contact as trigger for modification of virtual character behavior*. *Virtual Rehabilitation*, 205–211 (2008)
16. White, C.: *Classifying Human Motion in Virtual Environments (Unpublished paper describing the system)*, pp. 1–6 (2006)
17. Sebok, A., Nystad, E.: *Design and Evaluation of Virtual Reality Systems: A Process to Ensure Usability*. In: *Proceedings of the Virtual Reality Design and Evaluation Workshop* (2004)
18. Holm, R., Priglinger, M., Stauder, E., Volkert, J., Wagner, R.: *Automatic Data Acquisition and Visualization for Usability Evaluation of Virtual Reality Systems*. In: *Eurographics-Short Presentations* (2002)
19. Chernova, S., De Palma, N., Morant, E., Breazeal, C.: *Crowdsourcing human-robot interaction: Application from virtual to physical worlds*. In: *2011 IEEE RO-MAN*, pp. 21–26. IEEE (2011)
20. Bonebright, T.L., Miner, N.E., Goldsmith, T.E., Caudell, T.P.: *Data Collection and Analysis Techniques for Evaluating the Perceptual Qualities of Auditory Stimuli*. *ACM Transactions on Applied Perception (TAP)* 2, 505–516 (2005)

21. Borner, K., Lin, Y.C.: Visualizing Chat Log Data Collected In 3-D Virtual Worlds. In: Proceedings. Fifth International Conference on Information Visualisation, pp. 141–146 (2001)
22. Borner, K., Hazlewood, W.R., Sy-Miaw, L.: Visualizing the spatial and temporal distribution of user interaction data collected in three-dimensional virtual worlds. In: 2002 Proceedings of the Sixth International Conference on Information Visualisation, pp. 25–31 (2002)
23. Teixeira, L., Vilar, E., Duarte, E., Rebelo, F.: Comparing Two Types of Navigational Interfaces for Virtual Reality. *A Journal of Prevention, Assessment and Rehabilitation*, 2195–2200 (2012)
24. Griffiths, G., Nichols, S.S.n., Wilson, J.R.: Performance of New Participants in Virtual Environments: The Nottingham Tool for Assessment of Interaction in Virtual Environments (NAIVE). *International Journal of Human-Computer Studies* 64, 240–250 (2006)
25. Bruckman, A.: Chapter 58: Analysis of Log File Data to Understand Behavior and Learning in an Online Community. In: Weiss, J., Nolan, J., Hunsinger, J., Trifonas, P. (eds.) *The International Handbook of Virtual Learning Environments*, pp. 1449–1465. Springer, Dordrecht (2006)
26. Orkin, J., Roy, D.: The Restaurant Game: Learning Social Behavior and Language from Thousands of Players Online. *Journal of Game Development* 3, 39–60 (2007)
27. Hartson, H.R., Castillo, J.C., Kelso, J., Neale, W.C.: Remote Evaluation: The Network as an Extension of the Usability Laboratory. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground (CHI 1996), pp. 228–235. ACM (1996), 238511
28. Preece, J.: *Online Communities: Designing Usability and Supporting Sociability*. John Wiley & Sons, New York (2000)
29. Mikovec, Z., Maly, I., Slavik, P., Curin, J.: Visualization of Users' Activities in a Specific Environment. In: *The 2007 Winter Simulation Conference*, pp. 738–746 (2007)
30. Schroeder, R., Heldal, I., Tromp, J.: The Usability of Collaborative Virtual Environments and Methods for the Analysis of Interaction. *Journal of Presence* 15, 655–667 (2006)

User-Centric Recommendation-Based Approximate Information Retrieval from Marine Sensor Data

Zhao Chen¹, Md. Sumon Shahriar², and Byeong Ho Kang¹

¹ School of Computing and Information Systems
University of Tasmania, Sandy Bay, Hobart, Tasmania, Australia
zchen4@postoffice.utas.edu.au, Byeong.Kang@utas.edu.au

² Intelligent Sensing and Systems Laboratory (ISSL),
CSIRO ICT Centre, Hobart, Australia
mdsumon.shahriar@csiro.au

Abstract. Due to the wide use of low-cost sensors in environmental monitoring, there is an increasing concern on the stability of marine sensor network (MSN) and reliability of data collected. With the dramatic growth of data collected with high sampling frequency from MSN, the query answering for environment phenomenon at a specific time is inevitably compromised. This study proposes a simple approximate query answering system to improve query answering service, which is motivated by sea water temperature data collected in Tasmania Marine Analysis and Network (TasMAN). The paper first analyses the problems of special interest in missing readings in time series of sea water temperature. Some current practices on approximate query answering and forecasting are reviewed, and after that some methods of gap filling and forecasting (e.g. Linear Regression (LR), Quadratic Polynomial Regression (QPR), Moving Average (MA) and Autoregressive Integrated Moving Average (ARIMA)) are introduced in designing the simple approximate query answering system. It is followed by experiments on gap filling of time series with artificial noise made in the original time series. Finally, the comparison of different algorithms in terms of accuracy, computation time, extensibility (i.e. scalability) is presented with recommendations. The significance of this research lies in the evaluation of different simple methods in forecasting and gap filling in real time series, which may contribute to studies in time series analysis and knowledge discovery, especially in marine science domain.

Keywords: Sensor data, data and knowledge acquisition, approximate information retrieval, statistical and machine learning.

1 Introduction

With the emergence of low-cost and smart sensors, sensor networks offer the potential to facilitate scientific research such as environmental monitoring in periodically collecting sensor readings from remote locations. Furthermore, it is

predicted that there will be a growth in deployment of sensor networks because of gradually decreasing cost of sensors [1]. TasMAN [2] is a typical application of sensor network in marine research.

However, there are several issues about time series data gathered from low-cost sensor networks. The first one is the concern on the stability of data collection phase which is compromised by sensor node failures, transmission errors. In another word, there may be many missing readings which can appear in gaps or single breakpoints in time series. Another challenge is the gradual accumulation of data readings which result in expansion of database storage and furthermore slows down traditional query answering. The deployment of inexpensive sensors also raise the question of irregularity in observation interval because of the aging of device. Therefore, the study on how to technically deal with those problems in time series is of critical importance. The aim of this study is to design a simple approximate query answering system which can answer queries with approximate results. First, the gaps in time series are filled with the use of time series forecasting methods. Moreover, approximate results are stored in database which can answer the same queries for different users without duplicate computation. Currently, there are few applications of approximate systems in marine sensor networks.

The paper is organised as follows. It starts with briefly stating related research in Section 2. Then we present the description of data pre-processing in Section 3. Section 4 discusses various issues related to gap filling and approximation. Section 5 presents different methods used in designing approximate query answering system. Section 6 evaluates different methods in practice. Finally, in Section 7, it shows the implementation and user interface design in approximate query answering system.

2 Related Research

Acharya et al. [3] offered a detailed description of the first approximate query answering system (AQUA) which was designed to provide fast, highly accurate approximate answers. The mechanism is to employ an approximate query engine along with data warehousing. The approximate query engine has various synopses and those synopses are used to approximately answer queries without invocations to main data warehouses. Those synopses were updated periodically and the accuracy of approximate query answers is ensured by the use of join sampling and biased sampling. The deficiency of this system is the limitation of queries with select, aggregate, group by and/or joins. This research enlightened us in the design of approximate query answering system with pre-computed synopses to lessen response time and periodical maintenance to deal with data streaming issues.

Mestekemper et al. [4] evaluated some frequently used forecasting methods and proposed a suitable model to forecast water temperature on the basis of hourly data in their project. Based on the evaluation among some models usually applied in econometrics (i.e. Least squares estimation (LS), Maximum likelihood

estimation (ML) and Full maximum likelihood (FullML)), LS was noted as the best model to fit and forecast water temperature using water and air temperature observations in previous days. Moreover, the LS model is further compared with multiple regression analysis, second-order Markov process and Box-Jenkins model. It is indicated that LS model outperformed those three models in hydrologic field. The proposed model can forecast water temperature for 3 days ahead with reasonable accuracy. The significance of their research lay in the similar research domain as in our project. They also found that the water temperature for one hour is mainly related to the previous two to five earlier observations.

In general, there are insufficient researches in univariate time series gap-filling especially in marine science domain. Most researches address gap-filling for one time series in correlation to other related time series which may have some readings at different locations, depths or for different phenomena. However, there are tremendous studies in time series forecasting area which inspired us to apply forecasting techniques in the gap filling practice.

3 Data Preprocessing

In this study, observation on sea water temperature is collected from TasMAN marine sensor network.

3.1 Data Access

The original data in TasMAN is an open data source stored in XML format as web documents. It can be achieved by a web link as follows:

```
http : //www.csiro.au/tasman/WDS/wds?start = 20110801000000&end =
20110807000000&request = GetObservation&format = xml&sensors =
1.100.1
```

In the hyperlink, sensors=1.100.1 is in the format of $\langle network_id \rangle . \langle feature_id \rangle . \langle sensor_id \rangle$. There are several sensor node clusters in a sensor network and similarly each cluster contains multiple sensors for different phenomenon (i.e. water temperature, salinity, pressure, conductivity and etc.) at different depths. The duration of observation is specified with a start time and an end time (start=2011080100000 and end=20110807000000 stand for the start time and end time respectively in the format of YYYYMMddHHmmss). This research focus on a specific sensor 1.100.1(1.100.1 = $\langle TasMAN \rangle . \langle CMARwharfnode \rangle . \langle Temperature[EC250] \rangle$) which collects sea water temperature observations at the depth of 1 metre underwater from the year 2008. It is difficult to directly view observations in web browser because the large volume of data takes long time to be loaded.

3.2 Data Extraction

Data extraction is facilitated with a program due to large data volume and multilevel XML structure. Although the data source is theoretically approachable as

hyperlink by web browser, data loading is fairly slow and sometimes fails when a long duration is chosen which consequently results in large volume of XML document. Moreover, there are four levels of elements from networks, features, sensors and finally observations in the XML structure. There are six attributes (*id*, *time*, *value*, *qc_flag*, *qa_uncertainty* and *qa_algorithm_version*) in each observation element. The detailed description of the attributes are given in the Fig.1

Attribute	Description
id	Observation ID
time	Observation time (yyyyMMddTHHmss)
value	Water temperature value (degree C)
qc_flag	QC flag (0 to 4)
qa_uncertainty	QC/QA uncertainty using automated assessment algorithms
qa_algorithm_version	Automated assessment of data quality from CSIRO

Fig. 1. Attributes of the sensor data

The original format of time in XML file is not a standard time format for data storage. Therefore, it is converted in the program to delete letter 'T' in time string for every observation. In addition, a daily aggregation of maximum, minimum and average value has been made according to the original data when there are enough observations in that day.

3.3 Data Archive

Microsoft Visual Studio and Visual C# are the platform and programming language used in this research because their excellent performance in data presentation and processing. The original data and daily aggregation are stored in different tables in a MySQL database. With the original and pre-processed data archived in the local database, the water temperature time series is independent from the unstable web connection and can be manipulated locally without corrupting original XML file.

4 Data Analysis

The time series used in this research covers water temperature observation from February 19th, 2008 to March 20th, 2012. The standard sampling rate is every five minutes. Theoretically, there are exactly 1492 days with daily expected 288 observations which counts to 429696 records in total. However, the actual number of observations collected by the sensor is 387311, which means there are substantial missing values (42385) in the time series. Holistically, there are three apparent issues in terms of gaps, sampling numbers and sampling rate in the original time series after data analysis.

4.1 Gaps

It is found that missing observations mostly spread out as gaps which contain a large number of missing values instead of individual breakpoints evenly-distributed in the time series. When analysing the time series daily, there are 140 days with no records and many gaps persist for more than one week. Here is a list of gaps in the time series as shown in the Fig 2.

Gap start date	Gap end date	Duration (days)
2008-05-01	2008-05-06	6
2008-07-18	2008-08-04	16
2008-11-11	2008-11-13	3
2009-05-21	2009-05-24	4
2009-12-15	2010-01-05	22
2010-03-06	2010-03-21	22
2010-11-06	2010-11-06	1
2011-01-28	2011-02-06	10
2011-04-05	2011-04-06	2
2011-04-19	2011-05-12	23
2011-12-20	2012-01-07	19
2012-01-18	2012-01-31	14

Fig. 2. Statistics of gaps in the sensor data

Moreover, there are considerable amount of missing values unevenly-distributed in those days with observations.

4.2 Sampling

It is assumed that the sampling rate for observation to be five minutes, there should be 288 readings collected from the sensor every day. However, there are a considerable number of days which are under-sampled or over-sampled. Here is a pie chart to show the percentage of days which are under-sampled, well-sampled and over-sampled as shown in Fig 3.

It depicts that theoretically the number of under-sampled days (627) counts to approximately 46 per cent in the whole time series. More than half number of dates in the time series shows a good performance of the sensor in collecting sea water temperature observations. A few days (32) were over-sampled observations, which implies the sampling rate was changed during that time.

Under-sampling. In under-sampled dates, the number of samples also varies. The following pie chart illustrates that more than 75 per cent of under-sampled dates has a good coverage of samples (the sample number is greater than 250 per day) in spite of naming under-sampling. In contrast, there are 54 days (8.61%) in the chart that match the term of under-sampling in Fig 4.

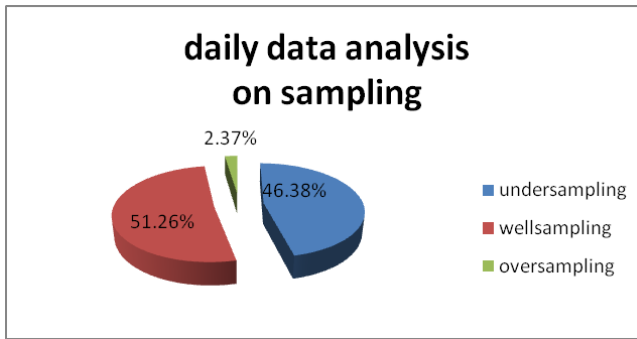


Fig. 3. Preliminary data analysis

Despite of slight variation in sampling frequency (seldom sampling interval is more than five minutes but less than six minutes.), a considerable number of slightly under-sampled days could be accepted when the data is used for daily-based approximation.

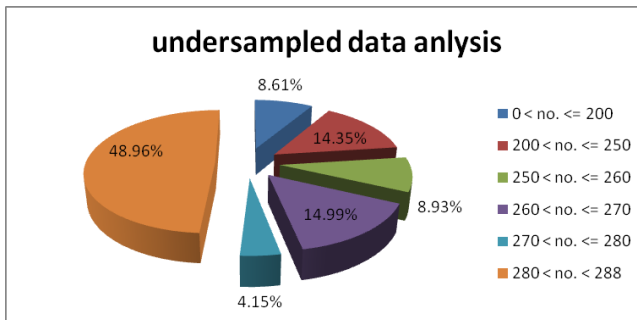


Fig. 4. Statistics of under-sampled data

Over-Sampling. Generally, the frequency of over-sampling in the water temperature time series is fairly low. There are 32 days in total (which is 2.37 per cent of dates in time series) when more samples are collected probably due to the change in sampling rate. The over-sampled days are distributed mainly in June and August in 2011 when the sampling rate was changed from five minutes to one minute. Furthermore, over-sampling would not make impact on daily aggregations of maximum, minimum and average values of sea water temperature.

4.3 Sampling Rate

Theoretically, an ideal sampling rate should retain the same in the time series, which is advantageous for data analyse in focusing on the variations of a

univariate time series. However, the sampling rate is changeable in water temperature time series. During the period when the sensor was deployed to August 2010, the frequency was mostly maintained well with the time interval shorter than five minutes and ten seconds. After that time, the sampling rate was relaxed to be irregular (sometimes less than one minute and sometimes greater than seven minutes). Another notable change in sampling rate was in June and August 2011 when it was reduced to around one minute.

4.4 Data Analysis by Quality Flags

Another noticeable figure is the quality control flag in the original observations. The quality flag represented automated quality assessment done with the use of the Fuzzy Set Theory in accordance to IODE flag standard used by both the Argo floats and the Australian Integrated Marine Observing System [5]. Here is a pie chart of data analysis according to quality flags as shown in Fig 5.

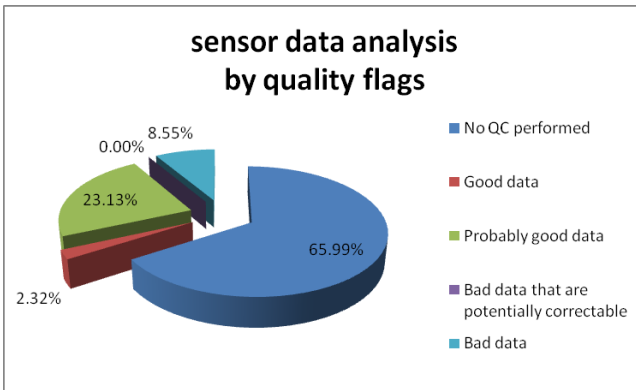


Fig. 5. Data analysis based on quality assurance and control

It shows that more than a quarter ($2.32\% + 23.13\% = 25.45\%$) of data readings are considered as good or probably good data by some QA/QC algorithms. In contrast, the already identified bad data is less than ten per cent (8.55%). In addition, about two thirds (65.99%) of original readings in time series have not been assessed by any kinds of QA/QC algorithms.

5 Methodology

This section reviews four basic forecasting methods used in the design of the simple approximate query answering system: Linear Regression (LR), Quadratic Polynomial Regression (QPR), Moving Average (MA) and Autoregressive Integrated Moving Average (ARIMA).

5.1 Linear Regression (LR)

Linear regression is the first type of regression analysis. It is a linear predictive model to identify the relationship between a scalar dependent variable y and one explanatory variable denoted X . In a simple linear regression model, there are two constant coefficients: the slope (a) and the intercept (b). After setting up the model, if there is an input value for X which is not accompanied by y , the model can predict the value of y with computation of the inputting X value, slope and intercept ($y = aX+b$).

5.2 Quadratic Polynomial Regression (QPR)

Quadratic polynomial regression is another form of linear regression. In QPR, the relationship between the scalar independent variable X and the dependent variable y is interpreted as a second order polynomial. With infinite independent variables, the second order polynomial represents a parabola. In QPR model, there are three constant parameters: parameter for square of X (a), parameter for X (b) and c term. After developing the model, if there is a input value X , the model can predict the value of y with computation of the X value and terms of a , b , c ($y = aX^2 + bX + c$).

5.3 Moving Average (MA)

Moving average is used to estimate the average value in the time series especially in removing the effect of random fluctuations. The calculation in moving average is straightforward and involves finding the average value for the n most recent time period and using that average as the forecast for the next time period. If the value of n is a constant, moving average is a recursive computation for the future time periods.

5.4 Autoregressive Integrated Moving Average (ARIMA)

In time series analysis, the Autoregressive Integrated Moving Average is a generalisation of an Autoregressive Moving Average (ARMA) model with an extra integrated differencing coefficient which is targeted to remove non-stationarity. In ARIMA(p,d,q), p stands the order of autoregressive part, d refers to integrated part and q represents moving average parts.

6 Implementation and Evaluation

In the design of simple approximate query answering system, several assumptions are made in terms of user, prediction pattern, sampling rate, number of data in approximation. First, the system caters for two categories of users: general users who are interested in the general sea water temperature information at a specific date (daily based) as well as specialists who are interested in water temperature at a time specified to hour and minute. For general users, the

daily water temperature is shown with maximum, minimum and average values. Therefore the original observation is transferred to maximum, minimum and average value pairs every day, which means if the number of observations is less than a fixed number (150) in a day, that day should be treated as no readings. (The percentage of days with readings less than 150 is less than 5% in the whole time series.) Second, the system applies a backward prediction pattern for gap filling. In other words, if there is no reading at user specified time, the program would compute the closest reading backward instead of finding the closest readings after that time.

For sake of simplicity, the sampling rate is assumed to be fixed, which means the system neglects the variance in seconds of sampling rate. (Sampling rate of 5 minutes 30 seconds and sampling rate of 5 minutes 1 second are treated as similar.) Furthermore, the number of readings required for approximate computation is assumed to be proportional to the interval of gap. In the system testing phase, some artificial gaps were randomly made in the time series. Therefore, different methods refilled the artificial gaps. The performance of different forecasting methods was evaluated in terms of accuracy, time consumed and scalability. The accuracy was determined by the difference of approximate results comparing with the backup of original data. Here are two tables in evaluations for daily-based approximation and sampling-rate-based approximation respectively as shown in Fig 6 and Fig 7. The standard of accuracy is explained as follows:

good: 70% of tests has variance between raw data and approximate result of 5%

reasonable: 70% of tests has variance between raw data and approximate results of 10%

poor: less than 70% of tests has variance between raw data and approximate results of 10 %

The standard of computation time is explained as follows:

short: less than 1500 milliseconds

medium: around 2000 milliseconds

long: greater than 3000 milliseconds

For general user in daily-based approximation, the performance of gap filling methods was gradually degraded with the expansion of gap dates. Every method was qualified in filling short gaps with fairly accurate results in a short time. For medium gaps (≤ 3), LR, QPR and ARIMA were competent in approximation with reasonable accuracy and short computation time. When the gap increased to more than 5 days, ARIMA was the only method that can provide reasonable accurate results in an acceptable time interval. Therefore, ARIMA was most scalable in daily-based gap filling in comparison to other methods. Another noticeable finding was that the performance of each method would be under-expected if the gap was longer than 7 days.

Gap date	Algorithms	Accuracy	Computation time	Scalability
1	Linear regression	good	short	low
	Quadratic polynomial regression	good	short	medium
	Moving average	reasonable	short	medium
	ARIMA	good	short	high
<=3	Linear regression	reasonable	short	
	Quadratic polynomial regression	reasonable	short	
	Moving average	poor	short	
	ARIMA	reasonable	short	
<=5	Linear regression	poor	medium	
	Quadratic polynomial regression	reasonable	medium	
	Moving average	poor	medium	
	ARIMA	reasonable	medium	
<=7	Linear regression	poor	medium	
	Quadratic polynomial regression	poor	medium	
	Moving average	poor	long	
	ARIMA	reasonable	medium	
>7	All	poor	long	

Fig. 6. Daily-based approximation

Gaps by Sampling rate	Algorithms	Accuracy	Computation time	Scalability
1	Linear regression	good	short	low
	Quadratic polynomial regression	good	short	medium
	Moving average	good	short	medium
	ARIMA	good	short	high
<=3	Linear regression	good	short	
	Quadratic polynomial regression	good	short	
	Moving average	good	medium	
	ARIMA	good	short	
<=5	Linear regression	reasonable	medium	
	Quadratic polynomial regression	reasonable	medium	
	Moving average	good	medium	
	ARIMA	good	short	
<=7	Linear regression	poor	long	
	Quadratic polynomial regression	reasonable	long	
	Moving average	poor	long	
	ARIMA	poor	medium	
>7	All	poor	long	

Fig. 7. Sampling-rate-based approximation

Similar to daily-based approximation, the performance of gap filling methods in sampling-rate based approximation was impaired with the increase of gap time interval. All methods can quickly fill short gaps with accurate results. LR, QPR and MA were further competent in recovering for medium gaps (≤ 3). When the gap exceeded 5 times of sampling rate, ARIMA and QPR can stably provide approximate results for gap filling. In scalability aspect, ARIMA once again outperformed other methods in scalability. When the gap was longer than 7 times of sampling rate, each method had a relatively poor accuracy and it took longer time for approximation.

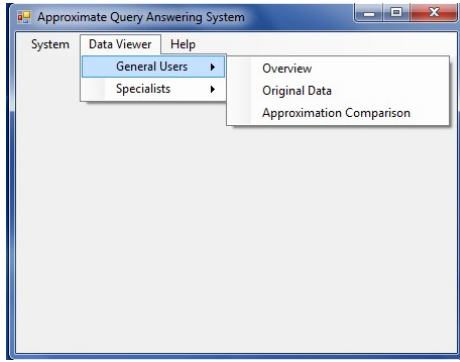


Fig. 8. Main Graphical User Interface

Because the daily data has been pre-processed to three figures (maximum, minimum and average), the fluctuation in maximum and minimum values was not well smoothed. Consequently, the water temperature values (especially maximum and minimum) fluctuated heavier than in sampling rate based approximation. In addition, the time series was locally linear in a short interval. Therefore, the performance of each method (especially MA) in daily-based approximation was not as good as it was in sample-rate-based approximation.

A noticeable feature of the simple approximate query answering system is that the gap filling process stores the approximate values and therefore if users' queries go to the same gap twice, the approximate results would not be double computed. Therefore, the approximate query answering system can reuse approximate results in the future, which is also recorded in other researches such as in [6,3,17].

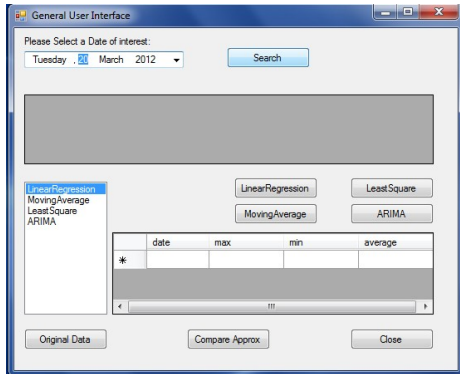


Fig. 9. Overview for general user

7 Demonstration of the System

In this section, the simple approximate query answering system is demonstrated with screen shots. As shown in Fig. 8, in the main user interface, users can select from general users who are interested in the general sea water temperature information at a specific date and specialists who are interested in water temperature at a specific time (specified to hour and minute). There are three interfaces they can enter: Overview (where approximation can be done), Original Data (raw data with no approximation in gaps) and Approximation Comparison (approximate result at a chosen time if available).

In the general user interface shown in Fig. 9, users can select a date from calendar picker. If the date is selected and the search button is clicked, the water temperature of that day would be shown in first viewer if available. Otherwise, users can choose a method for approximation at different button. After that, users can view the approximate results by selecting a method in the list box at left. Moreover, users can view original data at that day by clicking original data button left underneath and view approximate result comparison by clicking compare approx button.

The difference of specialist interface shown in Fig. 10 and general user interface is that specialists can specify their time of interest to minutes. When the original data at a date or time is not found in the original time series, there is a message box to remind users shown in Fig. 11.

A screen shot for original data as a Zed graph is shown in Fig. 12. The X axis is in the time format and Y axis is the water temperature in °C. The blue bubble in the line chart is the observations collected the sensor. The comparison of approximate results for different approximation methods is demonstrated in Fig. 13.

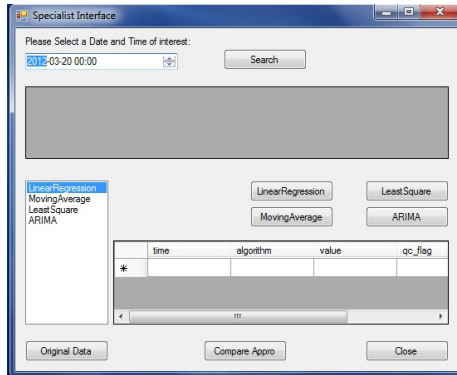


Fig. 10. Overview for specialist

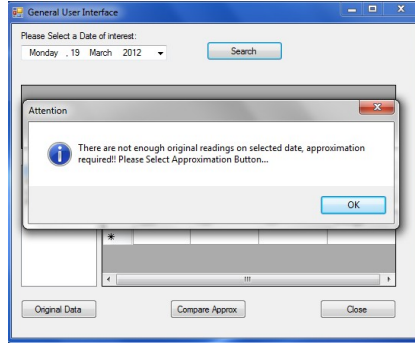


Fig. 11. Message box to remind users

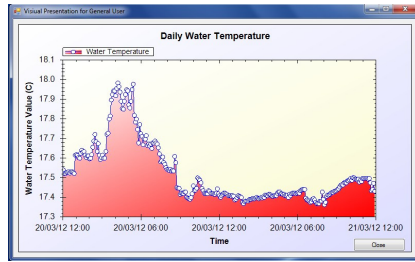


Fig. 12. Time series visualization

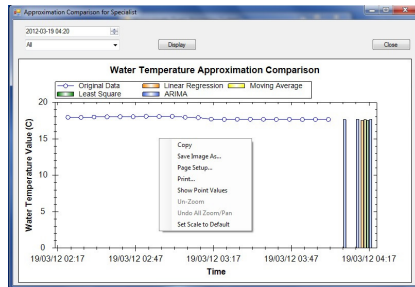


Fig. 13. Water temperature approximation comparison for specialist user

In the data, area users have options to save image and zoom in the image in a selected area if they would like to view the comparison more clearly. The enlarged image for demonstration is shown in Fig. 14.

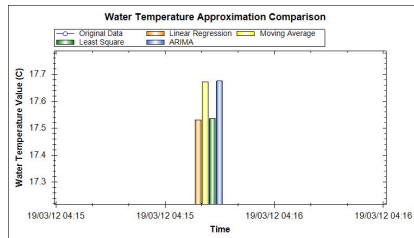


Fig. 14. Water temperature approximation comparison for specialist user(enlarged version)

8 Conclusion

In this paper, a simple approximate query answering system is proposed in answering queries with approximate results when queries go to gaps where observations are not found in the time series of sea water temperature. This paper first introduces data pre-processing which includes data accessing, extracting and archiving. Next, data analysis is made that found three main issues in terms of gaps, sampling and sampling frequency. The paper further reviews some general prediction methods which can also be applied in gap filling for missing values. It is followed by implementation of those methods in the simple approximate query answering system and experiments on performance of each method towards artificial gaps in terms of accuracy, computation time and scalability. Furthermore, the performance of the simple approximate query answering system is demonstrated in different scenarios. In summary, each approximation technique has its advantages and limitations in gap filling and therefore the approach to implement a hybrid model which combines different approximation methods is of constructive significance in gap filling for time series in marine science domain.

Acknowledgement. The Intelligent Sensing and Systems Laboratory at CSIRO ICT Centre is jointly funded by the Australian Government through the Intelligent Island Program and CSIRO. The Intelligent Island Program is administered by the Tasmanian Department of Economic Development, Tourism and the Arts. This research was conducted as part of the CSIRO Wealth from Oceans National Research Flagship.

References

1. Shahriar, M.S., de Souza Jr., P.A., Timms, G.P.: Smart query answering for marine sensor data. *Sensors* 11(3), 2885–2897 (2011)
2. Timms, G.P., McCulloch, J.W., McCarthy, P., Howell, B., de Souza Jr., P.A., Dunbabin, M.D., Hartmann, K.: The Tasmanian marine analysis network (TasMAN). In: *IEEE Oceans*, pp. 1–6 (2009)
3. Acharya, S., Gibbons, P.B., Poosala, V., Ramaswamy, S.: The Aqua approximate query answering system. *SIGMOD Rec.* 28(2), 574–576 (1999)
4. Mestekemper, T., Windmann, M., Kauermann, G.: Functional hourly forecasting of water temperature. *International Journal of Forecasting* 26, 684–699 (2010)
5. Timms, G.P., de Souza, P.A., Reznik, L.: Automated assessment of data quality in marine sensor networks. In: *Proceedings of Oceans 2010 IEEE-Sydney*, pp. 1–5 (2010)
6. Deng, F.: Approximation algorithms for frequency related query processing on streaming data. PhD thesis (2007) ISBN: 978-0-494-32950-4
7. Madden, M.R., Franklin, M.J., Hellerstein, J.M., Hong, W.: TinyDB: an acquisitional query processing system for sensor networks. *ACM Trans. Database Syst.* 30(1), 122–173 (2005)

Addressing Challenges for Knowledge Discovery from Data in the Domain of Seaport Integration

Ana Ximena Halabi Echeverry and Deborah Richards

Department of Computing
Macquarie University
NSW 2109, Australia

{ana.halabiecheverry,deborah.richards}@mq.edu.au

Abstract. Discovering knowledge from data for decision making is dependent on the existence of data relevant to the decision at hand. For decisions in domains that involve many different factors and concerns, such as seaport integration, data may exist across many repositories managed by different organizations with different goals and foci, not to mention different data structures, entities, labels, units of measurement, categories and time periods. To use this data for decision making, approaches to combine the data and handle missing values are two of the problems, among others, that need to be addressed. In this paper we discuss the need for managing micro and macro-level data and our approach to handle missing values.

Keywords: Seaport Integration, Data Aggregation, Missing Values.

1 Introduction

Discovering knowledge from data for decision making is dependent on the existence of data relevant to the decision at hand. For decisions in domains that involve many different factors and concerns, such as seaport integration, data may exist across many repositories managed by different organizations with different goals and foci, not to mention different data structures, entities, labels, units of measurement, categories and time periods. In this paper we present an approach to address two key issues which will affect the quality of decision making in seaport integration and other domains: data aggregation and missing values. We further discuss the notions of macro and micro data to allow strategic/high-level decision making to be conducted when only operational/low-level data is available. In Section 2 we discuss the need to aggregate data from multiple sources and the role of macro and micro data to support strategic and complex decision making. In Section 3 we consider how to handle missing values in the context of identification of ports who were leaders in compliance with environmental standards. Conclusions and future work appear in Section 4.

2 Aggregating Data from Multiple Sources

Port authorities (PAs) tend to be concerned with operational decisions and have tended to make local decisions [8, 9]. However, the increasingly competitive global

environment demands that PAs engage in longer-term and higher-level decision making be undertaken. Key reasons why strategic decision making does not occur includes the lack of available data and models or approaches to analyse the data. In our investigations concerning seaport integration, it became quickly apparent that potentially relevant data exists in many different locations. This data may use different labels/names, units of measurement and time frames. Some concepts may overlap [1] and be difficult to match. We see in Figure 1 examples of data from just four of the relevant sources in the US seaport domain: U.S. Army Corps of Engineers, U.S. Census Bureau, US Department of Homeland Security and US Department of Transportation. Each of those repositories offers a hierarchical structure or set of modules of information, which address a certain level of decision-making for each individual institution. If a seaport authority wishes to make any decision by analysing those data sets, the process will involve disaggregate analysis that unavoidably results in losing various degrees of information. In Figure 1 the data gathered/supplied by the US Census Bureau represents aggregated and summarised data (i.e. macro level data and abstract/high level concepts such as “people and households” and “geography”). Different colours indicate that some variables concern different types of decisions and different subsystems (discussed further below) that comprise the seaport domain.

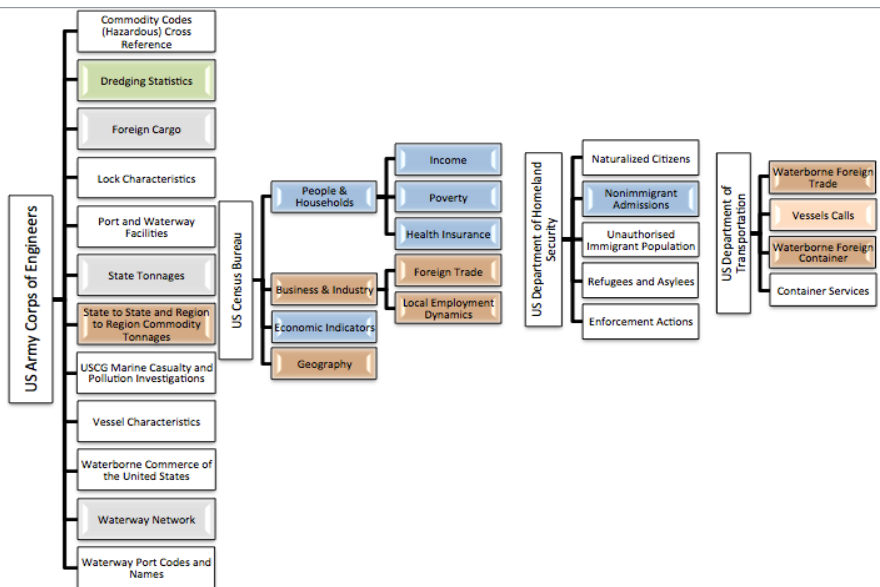


Fig. 1. Macro data repositories on the US Data websites¹

¹US Data sites: www.marad.dot.gov/library_landing_page/data_and_statistics/Data_and_Statistics.htm; www.bea.gov/international/index.htm; www.ndc.iwr.usace.army.mil/; www.dhs.gov/xlibrary/assets/statistics/yearbook/2008/ois_yb_2008.pdf

Figure 1 categorises the data according to its source. However, we could take an alternative approach which collects the data based on the type of decision that is to be made. We developed a systemic model which we call Port-Decision System Approach (PDSA) [3] which includes a number of subsystems to describe the seaport domain. Economic (ES) – shaded dark blue, Factors of productions and technology (FPT) – shaded brown, Global and environmental processes (GEP) – shaded green, Preference and experience (PE) – shaded skintone, Population and social structure (PSE) – shaded light blue and Political system institutions (PSI) – shaded purple. To make decisions concerning each of these subsystems it is necessary to extract the data from different sources and aggregate it by subsystem, shown for example in Figure 2.

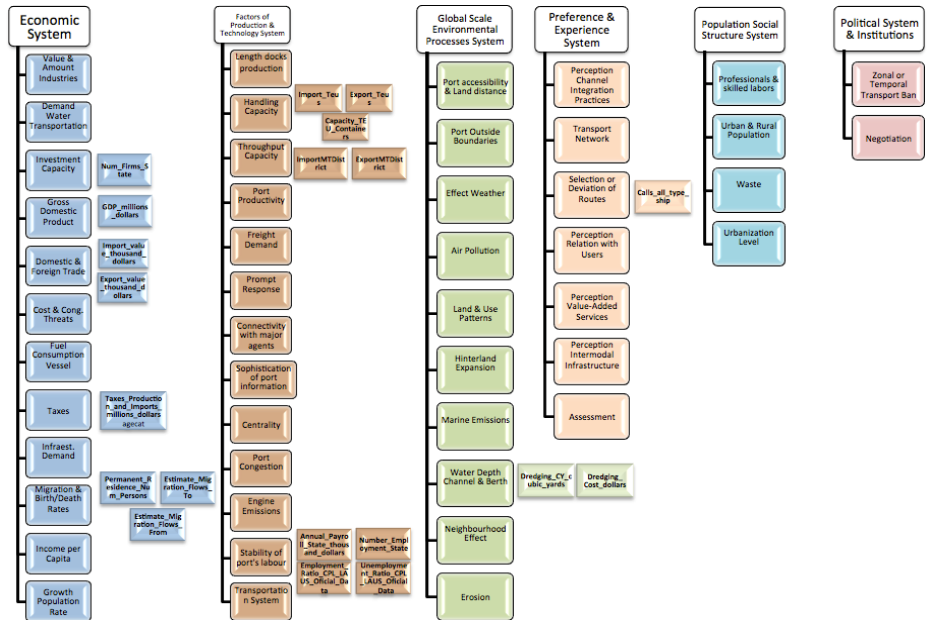


Fig. 2. Micro-level data repositories on the US Data websites

Looking closely at Figure 2 we can identify many low-level variables that have been compiled from multiple sources from the US data websites. Currently the decision maker is not necessarily aware that multiple hierarchies of data exist and would typically not have the skills or resources to combine the repositories to analyse the hierarchies. Our study involves exploration of these heterogeneous repositories in the quest for integrating data for analysis using data mining techniques so that evidence based guidance is provided for decision making.

Finding a way to connect macro and micro-level data will be important to aid strategic decision making. Strategic decisions, such as whether to expand the workforce, tend to concern macro level goals and data. However, data tends to be captured at the micro or operational level, such as number of employees and turnover rates. As a result there may be a mismatch between using micro level data for macro level decision making. On the one hand, it can be argued that the greater the level of abstraction of concepts represented in the model the more comprehensive the approach and

widely applicable the model will be to the phenomenon under study. However, a detailed representation of the model involving low level concepts (even instances) enhances its interpretability when implementing its outcomes in the real world. Table 1 shows how different level variables can map to subsystems and one another. In the next subsection we consider approaches in the literature and an approach using graph theory.

2.1 Data Aggregation Approaches in the Literature

There are techniques from the management field that consider how to handle the problem of data aggregation for decision making. Three of these techniques are: 1) multi-attribute value theory (MAVT), 2) aggregation of information based on indicators and 3) data level aggregation based on modelling abstraction. As described and used in [10], in this approach the attributes are associated to sub-attributes using expert weights ($W_{i,j}$) and an additive value function $Vc(a_i)$ that values an score between the preference of association and the given weight, illustrated in Figure 3.

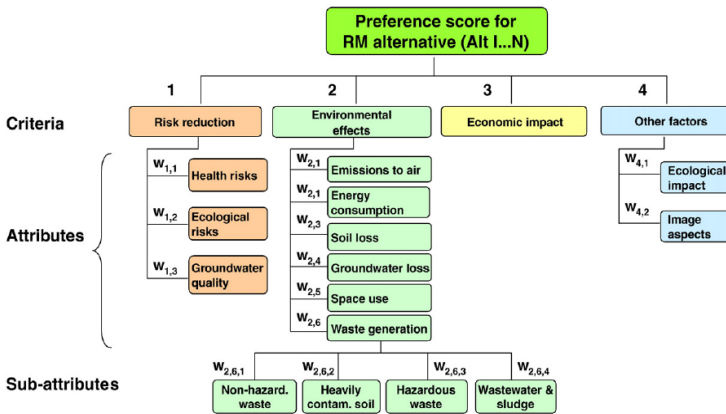


Fig. 3. Hierarchy of data aggregation based on MAVT [10]

A second approach on aggregation of information based on indicators has its roots in economic studies in which successive aggregation of scores are formed from different levels of indexes and sub-indexes. The 2011 World Economic Forum in their Global Competitiveness Report [12] uses this concept to report a structured computation of information. Formally, each sub-index represents a lower factor which can be measured from a data sample. The index is the weighted average of two or more sub-indexes. Finally, an indicator provides the higher factor which corresponds to an indication of the index worst and best possible outcomes. A third approach corresponds to typical data structures. Borshev & Filippov [2] state that in general, aggregate values are used to model higher abstraction problems such as transportation networks. A decrease in the aggregation is performed when modelling problems use data to model exact sizes, distances, velocities and timings matter, as illustrated in Figure 4.

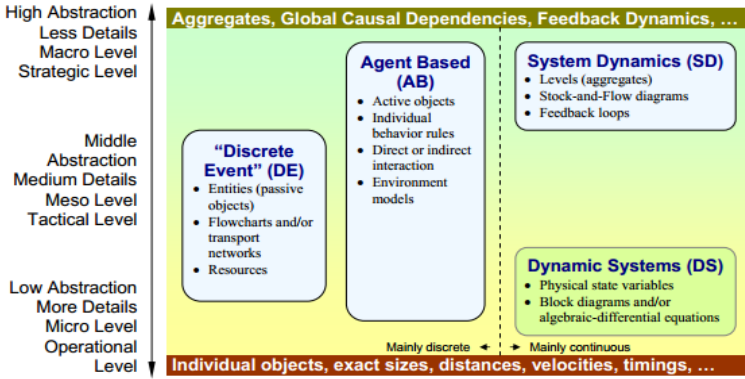


Fig. 4. Approaches in modelling according to the data level abstraction [2]

2.2 A Data Aggregation Approach Using Graph Notation

The previous approaches provide alternative solutions for data aggregation at different abstraction levels. Here to handle these different levels we suggest the use of graph theory to deal with hierarchical data structures. Graphs also provide visual benefits (Figure 5 shows a configuration based on the Table 1 formalisation)

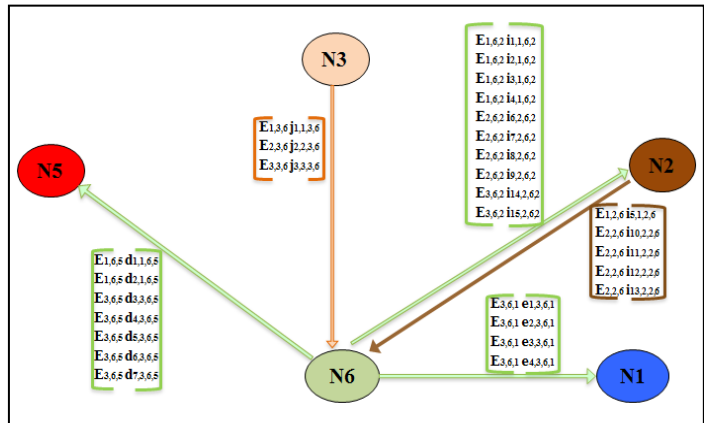


Fig. 5. Graph sample configuration

for understanding complex associations that otherwise need to be explained through complex analytical methods.

A mathematical definition of a graph G corresponds to a collection of vertices or nodes and edges that connect pairs of vertices. Suppose N denotes data at the macro level. This level aggregates concepts into categories that represent complex systems. $N_{i,j}$ is the pair of nodes denoting origin and destination of data in the macro level status, for example, $N_{6,2}$ traces a line from GEP (N_6) to FPT (N_2). E_k denotes the second-level data following the $N_{i,j}$ pathway, for example, $E_{1,6,2}$ denotes the concept for air pollution/emissions that relates with environment and production

systems. Finally, we can drill further down to find the micro-level data named here as edges: $a_{l,k,i,j}$, $b_{l,k,i,j}$, $c_{l,k,i,j}$, $d_{l,k,i,j}$, $e_{l,k,i,j}$, $f_{l,k,i,j}$, $g_{i,j}$, $h_{i,j}$, $i_{i,j}$, and $j_{i,j}$.

These edges display a cluster correlation of measurable variables which connect with the concepts described by the second-level data aggregation. We have been using data mining methods such as clustering and neural networks to identify relationships between variables and this work will be reported elsewhere.

Table 1. Formalisation - data hierarchies

Macro-level	N_i	Macro-level2	N_j	Second-level	$E_{-}(k,i,j)$	Micro-level	$a_{-}(l,k,i,j)$
GEP	N6	FPT	N2	Air pollution/emissions	E1,6,2	CO2	i1,1,6,2
						SO2	i2,1,6,2
						NOx	i3,1,6,2
						O3	i4,1,6,2
FPT	N2	GE	N6	Air pollution/emissions	E1,2,6	facilities	i5,1,2,6
PE	N3	GE	N6	Air pollution/emissions	E1,3,6	Scientist	j1,1,3,6
GEP	N6	PSI	N5	Air pollution/emissions	E1,6,5	O3comply	d1,1,6,5
						Inadequacies	d2,1,6,5
GEP	N6	FPT	N2	Water quality (Marine	E2,6,2	oils	i6,2,6,2
						chemicals	i7,2,6,2
						runoff	i8,2,6,2
						NMS	i9,2,6,2
FPT	N2	GEP	N6	Water quality (Marine	E2,2,6	dredgeOcean	i10,2,2,6
						needWtTreat	i11,2,2,6
						facilities	i12,2,2,6
						Inadequacies	i13,2,2,6
PE	N3	GE	N6	Water quality (Marine	E2,3,6	Scientist	j2,2,3,6
GEP	N6	ES	N1	Impacts of growth (land use patterns)	E3,6,1	CRP	e1,3,6,1
						MarketVal	e2,3,6,1
						LeaseNum	e3,3,6,1
						LeaseAccess	e4,3,6,1
GEP	N6	PSI	N5	Impacts of growth (land use patterns)	E3,6,5	GAPStatus1	d3,3,6,5
						GAPStatus2	d4,3,6,5
						GAPStatus3	d5,3,6,5
						GAPStatus4	d6,3,6,5
						CountyArea	d7,3,6,5
GEP	N6	FPT	N2	Impacts of growth (land use patterns)	E3,6,2	LandFarms	i14,3,6,2
						dredgeOcean	i15,3,6,2
PE	N3	GE	N6	Impacts of growth (land use patterns)	E3,3,6	Scientist	j3,3,3,6

3 Handling Missing Values

Most data integration systems focus on data aggregation. This issue is exacerbated by the fact, there are missing values affecting the different levels of aggregation. They are incorporated in any of the representations obtained and their analysis is useful to

facilitate knowledge discovery [7]. We discuss in this section our missing values approach after first describing the problem context of the example we provide.

3.1 Knowledge Discovery for the Environmental Dimension of Seaports

Many developments in methodology for incomplete data settings have predominately done in statistics. These methods need to be widely utilized in practice and thus we pose the question of how to arise new issues on missing values when conveying questions that PAs might want to answer in their deeds and duties. In previous work, we have identified data of a port with whom they should partner based on their compliance with environmental standards. Such a partnership can deliver competitive advantages and improved risk management performance. To identify who is compliant within the context of environmental management system standards (EMS) we need to identify what variables will be relevant. Key environmental issues are summarized in Table 2. The variables cover three main areas:

Reducing Air Pollution/Emissions including particulate matter (PM), nitrogen oxides (NO_x), sulfur oxides (SO_x), carbon dioxides (CO_2), nitrogen oxides (NO_x), sulphur dioxides (SO_2) and ozone expressed (O_3); **Improving Water Quality:**

Table 2. Observational dataset for missing value analysis

Selected Variable	Reducing Air Emissions	Improving Water Quality	Minimizing Impacts of Growth
needWtTreat		X	
facilities	X	X	
oils		X	
chemicals		X	
Inadequacies	X	X	
CO2	X		
O3comply	X		
O3	X		
O3cont	X		
SO2	X		
NOx	X		
CRP			X
LandFarms			X
Scientist	X	X	X
MarketVal			X
GAPStatus1			X
GAPStatus2			X
GAPStatus3			X
GAPStatus4			X
runoff		X	
CountyArea			X
LeaseNum			X
LeaseAcres			X
NMS		X	
dredgeOcean		X	X

Dredging activities (*dredgeOcean*), species habitat creation (national marine sanctuaries (*NMS*)); **Minimizing Impacts of Growth:** *CountyArea*. See Appendix in other paper by this author in this proceedings for descriptions of these variables.

Because these data are not always available, development of a missing value procedure is convenient for addressing several concerns caused by incomplete data. “Incomplete data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumption behind many statistical procedures is based on complete cases” [11. p.1]”

Table 3. Univariate Statistics for environment dataset

Univariate Statistics							
	N	Mean	Std. Deviation	Missing		No. of Extremes ^a	
				Count	Percent	Low	High
facilities	40	36.25	39.135	4	9.1	0	3
oils	40	153188.627	502043.6682	4	9.1	0	6
chemicals	40	3385.675	11996.3488	4	9.1	0	5
CO2	41	95060248.87	86657623.70	3	6.8	0	0
O3cont	41	77.4073	17.03626	3	6.8	4	3
SO2	41	224628.944	234729.5499	3	6.8	0	0
NOx	40	79626.03	67356.842	4	9.1	0	0
CRP	44	50.5568	23.01462	0	.0	0	6
LandFarms	44	111196.70	174025.981	0	.0	0	7
GAPStatus1	43	8191340.26	25835111.17	1	2.3	0	9
GAPStatus2	44	5205314.80	9915808.516	0	.0	0	9
GAPStatus3	44	12868766.57	21033110.76	0	.0	0	10
GAPStatus4	44	53293715.89	53579654.96	0	.0	4	9
runoff	44	331.2382	307.09730	0	.0	0	1
CountyArea	44	1770.34984	5120.920416	0	.0	0	4
LeaseNum	43	940.23	1775.410	1	2.3	0	5
LeaseAcres	43	5039253.37	9388547.853	1	2.3	0	5
dredgeOcean	40	8255772.28	10254649.39	4	9.1	0	2
LeaseArea	43			1	2.3		
needWtTreat	44			0	.0		
Inadequacies	40			4	9.1		
O3comply	41			3	6.8		
O3	41			3	6.8		
Scientist	44			0	.0		
MarketVal	44			0	.0		
NMS	43			1	2.3		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

3.2 Missing Value Analysis

In this section we want to consider the impact of missing covariate data in the analysis of data aggregation at different abstraction levels. Horton and Switzer [4] report in a review of missing data methods from 26 original articles, how infrequent a missing covariate data analysis (i.e. multiple imputation) appears in observational studies. The impact of missing values is embedded in the data structure and therefore its analysis is critical. Typically, the methodology of missing covariate data answers the following questions:

1. Where are the missing values located?
2. How extensive are they?
3. Do pairs of variables tend to have values missing in multiple cases?

4. Are data values extreme?
5. Are values missing randomly?

Table 3 displays a summary of missing values for the sample of variables considered. We see that some values are not missing at all, while other variables, such as *facilities* and *oils*, are missing around 9% of the time. We conveniently assessed the most common methods (i.e. listwise, pairwise, regression estimation) with the assumption that the pattern of missing values does not depend on the data values, i.e. the data is missing completely at random (MCAR). However, running Little’s [6] missing value test we conclude that significance value is less than 0.05 for our dataset. In this case data are not MCAR and then we need to use expectation-maximization (EM) estimation. EM depends on the assumption that the pattern of missing data is related to the observed data only (see Table 4). The overall summary of missing values is displayed in Figure 6 in three pie charts that show different aspects of missing values in the data. a) The variables chart shows that 14 of 24 variables have at least one missing value on a case. b) The cases chart shows that 11 of 44 cases have at least one missing value on a variable. c) The values chart shows that 40 of 1,056 values (cases x variables) are missing.

Table 4. Little’s MCAR test, EM means : Little’s MCAR test: Chi-Square=76.849, DF=56, Sig. = 0.34, The EM Algorithm failed to converge in 25 iterations

facilities	Oils	chemicals	CO2	O3cont	SO2
36.91	148940.68	3026.39	86651799.6	83.77	171032.42
NOx	GAPStatus1	LeaseNum	LeasesAcres	DredgeOcean	
63615.32	8192467.23	928.59	4976108.6	7062010.38	

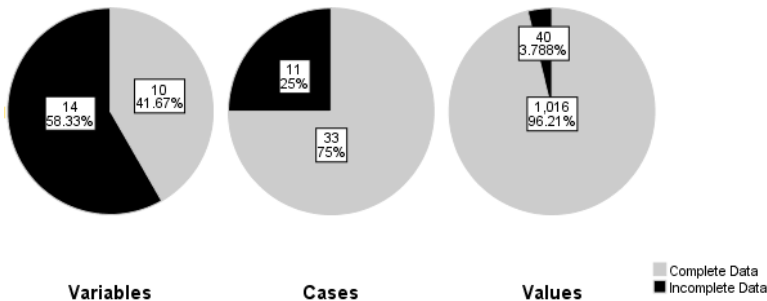


Fig. 6. Pie charts of summary of missing values

Table 5 indicates that three groups of variables record similar or related information: group 1 (*leaseNum, leaseAcres, NMS*), group 2 (*chemicals, oils, facilities, inadequacies*) and group 3 (*O3cont, SO2, CO2, Ocomply, NOx*). The table suggests that if we do not know the value of one variable within a group, probably we do not know the value for the other groups either.

The patterns chart in Figure 7 displays missing value patterns for the analysis variables. Each pattern corresponds to a group of cases with the same pattern of

Table 5. Patterns of missing data showing three groups. ^aVariables are sorted on missing patterns. ^bNo of complete cases if variables missing in that pattern (marked with X) are not used.

# Cases	Missing Patterns ^a														Complete
	GAPstat us1	Lease- Num	LeaseA- cres	NMS	O3Cont	SO2	CO2	O3Comp lv	Nox	chem- cials	oil	facilities	Inadequ- acies	dred- geOcean	
23															33
1									X						34
2													X		35
1	X												X		36
1										X	X	X	X	X	39
3										X	X	X	X		36
2					X	X	X	X	X						36
1		X	X	X	X	X	X	X	X						37

incomplete and complete data. For instance, pattern 4 represents cases that have missing values on group 3 (*O3cont*, *O3comply*, *O3*, *CO2*, *SO2*, *NOx*). The chart orders analysis and patterns to reveal where monotonicity exists. That is, there will be no “islands” of non-missing cells in the lower right portion of the chart and no “islands” of missing cells in the upper left portion of the chart. This dataset is nonmonotone and there are any values that would need to be imputed in order to achieve monotonicity.

The bar chart in Figure 8 shows that the majority of the cases in the dataset have pattern 1, i.e. the pattern for cases with no missing values. Patterns 2 and 4 represent missing values in around 5% of the cases. i.e., group 2 (chemicals, facilities, inadequacies, oils) and group 3 (*O3cont*, *O3comply*, *O3*, *CO2*, *SO2*, *NOx*) and pattern 6 that includes the variable *dredgeOcean*.

Estimated means are displayed in Table 5 for:

- The means from listwise deletion tend to be higher for group1 and group 2 whilst the means for chemicals, *CO2*, *CRP*, *GAPstatus1*, *GAPstatus3* and *LeaseNum* vary greatly. Because the data are not missing completely at random, estimates other than EM may be biased.
- The estimates for groups 2 and 3 with the greatest number of missing values include a large number of extreme values.

To observe if the distribution is more in line with the original data avoiding greater differences and random variations, it might be necessary to test the data to determine whether these values are not missing at random (MAR). Figure 9 displays multiple pairs of line charts, showing the mean and standard deviation of the imputed values of the variables chosen by the model as dependent at each iteration method for each of the 5 requested imputations. There should not be any patterns in the lines and look suitably random [10]. We see patterns that suggest the missing values are not random.

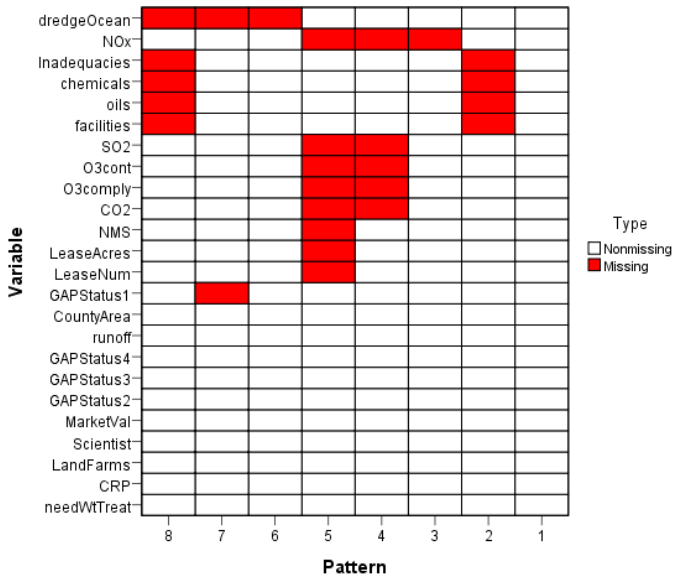


Fig. 7. Missing value patterns chart

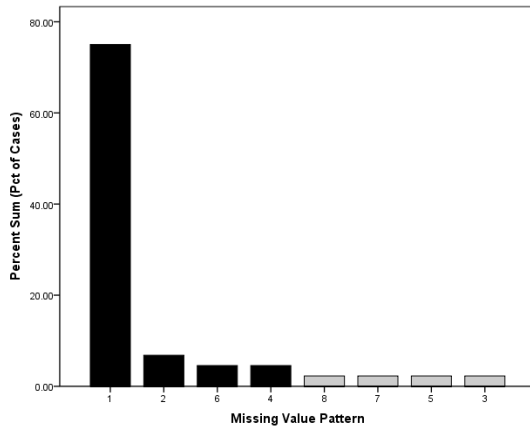


Fig. 8. Bar chart of missing value patterns

4 Conclusions and Future Work

Discovering knowledge from data for decision making is dependent on the existence of data relevant to the decision at hand. In the context of with whom PAs should partner based on their compliance with environmental management system standards (EMS), we have dealt with the maximum information from multiple levels and types of data, starting with macro-level data and ending with the micro-level data analysis.

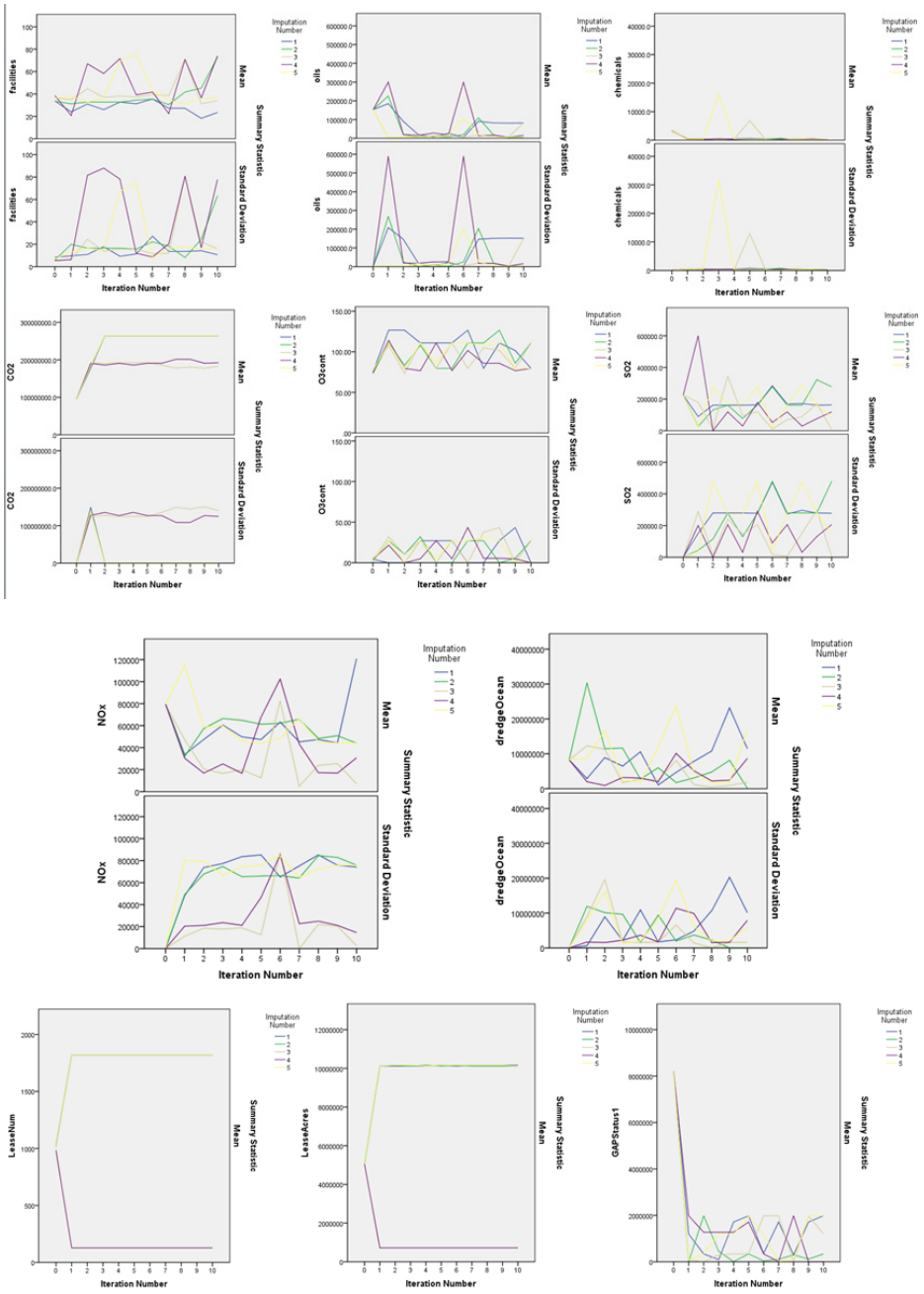


Fig. 9. Line charts to check if any patterns and that missing data are random

We are exploring the implementation of literature approaches on data aggregation such as [10], using graph notation to deal with hierarchical data structures and providing the visual benefits of graphs for understanding complex associations that otherwise need to be explained through complex analytical methods.

Missing value analysis suggests that if we do not know the value of one variable within a group, probably we do not know the value for the other groups either. The latter is corroborated our observations in that that dependency can be evident on variables pertaining to the same second-level of aggregation. That is, within the sample there is a correspondence of groupings displayed in the formalisation aggregation and the missing value pattern instances.

We will be conducting further analysis of the PDSA using time series data in a more comprehensive dataset for Latin American seaports in the quest to identify the legal, technical and political factors and associations that affect the decision making process of regional port authorities.

References

1. Bichou, K., Gray, R.: A critical review of conventional terminology for classifying seaports. *Transportation Research Part A* 39, 75–92 (2004)
2. Borshchev, A., Filippov, A.: From System Dynamics and Discrete Event to Practical Agent Based Modeling: Reasons, Techniques, Tools. In: *The 22nd International Conference of the System Dynamics Society*, Oxford, England (2004)
3. Halabi, A., Richards, D., Bilgin, A.: Proposing a port decision system approach for dynamic integration of South American sea ports. Paper Presented at the International Conference on Advances in ICT for Emerging Regions, ICTer (2011)
4. Horton, N.J., Switzer, S.S.: Statistical methods in the Journal. *New England Journal of Medicine* 353(13), 1977–1979 (2005)
5. Kruse, C.J.: Environmental Management Systems at Ports - A new initiative. In: *Proceedings of the 14th Biennial Coastal Zone Conference* (2005)
6. Little, R.J.A.: A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* 83, 1198–1202 (1988)
7. McClean, S., Scotney, B.: Using evidence theory for the integration of distributed databases. *Int. J. Intell. Syst.* 12, 763–776 (1997)
8. Notteboom, T.E.: Concentration and the formation of multi-port gateway regions in the European container port system: an update. *Journal of Transport Geography* 18(4), 567–583 (2010)
9. Notteboom, T.E., Rodrigue, J.-P.: Port regionalization towards a new phase in port development. *Maritime Policy and Management* 32(3), 297–313 (2005)
10. Sorvari, J., Seppälä, J.: A decision support tool to prioritize risk management options for contaminated sites. *Science of the Total Environment* 408, 1786–1799 (2010)
11. SPSS, Missing Value Analysis 16.0, Chicago, USA (2007)
12. World Economic Forum, *The Global Competitiveness Report*, Geneva, Switzerland (2011)

Data Envelopment Analysis for Evaluating Knowledge Acquisition and Creation

Chuen Tse Kuah and Kuan Yew Wong

Department of Manufacturing and Industrial Engineering, Faculty of Mechanical Engineering,
Universiti Teknologi Malaysia, 81310 UTM Skudai, Malaysia
joseph.kuah84@gmail.com, wongky@fkm.utm.my

Abstract. From a managerial perspective, a model to measure the performance of knowledge acquisition and creation in organizations has been created based on the Data Envelopment Analysis (DEA) methodology. An application in higher educational institutions (HEIs) is shown. The model is found suitable for this purpose and is able to give some important insights to managers on what areas and to what extent they should improve in order to become efficient.

Keywords: Data Envelopment Analysis (DEA), Performance measurement, Knowledge creation, Knowledge acquisition, Knowledge management.

1 Introduction

An efficient knowledge management is crucial for an organization to achieve sustainable competitive advantages. Knowledge acquisition and creation are two of the most important elements in knowledge management. Knowledge acquisition is the process where an organization imports knowledge and expertise from external sources. On the other hand, knowledge creation refers to the process where the workers generate new knowledge, ideas, solutions, products, and services.

In this paper, these two elements are assessed collectively based on the fact that the ultimate outcome of knowledge acquisition is the creation of new knowledge. Evaluating them together would give management an overall picture on both areas.

The goal and originality of this paper is to develop a measurement model based on the Data Envelopment Analysis (DEA) methodology to evaluate these two elements in organizations. Some basic concepts of DEA are reviewed next. An explanation of the developed model follows. An actual application is then elucidated and discussed. Finally, conclusions and future research directions are drawn based on the findings.

2 Original DEA Models

DEA is a mathematical model for measuring relative efficiencies of a group of homogeneous Decision Making Units (DMUs). It minimizes subjective judgments and is capable of handling multiple inputs and outputs. Assuming that there are n DMUs,

each with m inputs and s outputs, the relative efficiency score of a test DMU₀ is obtained using the following model [1]:

$$\begin{aligned}
 \text{Max } \varepsilon_0 &= \frac{\sum_{r=1}^s u_r y_{r0}}{\sum_{i=1}^m v_i x_{i0}} \\
 \text{s.t. } \frac{\sum_{r=1}^s u_r y_{rj}}{\sum_{i=1}^m v_i x_{ij}} &\leq 1, \quad \forall j \\
 u_r, v_i &> 0, \quad \forall r, i
 \end{aligned} \tag{1}$$

where,

- $r = 1$ to s ,
- $i = 1$ to m ,
- $j = 1$ to n ,
- y_{rj} = amount of output r produced by DMU _{j} ,
- x_{ij} = amount of input i consumed by DMU _{j} ,
- u_r = weight assigned to output y_r ,
- v_i = weight assigned to input x_i .

Fundamentally, for a test DMU₀, Model (1) compares the inputs and outputs among all DMUs and determines the optimum set of weights (u_r and v_i) which would give DMU₀ the highest possible efficiency score ε_0 , while constraining the efficiency scores of all DMUs to be within 1. The model is run n times to determine the efficiency scores for all DMUs. $\varepsilon_0 = 1$ indicates that a particular DMU is efficient, while a value less than 1 means it is inefficient.

Model (1) can be converted into its dual form, Model (2), which is also known as the envelopment form in DEA [1]. For a guideline on how to transform Model (1) into Model (2), readers are referred to [2].

$$\begin{aligned}
 \text{Min } \theta_0 \\
 \text{s.t. } \sum_j \lambda_j x_{ij} - \theta_0 x_{i0} &\leq 0, \quad \forall i \\
 \sum_j \lambda_j y_{rj} - y_{r0} &\geq 0, \quad \forall r \\
 \lambda_j &\geq 0
 \end{aligned} \tag{2}$$

Model (2) has a feasible solution of $0 < \theta \leq 1$ and the optimal solution of a test DMU₀ is $\theta_0 = 1$, $\lambda_0 = 1$, and $\lambda_j = 0$ ($j \neq 0$). In other words, an efficient DMU has a score of $\theta = 1$; while inefficient DMUs have scores of $\theta < 1$. For each inefficient DMU, Model (2) identifies a set of corresponding efficient DMUs as benchmarks for improvement. The reference sets for inefficient DMUs are identified from the non-zero λ values. In addition, for an inefficient DMU, DEA proposes improvement targets either by reducing the inputs by multiplying with θ_0 while maintaining the output levels, or by increasing the outputs by multiplying with $1/\theta_0$ while maintaining the input levels.

Model (2) is generally preferred than Model (1) because it is less computational cumbersome. This can be reflected from the constraints of the models. The constraints of Model (1) are more complicated than those of Model (2). Furthermore, Model (2) is favored because it can identify reference sets for the DMUs as described above. It should be noted that both efficiency scores, ε_0 and θ_0 , obtained from the two models are identical.

In short, the main function of DEA is as an analytical tool to assess and benchmark the performance of various DMUs.

3 Developed Model for Knowledge Acquisition and Creation Performance Measurement

DEA serves as a suitable tool to evaluate the performance of knowledge acquisition and creation in an organization by viewing it as a process that converts multiple inputs into multiple outputs. These input and output data are analyzed using a performance measurement model developed based on Model (2). The results of the analysis will be the performance scores of all DMUs and improvement targets for those inefficient ones. The conceptual framework of the evaluation model is illustrated in Fig. 1.

One important issue in performing an analysis using DEA is determining what input and output data to be used. Thus, a review on the past literature has been done. Tables 1 and 2 summarize the measures, their references, and descriptions. Note that the list is not meant to be distinctive and can be edited based on managerial opinions.

Next, to propose improvement targets for inefficient DMUs, the outputs are to be increased by multiplying with $1/\theta_0$, while the inputs remain unchanged. Reducing inputs is undesirable because knowledge workers, as an input, are one of the most valuable assets of an organization. The improvement targets are formulated as:

$$\hat{y}_r = y_r \times \frac{1}{\theta_0} \quad (3)$$

4 An Application

An application of the developed model will be demonstrated in higher educational institutions (HEIs). Higher education is a knowledge-intensive industry and thus it

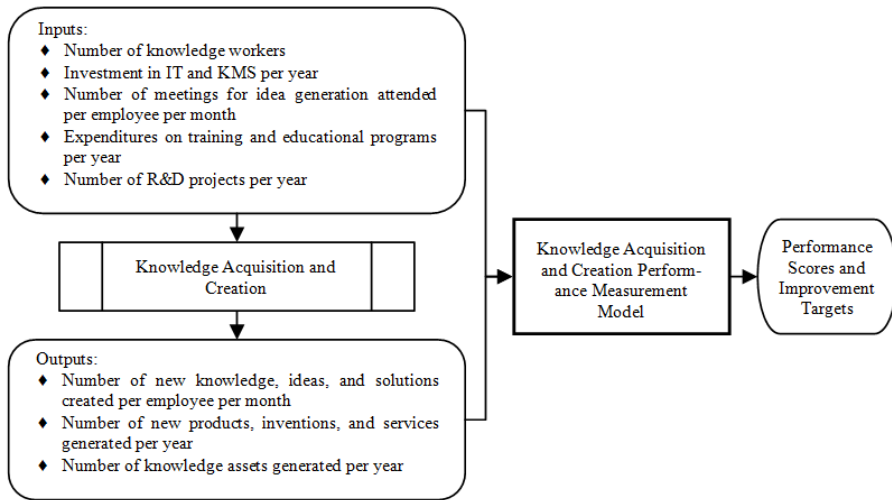


Fig. 1. Conceptual framework of knowledge acquisition and creation performance measurement model

Table 1. Input measures

Measures and References	Descriptions
x_1 : Number of knowledge workers [3-9]	Knowledge workers are one of the fundamental elements of knowledge acquisition and creation. They acquire and generate new knowledge, ideas and solutions. A worker's mind itself is a developer and reservoir of tacit knowledge. They solve problems and make important decisions to improve the organizational performance.
x_2 : Investment in IT and KMS per year [4-5], [8-11]	Information technology (IT) and knowledge management system (KMS) are the two basic architectures of knowledge discovery. With these, workers can rapidly search, acquire, extract, and retrieve knowledge. Moreover, IT and KMS support the collaborations and communications among the workers and enable the formation of virtual communities of practice (CoPs) both internally and externally which are important for knowledge acquisition and creation.
x_3 : Number of meetings for idea generation attended per employee per month [3], [9], [12-15]	Examples of idea generation meetings are brainstorming and strategic meetings. In such meetings, new knowledge and ideas would be sparked and generated through interactions and discussions among the workers.
x_4 : Expenditures on training and educational programs per year [3-11], [13-14]	Ongoing training and educational programs are means to transfer up-to-date knowledge to the workers. External trainers can also be hired to give training sessions on special knowledge. This has proven to be an effective way of acquiring external knowledge and diffusing it to the target audiences. After the workers have acquired new knowledge, their personal knowledge bases are enhanced and more new ideas and knowledge can be generated.
x_5 : Number of R&D projects per year [4-5], [9], [16]	An organization's success is greatly influenced by its innovations. R&D projects are necessary for an organization to create new products, inventions and services. The number of R&D projects serves as a proxy measure for the level of effort of an organization in developing new knowledge.

Table 2. Output measures

Measures and References	Descriptions
y_1 : Number of new knowledge, ideas, and solutions created per employee per month [3], [6-7], [15]	New knowledge, ideas, and solutions are created by the knowledge workers via the process of knowledge creation. In addition, by acquiring knowledge externally, new knowledge, ideas and solutions may be imported into a company as well.
y_2 : Number of new products, inventions, and services generated per year [3-6], [8-9]	New products, inventions, and services can be generated via knowledge acquisition and creation. Particularly, the outcomes of R&D projects are new products and services which can improve an organization's competitiveness and increase its market share.
y_3 : Number of knowledge assets generated per year [3-9], [16]	Another output of knowledge acquisition and creation is the generation of knowledge assets such as patents, copyrights and scientific publications. By leveraging its knowledge assets, an organization can achieve sustainable competitiveness.

serves as a perfect test subject for the model. This section explains the implementation of the model to assess HEIs' knowledge acquisition and creation performance.

A survey was conducted using a specially designed questionnaire to collect the data needed. It was conducted through mails within Malaysia. Firstly, the recipients were sampled from the Malaysian Ministry of Higher Education's online database. Next, the questionnaire was sent to potential respondents along with an explanation cover letter. The respondents chosen were presumably in a position to comment on their institutions' knowledge management and have access to the information needed.

At the end of the survey, 23 usable responses were obtained. In this study, the data were used to compute relative efficiencies of the HEIs. Response rate does not have effects on the results' accuracy, and thus it is not a concern as long as the responses are sufficient for the analysis.

A MATLAB program was written based on the developed model. The data were analyzed using the program to obtain the performance score of each HEI. Results are summarized in Tables 3 and 4. Table 3 shows the performance score and ranking along with the reference set for each DMU. Table 4 presents the improvement targets for the inefficient DMUs.

DMUs with a score of 1 are efficient, while those score less than 1 are considered inefficient. From Table 3, it can be observed that performance scores of the DMUs range from 0.1501 to 1, with an average score of 0.6431. Out of 23 DMUs, 7 are efficient and 16 are inefficient. As additional information, the third column shows the ranking of the DMUs based on their scores. From this piece of information, the organizations can know where they are positioned relatively to their competitors in the same industry and take it as a motivation to improve their performance.

Also recorded in Table 3 are the corresponding λ values of the reference sets. The greater the λ value means the referred DMU is closer to the DMU under evaluation in terms of their input-output data. This information is useful for an organization to know which efficient DMUs it is being benchmarked with, so that it can improve

itself by learning from them. For example, DMU₃'s performance score is 0.64, and from Table 3, its manager can know that the efficient DMUs it is being benchmarked with are DMU₁₆ and DMU₂₂ with λ values of 0.69 and 1.25 respectively. By understanding the operations of these 2 DMUs, appropriate strategies can be devised to improve its knowledge acquisition and creation. Furthermore, the manager can choose to focus more on DMU₂₂ because of its larger λ value.

Table 3. Performance scores and reference sets of DMUs

DMU	Score	Rank	Reference Set											
			DMU	λ	DMU	λ	DMU	λ	DMU	λ	DMU	λ		
1	1.0000	1	1	1.00										
2	1.0000	1	2	1.00										
3	0.6400	11	16	0.69	22	1.25								
4	0.3345	20	5	0.02	22	0.49								
5	1.0000	1	5	1.00										
6	0.5054	13	2	0.01	5	0.01	13	0.38	22	1.29				
7	0.8157	10	1	0.11	13	1.27	22	0.09						
8	0.2545	21	13	0.01	16	0.47								
9	0.8341	9	13	0.27	16	0.47								
10	0.2400	22	13	0.24	16	0.10	22	1.99						
11	1.0000	1	11	1.00										
12	0.5555	12	2	0.12	11	0.01	13	0.51						
13	1.0000	1	13	1.00										
14	0.3604	19	2	0.01	11	0.01	13	0.28	22	0.35				
15	0.1501	23	13	0.98	16	0.22	22	0.64						
16	1.0000	1	16	1.00										
17	0.8889	8	16	0.25										
18	0.4308	17	5	0.01	13	0.05	16	0.54	22	0.32				
19	0.4630	15	2	0.02	11	0.02	13	0.10	16	0.10	22	0.47		
20	0.4551	16	1	0.01	16	0.18								
21	0.4895	14	5	0.01	13	0.01	22	1.49						
22	1.0000	1	22	1.00										
23	0.3735	18	1	0.35	13	0.95	22	0.60						

Improvement targets were determined for every inefficient DMU as recorded in Table 4. These targets can be used by an institution as a guideline for future improvements. Take DMU₇ as an example, its performance score is 0.8157, thus the output levels have to be improved by 22.6% ($1/0.8157 = 1.226$). Its improvement targets are therefore $\hat{y}_1 = 8$, $\hat{y}_2 = 24$, and $\hat{y}_3 = 323$. In order for DMU₇ to be efficient, it has to increase these measures respectively while maintaining the same input levels. With this information on hand, the manager can then decide on how to channel the resources into specific improvement initiatives.

Table 4. Improvement targets for inefficient DMUs

DMU	Improvement Targets		
	\hat{y}_1	\hat{y}_2	\hat{y}_3
3	55	18	8
4	15	3	30
6	40	26	111
7	8	24	323
8	12	8	8
9	4	4	70
10	63	21	63
12	2	64	182
14	12	14	14
15	27	54	200
17	3	3	3
18	24	12	24
19	18	22	11
20	3	9	3
21	45	5	21
23	27	27	268

5 Conclusions

This paper has presented a performance measurement model for knowledge acquisition and creation using DEA. It proves to be a suitable model to evaluate these aspects effectively and conveniently. The information obtained from the developed model could help organizations to identify the inefficient areas and improvement targets in order to become efficient. These can be done by referring to their corresponding efficient benchmarked DMUs and the improvement targets.

The model has been tested in HEIs, which represent a highly knowledge-based industry. However, since knowledge acquisition practices may vary from one industry to another, it is necessary to test the model in other industries. In addition, though the measures proposed in this paper are as generic as possible to ease their future applications in other areas, they should be reevaluated based on different industries and modified wherever necessary.

Another element that can be included in future studies is finding the best practices and critical success factors of knowledge acquisition and creation in one industry. By collecting additional information such as what techniques and practices that organizations have implemented and upon obtaining their performance scores, it should shed some lights on which of the techniques and practices are indeed leading the organizations toward effectiveness and sustainable competitive advantages.

Acknowledgements. The authors would like to thank Universiti Teknologi Malaysia (UTM) for supporting this research.

References

1. Charnes, A., Cooper, W.W., Rhodes, E.L.: Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research* 2, 429–444 (1978)
2. Vanderbei, R.J.: *Linear Programming: Foundations and Extensions*, 3rd edn. Springer Science+Business Media, New York (2008)
3. Ross, J., Ross, G., Dragonetti, N., Edvinsson, L.: *Intellectual Capital: Navigating in the New Business Landscape*. New York University Press, New York (1998)
4. Von Krogh, G., Roos, J., Kleine, D.: *Knowing in Firms: Understanding, Managing, and Measuring Knowledge*. Altamira Press, Walnut Creek (1999)
5. Ahmed, P.K., Lim, K.K., Zairi, M.: Measurement Practice for Knowledge Management. *The Journal of Workplace Learning* 11(8), 304–311 (1999)
6. Robinson, H.S., Carrillo, P.M., Anumba, C.J., Al-Ghassani, A.M.: Performance Measurement in Knowledge Management. In: Anumba, C.J., Egbu, C.O., Carrillo, P.M. (eds.) *Knowledge Management in Construction*, pp. 10–30. Blackwell Publishing, Oxford (2005)
7. Wen, Y.F.: An Effectiveness Measurement Model for Knowledge Management. *Knowledge-Based Systems* 22(5), 363–367 (2009)
8. Holtshouse, D.: Knowledge Work 2020: Thinking Ahead about Knowledge Work. *On the Horizon* 18(3), 193–203 (2010)
9. Jafari, M., Rezaenour, J., Akhavan, P., Fesharaki, M.N.: Strategic Knowledge Management in Aerospace Industries: A Case Study. *Aircraft Engineering and Aerospace Technology* 82(1), 60–74 (2010)
10. Wong, K.Y., Aspinwall, E.: An Empirical Study of the Important Factors for Knowledge Management Adoption in the SME Sector. *Journal of Knowledge Management* 9(3), 64–82 (2005)
11. Wong, K.Y.: Critical Success Factors for Implementing Knowledge Management in Small and Medium Enterprises. *Industrial Management and Data Systems* 105(3), 261–279 (2005)
12. Wiig, K.: *People-focused Knowledge Management*. Elsevier, Oxford (2004)
13. Dalkir, K., Wiseman, E., Shulha, M., McIntyre, S.: An Intellectual Capital Evaluation Approach in a Government Organization. *Management Decision* 45(9), 1497–1509 (2007)
14. Geisler, E.: The Metrics of Knowledge: Mechanisms for Preserving the Value of Managerial Knowledge. *Business Horizons* 50(6), 467–477 (2007)
15. Josune, S., Nekane, A., Olga, R.: Knowledge Sharing and Innovation Performance: A Comparison between High-tech and Low-tech Companies. *Journal of Intellectual Capital* 10(1), 22–36 (2009)
16. Jiang, X., Li, Y.: An Empirical Investigation of Knowledge Management and Innovative Performance: The Case of Alliances. *Research Policy* 38(2), 358–368 (2009)

A High-Order Hidden Markov Model for Emotion Detection from Textual Data

Dung T. Ho and Tru H. Cao

Ho Chi Minh City University of Technology
and John von Neumann Institute - VNUHCM
hotrdung@gmail.com, tru@cse.hcmut.edu.vn

Abstract. Emotion detection from text is still an appealing challenge. The approaches to this problem have been done firstly based on just emotional keywords, and then extended with utilizing also other generic terms. However, they still lack of some useful semantic features, such as a psychological characteristic that emotion is the result of a mental state sequence. Recent works focus on using rules to exploit those features, but have the coverage problem. In this paper, we propose a method using the high-order Hidden Markov Model whose states are automatically generated to model the process that a mental state sequence causes an emotion. Our experiments on the ISEAR dataset have shown a better result in comparison with the state-of-the-art methods.

1 Introduction

Nowadays, computers are good to a certain degree at understanding human natural language, even at the semantic level, thanks to the achievements in the field of natural language processing. However, computers still misunderstand human language due to many obstacles. One of them is that what a human expresses may have various meanings depending on his emotion. Hence, it is necessary to make computers able to recognize human emotion so that they could understand human language better. This emotion detection problem, firstly introduced by Picard in 1997 as Affective Computing [13], is an appealing challenge. It can be used for various applications such as improving human-computer interaction, computer tutors, expressive text-to-speech engine and games, etc. [13] Nevertheless, though there have been many works with different approaches to this problem, more improvements are still required.

One criterion that can be used to classify those approaches is the kind of source they use to detect emotion. An intuitive source is multi-modal data, which includes voices, facial expressions, gestures, etc. Besides, emotion detection from textual data still attracts many works, because a large proportion of information stored in computers, as well as on the Internet, is in the textual form [9].

Emotion is such an abstract concept that there is still no proper definition for it [8]. Thus, the representation for emotion that computers can understand and evaluate should be the next thing to consider. For emotion representation, many studies in psychology have agreed on the categorical model, which classifies emotion into

discrete categories. A well-known example for this is Ekman's six basic universal emotions [5]. The more recent one is the dimensional model in [15], which considers an emotion as an entity constituted by some particular features. Each feature is modeled by one dimension, so each emotion can be represented as a point in that dimensional space based on its feature values. Generally, most of works focus on the categorical model, but how many and which categories should be used is commonly agreed to be different depending on a particular application domain [10].

Intuitively, the first approach to detect emotion from text is to identify emotional keywords. However, there are sentences that have emotion but contain no keyword so that cannot be detected by this way [9]. A remarkably successful approach to solve this problem is to use the LSA model to exploit the hidden semantic relation among keywords and other "generic" terms, thereby the emotion for those terms can be estimated so that they can also be used as marks to identify emotion [17]. However, since this is a bag-of-words model, the order of those marks is ignored. According to [13], the works in psychology has shown that which emotion a human would have depends on his mental state at that time, and more general, the sequence of prior mental states. Therefore, that order is significant to determine the emotion. Such a psychological characteristic of emotion is the focus of recent works. Those approaches use manually deduced detection rules that examine those characteristics to determine the emotion. Their drawback is low coverage, because to define a sufficient rule set manually, even for a specific domain, is not a trivial task.

In this paper, we propose a method that uses a high-order Hidden Markov Model (HMM) for emotion detection from text. Although the HMM has been used for emotion detection from multi-modal data, there has been no work that uses it for textual data. The key idea is to transform the input text into a sequence of events that cause mental states, then use the HMM to model the process that state sequence causes the emotion. The HMM is automatically constructed based on a training dataset.

The rest of this paper is organized as follows. In the next section, we describe more detail about related works on emotion detection from text and indicate their advantages as well as their drawbacks. Section 3 presents our emotion detection method with details of the important steps in the process. Section 4 describes our experiments and presents the result of the system evaluated with the ISEAR dataset. The last section summarizes our contribution and future work.

2 Related Works

In the keyword-based approach, the first task to do is to construct the lexicon of emotional keywords. This can be done by picking up from the dictionary for each emotion a set of words that express that emotion obviously. Then, these sets can be expanded based on some of word relations (synonym, hypernym in WordNet¹). Many publicly available resources have been created such as WordNet-Affect [18] and

¹ The information about WordNet can be found at <http://wordnet.princeton.edu>

SentiWordNet [7]. After that, emotion detection for a text is done by identifying emotional keywords in that text. The most recent works have also taken into account some basic linguistic information of the text to improve the detection. [1] makes use of a parser to analyze the sentence structure, the tense, the referred person to apply some fixed rules to determine whether a sentence with emotional keywords is actually has emotion. [2] uses a parser to find the head word of the sentence, which has a major impact on the emotion expressed in that sentence, as well as the contrasts such as negative words that may change the emotion. This approach is straightforward, but its accuracy highly depends on the quality of the emotional keyword sets. Besides, the problem of polysemy also affects the detection. However, the most weakness of this approach is that those sentences that have emotion but do not contain any emotional keyword cannot be detected [9].

To overcome that weakness, [17] proposes a method that uses LSA to reduce the dimension number of the Vector Space Model (VSM) to exploit the hidden semantic relation among terms. Hence, by evaluating the semantic similarity among “generic” terms, which have no explicit emotion, and predefined affective terms (in WordNet-Affect [18]), those “generic” terms could also be assigned an emotion. In particular, each emotion has a pseudo-document constructed from the affective terms that express that emotion directly (together with their synset), and then the representing vector of that document, as well as that emotion, is calculated. An input text is also represented by an LSA vector to be compared to the representing vector of each emotion by the cosine similarity. Thereby the most likely emotion for it can be determined. After that, [10] compares the LSA with other dimension reduction variants, namely, PLSA and NMF. Their evaluation on the ISEAR dataset shows that using PLSA yields the best result. In general, by having “generic” terms as extra marks, this approach can yield a better result than the first approach that uses only emotional keywords. However, as mentioned above, this approach has not taken into account psychological characteristics of emotion.

Based on the works in psychology, [19] constructs a predefined set of Emotion Generation Rules (EGR). From a training dataset in which each sentence is tagged with the most suitable EGR, they manually deduce a set of sequences that consist of semantic labels and concepts. Each concept is then replaced by the attributes defined for it in a domain-specific ontology. Afterwards, a set of Emotion Association Rules (EAR) can be extracted from that sequence set using the a-priori algorithm. Finally, the most appropriate rule to identify emotion for an input text is determined by the Separable Mixture Model. This is a novel method, but has the problem of coverage, because it is hard to have a sufficient and precise ontology, as well as to define EGR manually for every situation in practice and to construct an annotated dataset for those EGRs. Another work in [8] shows that most emotions are expressed with the presence of causes. Therefore, they use some predefined rules based on syntactic structures to extract emotion causes from the sentences that contain an emotional keyword. Then those emotion causes are used as a clue to recognize emotion. It requires linguistic analysis to derive emotion causes extracting rules, thus also costs much manual labor.

3 Proposed Method

Our method is based on the idea that the emotion depends on the human mental state [13], and the other idea that it is caused by emotional events [8]. We formalize these two ideas as the process of emotion invocation. This process starts with a certain mental state and transitions to another state when an event occurs based on both the current state and that event, and so on. In each mental state, an emotion might be invoked or not.

To model this process, we use a high-order HMM. Each HMM state corresponds to a certain mental state, and the symbols that it can emit correspond to the events that could cause that mental state. For every text, we consider it as a sequence of events. In other words, we consider the whole-idea that the text expresses is constituted by several sub-ideas, each of which is a part of that text and describes a certain event. Using the Viterbi algorithm with the HMM, we could find the most appropriate state sequence corresponding to that sequence of events, and thereby determine the most probable emotion of that text.

We implement our method as two phases. The preceding phase is to construct the high-order HMM. At first, the states of the HMM are automatically generated, and then its parameters are estimated. Both of these tasks are based on the same training dataset, in which each pattern is tagged with its right emotion. The following phase is to use the constructed HMM to detect emotion for an input text. The details of these two phases are described in the next sections.

3.1 Constructing the HMM

The first task in this phase is to generate the states of the HMM. It begins by introducing each training pattern text, from an emotion-tagged corpus, to a syntax parser. By analyzing its syntactic structure, the syntax parser identifies grammatical components of the text and outputs the stemmed form as well as syntactic function for each component. Based on this output, we could split the text into grammatically separate parts, each of which is either a sub-idea or a grammatical element. More clearly, each sub-idea corresponds to a clause or a phrase in the text, and grammatical elements correspond to the function words (such as conjunction, adverb...), which are used to link those sub-ideas together. Then, every sub-idea extracted from the corpus is represented by one LSA vector. By clustering those vectors based on their semantic similarity, the HMM states are generated automatically, each of which corresponds to one obtained cluster. There are also a number of special states to model the grammatical elements. After that, the last step is to use the emotion-tagged corpus again to estimate the parameters of the HMM.

Besides the main tasks above, a preliminary task is to determine which sub-ideas are semantically closely related, thus should belong to the same HMM state. For this, we propose to use the VSM enhanced by LSA to reduce its dimension, and acquire its parameters through a large plain text corpus. However, we note that LSA is not mandatory and other techniques could be employed instead for the same purpose.

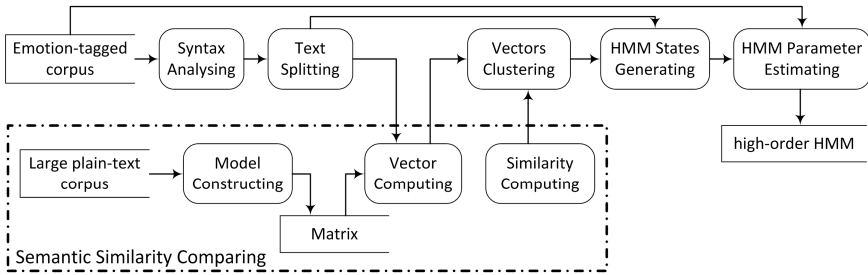


Fig. 1. HMM Constructing Process

Figure 1 describes the whole process of this phase. The rest of this section describes more detail about its essential steps. The first is how to split the text, the following is how to generate states for the HMM, and the last is how to estimate the parameters of that high-order HMM.

Text Splitting. For each training pattern, the first thing to do is to split its text to transform it into a sequence of sub-ideas. Those sub-ideas expressed in the text should be somehow separable, and we assume that they are grammatically separable. Therefore, to split the text, we have to analyze its syntactic structure. In addition, a text might have various kinds of syntactic structures, which might be simple such as a phrase, a simple sentence that has only one clause, or might be more complex such as a complex sentence with more than one clause, a paragraph, or even a document. Generally, we consider that every text contains one or more phrases and clauses, as well as the function words that linked them together. Each phrase or clause corresponds to one sub-idea, which may describe a primary event that directly causes the emotion of the text, or a related event as well as a contextual factor that has an influence on the emotion. Each function word is a conjunction, a preposition, an adverb, ..., which has influence on the meaning expressed by those sub-ideas. We group grammatically similar function words together, and call those groups grammatical elements. Consequently, we have a unified view in which every text is a sequence of sub-ideas and grammatical elements. This unified view allows our method to handle any kind of input text, as long as its syntactic structure is analogous with the training patterns.

We accomplish this step by utilizing a syntax parser to identify the syntactic function for each element of the text. Then, by analyzing this result, we find appropriate splitting points to split the text into a sequence of separate and integral parts, each of which corresponds to one phrase or one clause or one function word.

To have a better result, the text should be stemmed. The stemming could be done simultaneously with the syntax parsing by the same syntax parser, or by another engine separately.

Semantic Similarity Comparing. After having all the text patterns in the training dataset separated into sequences of sub-ideas and grammatical elements, a set of

HMM states is generated so that every sub-idea could be assigned to one appropriate state. Moreover, the sub-ideas that are closely related by semantics should belong to the same state, while semantically different sub-ideas should belong to different states. Hence, at first, it demands a mechanism to compare those sub-ideas semantically. To fulfill this demand, we propose to use the VSM in combination with the LSA.

Every sub-idea is essentially a text, we consider it as a document for convenience. We recall that the VSM uses term-by-document matrix to represent the occurrence of terms in each document. Then each document is represented by a vector, which corresponds to one column of this matrix. Each element of this vector corresponds to a distinct term and represents the weight of that term in that document. The weight must reflect the term frequency in the document, and is usually evaluated with the tf-idf weighting schema, which considers a term important to a document only if it has high frequency in that document while is not so common with respect to the whole set of documents. In practice, the size of the term-by-document matrix is often very large, because there are about tens of thousands of documents and terms to be modeled by the VSM.

The LSA, in fact, is a dimension reduction method for the VSM. This method is based on singular vector decomposition to decompose a large original matrix into a set of a much smaller number of orthogonal factors so that it can be approximated from that set by linear combination. Each document, therefore, can be represented by a vector of about 200 to 300 weights instead. This not only makes computational operations of the VSM less time-consuming, but also helps the VSM less sensible with noise. However, the most important effect of this reduction is that it could take advantage of an implicit higher-order structure in the association of terms with documents to identify semantically related texts accurately even in the case they have no common words [4]. Besides, it could partly solve the problem of polysemy.

The semantic similarity sim of two documents can be evaluated by the cosine of the angle between the two representing vectors of them in the dimension-reduced space. It can be used to obtain the distance d between two documents, which is required by a clustering algorithm. The value of d must be opposite to the value of sim and is not negative. Therefore, since the value of sim is in the range $[-1, 1]$, the value for d is computed by formula (1), so that the range of d is $[0, 1]$.

$$d = \begin{cases} 1 - sim, & sim \geq 0 \\ 1, & sim \leq 0 \end{cases} \quad (1)$$

Generation of HMM States. After having a mechanism to compare semantic similarity between two sub-ideas, the HMM states are generated by clustering the sub-ideas acquired from the training dataset based on their semantic similarity. We consider two clustering algorithms to use in our method. The first is the DBScan algorithm, by which the clustering is based on the density of these sub-ideas [6]. The advantage of this algorithm is that it does not require preselecting the cluster number, and the average distance from each vector to the center of its cluster is much smaller. Nevertheless, this algorithm also has a drawback that in practice there is often a large

number of isolated sub-ideas, thus using it would result in too many clusters. This is a serious problem due to the explosion of required memory space and calculation time for the high-order HMM, which are exponents of the cluster number. Therefore, we propose to use the k-Means algorithm to have a reasonably small number of clusters.

For each obtained cluster, we have a corresponding state generated for it in the high-order HMM. The representing vector for that state is chosen as the mean of all vectors representing the sub-ideas in that cluster. Hence, the emission probability of a symbol by that state can be calculated based on the similarity measure between the representing vectors of that symbol and that state. Beside them, each grammatical element also has a special state generated for its own. For example, each conjunction “and”, “or”, “not” belongs to a different special state, while “but”, “yet”, “however” correspond to the same special state.

Estimation of HMM Parameters. The Hidden Markov Model is a statistical model to model a system that, beside the generated set of states, consists of a set of symbols, such that a directly observable sequence of symbols is emitted by a hidden sequence of those states. The transitions between states and the emission of a symbol by a certain state obey a certain probability distribution. The first-order HMM is the simplest form of HMM, in which the probability that the system transition to a next state depends only on the current state, and the emission probability of a symbol at each step depends only on the current state in that step. In a more complex form, n -order HMM with $n \geq 1$, the transition probability of a state, as well as its emission probability is dependent on the $n-1$ preceding states of the current state.

In our method, each symbol is a sub-idea, which is a clause or a phrase. In other words, a symbol is a combination of arbitrary number of terms. Thus, the symbol set is infinite, so that we cannot define it directly, as well as the symbol emission probability distribution. Instead, we assume that a state could emit any symbol with a certain probability, and that probability is derived from the semantic similarity that is the cosine of the angle between its representing vector and the mean vector representing that state.

According to the emotion invocation process mentioned above, the target state of each state transition is determined based on the sequence of prior states. Hence, we have to choose a high-order HMM to handle this dependency. Moreover, with the special HMM states corresponding to grammatical elements, our HMM could also handle the grammatical relations in the text, such as negative relation, conditional relation, causal relation, ... in emotion detection.

We implement the high-order HMM by using a modification of a method that transforms a high-order HMM into an equivalent 1st-order HMM [21], so that the algorithms for the 1st-order HMM can be used directly. We have to modify the transforming method because, in a proper high-order HMM, the symbol emission probability depends on a number of previous states while, in the high-order HMM we used, it depends only on the current state. For one reason, it is more intuitive and straightforward to compute this probability based on only one current state. Another reason is that, in our opinion, this helps the HMM parameters estimation more tolerant with a small dataset.

To estimate the parameters for the HMM, at first, for each emotion we add a special state tagged with its name, which can emit only symbol “EMOTION”. Then we add to the state sequence of each training pattern the special state corresponding to its emotion. After that, the parameters of the model, which are state transition probability and symbol emission probability, are computed simply by counting from those state sequences.

3.2 Using the Constructed HMM for Emotion Detection

In the second phase of our method, we use the high-order HMM constructed in the first phase for emotion detection. The process diagram of this phase is shown in Figure 2. There are only two new processes, one generates a corresponding symbol sequence for an input text, and the other detects the emotion. All other processes are the same as the corresponding ones in the first phase.

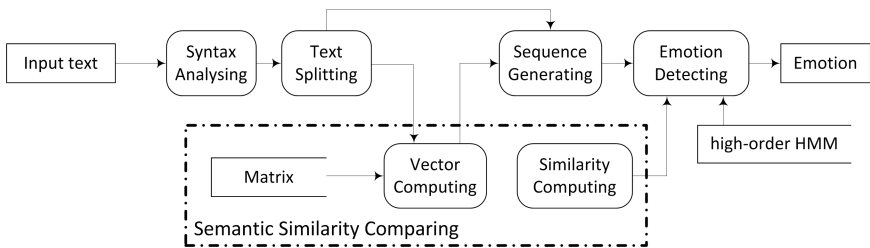


Fig. 2. Emotion Detection Process

For every input text, first it is transformed into a sequence of symbols by the same way as with training patterns. Then we add the symbol “EMOTION” to the end of that sequence. Finally, using the Viterbi algorithm on the constructed high-order HMM, the most probable state sequence to generate that symbol sequence would be found. Hence, the tag of the last state, which corresponds to the symbol “EMOTION”, is the name of the most probable emotion for that input text.

4 Experiments

4.1 Training the LSA Model

For the experiments, we use an already available implementation of LSA, the GenSim packet [14]. To train the LSA model, we choose the British Academic Written English (BAWE) corpus, which is freely available for academic research purpose and totally comprises about 6.5M words². Furthermore, all stop-words, which are commonly used in most sentences, are filtered out from both training patterns and input

² <http://www.coventry.ac.uk/researchnet/BAWE/Pages/BAWE.aspx>

texts before introduced to the LSA. We use the SMART’s stop list³, which contains 570 words. One more remark is that we choose the number of dimensions of the LSA model as 200, a typically used value.

4.2 ISEAR Dataset

For constructing and training the high-order HMM, we use the ISEAR (International Survey on Emotion Antecedents and Reactions) dataset [16]. This dataset consists of 7666 textual pieces tagged with the most appropriate of seven major emotions (joy, fear, anger, sadness, disgust, shame, and guilt) chosen by close to 3000 students. About half of these pieces are complex sentences or have multiple sentences that can be transformed into a sequence of more than one symbol. Thus, the interaction among multiple states that affects the emotion can be considered to be existent.

More importantly, using the ISEAR dataset for evaluation allows us to compare our method with the best methods mentioned in [10], which also uses this dataset for evaluating. Besides, to have an equal comparison, we also take into account only four emotions (anger (includes both anger and disgust), fear, joy and sadness) as [10] does. The number of patterns for each emotion is shown in Table 1.

Table 1. Number of textual pieces for each emotion in ISEAR dataset

Emotion	Anger	Fear	Joy	Sadness	Total
Number	2,192	1,095	1,094	1,096	5,477

4.3 Stanford Syntax Parser

To split the input text, we need a syntax parser to analyze its grammatical structure. In our experimentation, the syntax parser we choose is the Stanford Parser (SP), because it is highly tolerant with grammatical mistakes [2]. The SP provides various kinds of output, but we use only three of them.

The most important one is the Penn Treebank output [11], in which the syntax components of the input text are tagged with its syntactic function and are organized in a hierarchical tree. Traversing that tree top-down, starting from its root, for each node, we find which of its child-nodes is the “S” node (which stands for Simple declarative clause). The traversing is continued recursively with only those child-nodes. After the traversing has stopped, the last traversed node of each sub-branch corresponds to one part of the input text to be split, and the text of that part can be obtained by traversing that node in the deep-first order. The input text is split this way into separate sub-ideas and grammatical elements. More clearly, each clause (“S” node) or phrase (“PP”, “NP”... node) corresponds to one sub-idea, and each function word identified by the SP, which is a conjunction such as “and”, “or”, “but”... (“CC” node), “if”, “although”... (“IN” node), corresponds to one grammatical element.

³ <http://jmlr.csail.mit.edu/papers/volume5/lewis04a/all-smart-stop-list/english.stop>

The next one is the stemming output for both of the BAWE corpus and the ISEAR dataset. With stemming, the accuracy of the LSA model is slightly improved. The last one is the type dependency output of the SP, which describes grammatical relationships between syntax components in a sentence [12]. Utilizing those relationships could help to integrate deeper syntactic information into the emotion detection. For experiments, we utilize the negation modifier relationship to add the NOT state into the state sequence generated from the input text.

4.4 Results

To evaluate our method, we divide the ISEAR dataset randomly into two parts: one consists of 2/3 pieces used for training, and the other consists of the remaining pieces used for testing. Besides, we choose to use the average distance between vectors and the center of its cluster as the threshold, so that a state can emit a symbol only if the similarity measure between their representing vectors is larger than this threshold.

With the 2nd-order HMM, we conduct experiments with various values for the number of the states generated by clustering to compute the average result over 5-fold cross validation for each, which is shown in Table 2. We assume that the high-order HMM must have at least 5 states, one for each of the four emotions and one more for the neutral state, which invokes no emotion. Therefore, we choose the number of states so that it is a multiple of 5, starting with the value of 5 and then increasing gradually. The best result is obtained when the state number is 45.

Figure 3 is the chart of the result shown in Table 2. At first, when the state number is still too small, the precision is low since the HMM could not model enough emotion detection cases. The best result is achieved when the state number is 45. After that, while the precision just increases slightly, the recall falls down, so that the result is not improved. The reason is that for the high-order HMM with too many states, the training dataset is no longer large enough to train it well. For the same reason, the average result obtained over 5-fold cross validation using the 3rd-order HMM is worse than 2nd-order with the same state number of 45.

Table 3 shows the comparison between our method using the 2nd-order and 3rd-order HMMs, which both have 45 states, and other methods using LSA or PLSA [10]. The result shows that our method using the automatically constructed high-order HMMs is better than the method using LSA or PLSA.

Table 2. Average result of 5-fold cross validation with the 2nd-order HMM

N	5	25	45	65	85	105	125	145
Precision	0.398	0.434	0.449	0.465	0.466	0.467	0.476	0.479
Recall	0.278	0.292	0.291	0.284	0.269	0.263	0.265	0.262
F1	0.327	0.349	0.353	0.352	0.341	0.336	0.340	0.339

Table 3. Comparing the results of the high-order HMMs and LSA/PLSA methods

Method	2nd-order HMM	3rd-order HMM	LSA	PLSA
F1	0.353	0.341	0.228	0.270

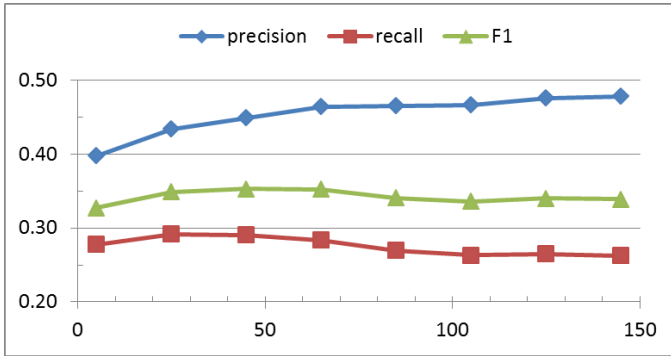


Fig. 3. Average result chart of 5-fold cross validation using 2nd-order HMM

5 Conclusion and Future Work

We have presented our proposed high-order HMM for emotion detection from text. On one hand, it can take into account both the psychological characteristic of emotion that is the process of emotion invoking and linguistic information that are grammatical relations of the input text. On the other hand, the proposed method could detect emotion expressed by “generic” terms, by integrating the VSM with LSA as a semantic similarity comparing mechanism for both constructing HMM states and matching parts of the input text to those states. Evaluation by cross validation on the ISEAR dataset shows a promising result.

This method could be further improved by using a better dimension reduction method such as PLSA, and by taking into account more linguistic information. Besides, we will also improve the HMM or try other probabilistic models to handle higher order dependencies between states and learn more efficiently from a spare training dataset.

References

1. Boucouvalas, A.C.: Real Time Text-to-Emotion Engine for Expressive Internet Communications. In: Riva, G., Davide, F., IJsselsteijn, W.A. (eds.) *Emerging Communication: Studies on New Technologies and Practices in Communication*, pp. 305–318. IOS Press, Amsterdam (2003)
2. Chaumartin, F.R.: UPAR7: A Knowledge-based System for Headline Sentiment Tagging. In: *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 422–425 (2007)
3. Danisman, T., Alpkocak, A.: Feeler: Emotion Classification of Text Using Vector Space Model. In: *AISB 2008 Convention, Communication, Interaction and Social Intelligence*, vol. 2, pp. 53–59 (2008)
4. 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(6), 391–407 (1990)

5. Ekman, P.: An Argument for Basic Emotions. *Cognition & Emotion* 6(3), 169–200 (1992)
6. Ester, M., Kriegel, H.-P., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 226–231 (1996)
7. Esuli, A., Sebastiani, F.: SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In: *Proceedings of the 5th Conference on Language Resources and Evaluation*, pp. 417–422 (2006)
8. Huang, C.-R., Chen, Y., Lee, S.Y.M.: Textual Emotion Processing from Event Analysis. In: *Proceedings of the Joint Conference on Chinese Language Processing* (2010)
9. Kao, E.C.-C., Liu, C.-C., Yang, T.-H., Hsieh, C.-T., Soo, V.-W.: Towards Text-based Emotion Detection: A Survey and Possible Improvements. In: *Proceedings of the 2009 International Conference on Information Management and Engineering*, pp. 70–74 (2009)
10. Kim, S.M., Valitutti, A., Calvo, R.A.: Evaluation of Unsupervised Emotion Models to Textual Affect Recognition. In: *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 62–70 (2010)
11. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a Large Annotated Corpus of English: the Penn Treebank. *Comput. Linguist* 19(2), 313–330 (1993)
12. Marneffe, M.-C., Manning, C.D.: *Stanford Typed Dependencies Manual* (2011)
13. Picard, R.W.: *Affective Computing*. MIT Press (1997)
14. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: *Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks*, pp. 45–50 (2010)
15. Russell, J.A.: Core Affect and the Psychological Construction of Emotion. *Psychological Review* 110(1), 145–172 (2003)
16. Scherer, K.R., Wallbott, H.G.: Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of Personality and Social Psychology* 66, 310–328 (1994)
17. Strapparava, C., Mihalcea, R.: Learning to Identify Emotions in Text. In: *Proceedings of the 2008 ACM Symposium on Applied Computing*, pp. 1556–1560 (2008)
18. Strapparava, C., Valitutti, A.: WordNet-Affect: An Affective Extension of WordNet. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1083–1086 (2004)
19. Wu, C.-H., Chuang, Z.-J., Lin, Y.-C.: Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models. *ACM Transactions on Asian Language Information Processing* 5(2), 165–183 (2006)
20. Yang, C., Lin, K.H.-Y., Chen, H.-H.: Emotion Classification Using Web Blog Corpora. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 275–278 (2007)
21. Ye, F., Yi, N., Wang, Y.: EM Algorithm for Training High-order Hidden Markov Model with Multiple Observation Sequences. *Journal of Information & Computational Science* 8(10), 1761–1777 (2011)

A Lazy Man's Way to Part-of-Speech Tagging

Norshuhani Zamin¹, Alan Oxley², Zainab Abu Bakar³,
and Syed Ahmad Farhan⁴

^{1,2} Faculty of Science and Information Technology

³ Faculty of Computer and Mathematical Sciences,
Universiti Teknologi Mara,
40000 Shah Alam, Selangor, Malaysia

⁴ Faculty of Engineering,
Universiti Teknologi PETRONAS, Bandar Seri Iskandar,
31750 Tronoh, Perak, Malaysia
{norshuhani, alanoxley}@petronas.com.my,
zainabcs@salam.uitm.edu.my,
syfarisk@gmail.com

Abstract. A statistical-based approach to word alignment involving automatically projecting part-of-speech (POS) tags is presented. The approach is referred to as the “lazy man’s way” because it improves POS assignment for a resource-poor language by exploiting its similarity to a resource-rich one. This unsupervised learning method combines the N-gram and Dice Coefficient similarity functions in order to align English texts with Malay texts thus projecting the POS tags from English to Malay. It is a quick method that does not require the laborious effort needed to annotate the Malay dataset. A case study, an experiment done on 25 terrorism news articles written in Malay, has shown that leveraging pre-existing resources from a resource-rich language, i.e. English, to supplement a resource-poor language, i.e. Malay, is feasible and avoids building new text-processing tools from scratch. The system was tested on the Malay corpus, consisting of 5413 word tokens. The results reached values of 86.87% for precision, 72.56% for recall and 79.07% for F1-Score. This shows that the “lazy man’s way”, where a resource-poor language just exploits the rich linguistic information available in English, increases bitext projection accuracy significantly.

1 Introduction

The Malay language is widely used in Malaysia, Brunei, Singapore and Indonesia with around 300 million users approximately [1]. It is a type of Indo-European language and has relatively few resources and has small corpora of texts. (A corpus is a collection of various written or spoken texts in machine-readable forms.) This has been a hurdle for researchers to investigate the language computationally [2]. The characteristics and uniqueness of the Malay language attract linguists to explore the underlying challenges and opportunities. Malay is inflectional language in which the language performs massive affixation, reduplication and composition [3]. Annotated textual data in Malay are currently scarce. Available data are limited and not publicly

available. Examples of private data include the Malay Practical Grammar Corpus [4], the Dewan Bahasa Pustaka (DBP) Database Corpus¹, the Malay Corpus by Unit Terjemahan Melalui Komputer from the University Science of Malaysia [5] and, more recently, the MALay LEXicon (MALEX) [6]. The freely available Malay Concordance Project Corpus² is a collection of 3 000 000 words extracted from classical Malay texts, ones that are not related to this research domain.

Text-processing tools such as lemmatisers, part-of-speech taggers (POS-taggers), analysers, stemmers and parsers are available for only a few resource-rich languages, such as English, German and Japanese. Parallel corpora or bitext are extensively studied in the machine translation field where the aligned phrases and words are used to create translation models [7]. A parallel corpus appears to be a good statistical resource. Hence, this research is conducted to exploit the linguistic information from a resource-rich language for the benefit of the Malay language by using a bitext mapping, thus avoiding building a Malay tagger from scratch. In this study, English is chosen as its taggers are very well established with an accuracy of up to 98% accuracy, and therefore there is almost no room for improvement [8, 20].

Although it is acknowledged that different languages encode types of grammatical information differently, the proposed technique is able to resolve most syntactic disambiguation between English and Malay. A model comprising of the N-grams of two characters and the Dice Coefficient similarity function is used to leverage the pre-existing resource. Reuse of resources helps to reduce costs and overheads in system development. The aim of this paper is to develop an automated POS-tagger for Malay by projecting the linguistic resources from a resource-rich language, i.e. English. The system takes as input 1) Malay terrorism text, 2) its translated English text tagged with an open-source tagger and 3) a manually generated dictionary look-up of Malay-English words relevant to the domain. The output is Malay text with projected POS tags from the annotated English corpus. In comparison with our previous research [9], significant improvements of 11% and 6% have been achieved for precision and recall, respectively.

2 Related Work

Limited research is made available on POS tagging for Malay. Lack of linguistic tools and limited access to computational resources daunt researchers from conducting further investigation on this language. Research on Malay linguistics has been explored thoroughly by Ranaivo-Malaicon in a series of publications [10, 11, 12]. The studies include lexical and morphological analyses, and tagging. The POS tags are inferred from the rule-based morphological analyser. Building a morphological analyser is computationally expensive and laborious. A study on Malay POS tagging to complement MALEX, the annotated Malay lexicon, focusing on the problem of syntactic drift has been conducted [13]. The tagsets are identified from a data-driven approach and the study presented a list of possible syntactic drifts

¹ <http://www.dbp.gov.my>

² <http://mcp.anu.edu.au/>

for the Malay language. Further in this area, the researchers performed a corpus-based approach to the analysis of grammatical class in Malay [14]. An analysis of four DBP novels involving around 120 000 words was conducted over this supervised system. Each of these words was laboriously tagged using the DBP tagsets, consisting of 71 different word classes. Then, the annotated corpus was used to predict unseen words in a totally new novel of the same genre.

An open source corpus with over 26 000 Malay words extracted from modern Malay texts on the World Wide Web was used to develop a Malay sentence tokeniser, lemmatiser and POS-tagger [15]. However, the tags were not purely generated but partially taken from the KAMI Malay-English lexicon of various genres [16] and the work was reported incomplete. Nevertheless, this supervised approach of lemmatisation achieved 94.5% overall accuracy. A closely related work on bitext mapping is reviewed in [17]. A pattern recognition algorithm known as the Smooth Injective Map Recognizer and the Geometric Segment Alignment algorithm are used to align English and Malay texts. Additionally, the prototype required a translation lexicon constructed from a machine-readable English-Malay dictionary and a lemmatiser to lemmatise texts. Tagging and lemmatisation are performed using Brill's tagger [23]. However, no work on POS tagging is involved in this research. More recently, a trigram Hidden Markov Model (HMM) for tagging Malay texts was introduced in [18]. It is a supervised statistical tagger that learns from a tagged bilingual Malay-English dictionary which contains only 576 words [19]. The accuracy of the tagger reached up to 67.9% for an average of 1840 test tokens. The results show that HMM is a promising method to predict tags for Malay words but the overall process to prepare the Malay corpus involves a highly expensive morphological analysis.

3 POS Tagging

POS tagging is the first stage in automated text analysis. The development of language technologies can scarcely begin without this initial phase. A POS tagger is software that reads text in some language and associates a POS tag to each word. POS tags represent syntactic and morphological categories. It is a significant step in semantic disambiguation. The term "POS tag" is often used interchangeably with the terms "category," "word class," and "lexical category" in linguistic publications.

There are three ways to conduct POS tagging - supervised, semi-supervised and unsupervised [20]. Supervised learning requires a collection of sample data to be learnt and the pattern found is used to determine new instances. Semi-supervised learning takes very little data to initiate the learning process. This approach is suitable when the nature of the domain possesses a limited dataset. Unsupervised learning has no target attribute and this leaves a challenge to the algorithm to explore the data to find intrinsic structures in it. A comprehensive review finds that supervised and semi-supervised POS taggers generate better results than unsupervised methods. This is evidenced by the Stanford POS tagger which was trained on the Wall Street Journal corpus. It records up to 97.24% accuracy [21]. A range between 96% and 97% accuracy is achieved for most Indo-European languages such as English, French, Dutch and German [20]. There is

also an error-driven transformation-based tagger for English, known as Brill's tagger, which automatically learns and induces tagging rules from a pre-tagged English corpus [23, 24]. It is the first widely used tagger to have an accuracy of above 95%. A mirror of Brill's tagger is the latest version and is known as the CST tagger³. Unsupervised learning methods are only claimed to be an alternative solution to older supervised and semi-supervised algorithms. Christodouloupoulos et al. [25] uncovered further evidence for this analysis. The study showed that the best accuracy i.e. the F1 score recorded for unsupervised POS taggers, for less-studied languages, was only 76.1%.

It is interesting to highlight that this research focuses on a statistical unsupervised learning approach, an approach which is new to Malay. The unsupervised learning method is opted for due to: 1) the relative successes of such a method in other foreign languages, 2) the lack of annotated data to build a supervised classifier and 3) the desire to come up with a quick-turnaround solution. A bilingual or parallel corpus is proposed as a possible solution to the issue of the non-existence of a Malay corpus. Also, as a Malay terrorism corpus is not available, there is no training data. This proposed research is hoped to bridge the gap in performance between supervised learning POS tagging and unsupervised learning POS tagging.

Ambiguity is the challenge to POS tagging. In this context, ambiguity refers to a word with multiple POS tags. However, this problem occurs more frequently in English than Malay. For example, the common tag for the word *bank* is a noun as in the *river bank*. However, in the phrase *bank the money*, the word *bank* is tagged as a verb. The phrase *we can can the can* illustrates a more complicated example where the three occurrences of the word *can* correspond to the auxiliary, verb and noun categories respectively. Using contextual rules is one way of solving this problem.

4 Proposed Approach

This paper presents an unsupervised POS tagging approach to tag Malay terrorism texts using a bilingual Malay-English corpus. An N-gram scoring method for two characters is integrated with the Dice Coefficient function [26] in order to calculate the probability distribution of letter sequences between Malay and English texts. (An N-gram for $N = 2$ is commonly referred to as a "bigram." A bigram is a sequence of graphemes, the smallest semantically distinguishing unit in a written language, i.e. alphabetical letters) N-grams are a measure of assessing the similarity of two strings [28]. They perform well on languages of different structures and are widely implemented in much text-mining research. The use of the Dice Coefficient function in bitext alignment research is referred to in [27]. It is a simple statistical method to measure the string closeness of two different texts and gives good results.

The use of bigrams is proposed as a significant improvement to several limitations observed in the research: 1) the morpheme similarity measure requires a morphological analyser and 2) the effort to pre-align the bilingual corpus is an expensive option. An N-gram based model is faster and smaller at matching two languages with dissimilar

³ http://cst.dk/online/pos_tagger/uk/

patterns [29] and performs better than the sequence of morphemes approach as experimented in [30]. A dictionary look-up of different lexemes is introduced in the framework. (In linguistics, a lexeme refers to a variation of a word. It is the minimal unit of language that has a semantic interpretation. For example, the words “cry,” “cries,” “cried” and “crying” are lexemes for the lemma “cry”. A lemma is the basic morphological unit of a language, which is often referred to as the “base word.”) The dictionary developed in this research is a lexical database that is organized around lexemes, which are listed in the dictionary as separate entries. Bigram scoring is used to pick the English lexeme with the highest similarity score to a given Malay word. This is a simple and faster solution to building a morphological analyser for Malay. Developing a Malay morphological analyser is not easy since the language is agglutinative and inflectional and performs massive affixation, reduplication and composition. The proposed framework is shown in Fig.1.

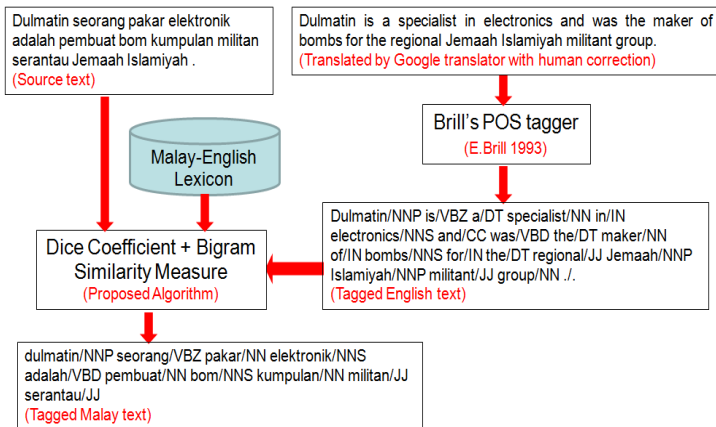


Fig. 1. Unsupervised Malay tagger framework [9]

The output of the framework is automatically tagged Malay text. This is done by projecting the associated English POS tag to the matched Malay word. A fully unsupervised Malay tagger is introduced in this paper.

5 Word Alignment Algorithm

Consider as input a Malay sentence: “Peristiwa keganasan sukar dipercayai itu yang berlaku di Indonesia adalah tragik.” The text “The unbelievable/JJ terrorism/NN events/NNS that/WDT happened/VBD in/IN Indonesia/NNP were/VBD tragic/JJ” is the translated and tagged version of the sentence (Note: DT = Determiner; JJ = Adjective; NN = Singular Noun; NNS = Plural Noun; WDT = Wh-Determiner; VBD = Past Tense Verb; IN = Preposition, Conjunction; NNP = Singular Proper Noun). The Malay sentence is tokenized and the closest translation for each word is calculated and selected from the lexicon. Let us illustrate this process using the Malay phrase “sukar dipercayai.” This is an example of many-to-one text mapping as

opposed to the previous example of one-to-one mapping [9]. D_{ME} is a Malay-English dictionary of terrorism related words and phrases. Lexemes for “percaya” and its related phrases are stored with their translation. This information serves as the thesaurus of the word, as in Fig. 2.

percaya = believe, trust	boleh dipercayai = reliable, trustworthy, believable
dipercayai = believed, trusted	sukar dipercayai = unbelievable, unreliable
mempercayai = trust, beguiling, credence	kepercayaan = reliance, trust, credo, credibility

Fig. 2. Sample lexicon

Lexemes are used in the lexicon to avoid building a Malay morphological analyser for lemmatisation. This is one of highly computational task in NLP. Using the example discussed earlier, the two lists of words are defined as follows:

$$\begin{aligned}
 W_M &= \{\text{peristiwa, keganasan, sukar dipercayai, itu, yang, berlaku, di, Indonesia, adalah, tragik}\} \\
 W_E &= \{\text{the/DT, unbelievable/JJ, terrorism/NN, events/NNS, that/WDT, happened/VBD, in/IN, Indonesia/NNP, were/VBD tragic/JJ}\} \\
 &= \{E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9, E_{10}\}
 \end{aligned}$$

The algorithm searches the best matching word in W_E for each word in W_M using the lexicon. The lexicon is referred to as D_{ME} in this working example. As for the Malay phrase “sukar dipercayai,” the D_{ME} list is extracted as follows:

$$D_{ME}(\text{sukar dipercayai}) = \{\text{unbelievable, unreliable}\} = \{d_1, d_2\}$$

The morpheme similarity of each word d_i with each of the English words in W_E is calculated using a combination of bigram score and the Dice Coefficient function as formalised below:

$$Sim(d_i, E_j) = \frac{2 \times N_{d_i \cap E_j}}{N_{d_i} + N_{E_j}} \quad (1)$$

where $N_{d_i \cap E_j}$ is the number of bigrams common to both d_i and E_j , N_{d_i} is the number of bigrams found in d_i , and N_{E_j} is the number of bigrams found in E_j .

Given the phrase ‘sukar dipercayai’, the translations found in the lexicon, D_{ME} are *unbelievable* and *unreliable*. Let d_1 represents *unbelievable* and d_2 represents *unreliable* for the D_{ME} (sukar dipercayai). The algorithm compares each of the lexicon word in the D_{ME} , in this case is the d_1 and d_2 with the words $E_1, E_2, E_3, E_4, E_5, E_7, E_8, E_9$ and E_{10} . In the first iteration, the calculation of the values $Sim(d_1, E_1)$ and $Sim(d_2, E_1)$ are performed to find the closest match. These are referring to the $Sim(\text{unbelievable, the})$ and $Sim(\text{unreliable, the})$ respectively. Next, the bigrams of each d and E are generated. This process produces the bigram strings of *un,nb,be,el,li,ie,ev,va,ab,bl,le*, *un,nr,re,el,li,ia,ab,bl,le* and *th,he*. The number of

bigrams for each word is calculated. In this iteration, d_1, d_2 and E_1 gained 11, 9 and 2 bigrams respectively. Then, the algorithm performs pair-wise matching of the list of bigrams as shown in the Fig. 3:

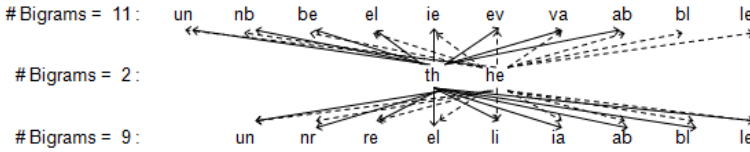


Fig. 3. Bigrams pair-wise matching of the English words / phrases

Using (1), the similarity score for the word *the* given *unavailable* as in $Sim(unavailable, the)$ is 0. Similarly, very poor match of $Sim(unreliable, the)$. The second iteration continues with the word $E_2(unbelievable)$ and this algorithm iterates until $E_{10}(tragic)$. However, in this example, the second iteration returns the highest correlation score. The word *unbelievable* appears to be the most plausible translation of the Malay phrase ‘sukar dipercayai’ as the calculation returned the highest probability score i.e. 1. After all words in W_M and W_E are statistically compared, the bitext are mapped as depicted in Fig.4.

peristiwa	keganasan	sukar dipercayai	itu	yang	berlaku	di	Indonesia	adalah	tragik
the	unbelievable	terrorism	events	that	happened	in	Indonesia	were	tragic
DT	JJ	NN	NNS	WDT	VBD	IN	NNP	VBD	JJ

Fig. 4. Malay-English bitext mapping

Finally, the POS tag of each English word is projected to its corresponding Malay word so as to create the annotated Malay terrorism corpus: *The/DT unbelievable/JJ terrorism/NN events/NNS that/WDT happened/VBD in/IN Indonesia/NNP were/VBD tragic/JJ*. In this work, all delimiters such as commas, exclamation marks, question marks and other symbols are ignored.

6 Experimental Results

Initially, the dataset for the experiment was only available in hardcopy form. Among these were 25 news articles on Indonesian terrorism. These articles were digitized and formatted accordingly. The experimental setup involved the following processes:

- i. The 25 Malay terrorism news articles were translated into English using Google Translate⁴. The accuracy of the translation was checked by a human expert an amended where necessary.

⁴ <http://translate.google.com/#>

- ii. The translated corpus was then tagged using a free version of Brill's transformation-based learning tagger known as the CST tagger (referred to as the annotated English corpus). Brill's tagger is a rule-based tagger with 48 different tagsets and expressive disambiguation rules.
- iii. A Malay-English terrorism lexicon was hand-crafted using the online 'Dictionary Malay & English'⁵ and this serves as the vocabulary of the research. (Within this paper we often refer to it as the "list of words" or the "dictionary look-up.") Although this is the most laborious step, it requires significantly less effort than building a morphological analyser.

An English - Malay bilingual corpus of 25 news articles on Indonesian terrorism, with 263 sentence pairs and 5413 word tokens, was used to evaluate the framework. The Malay texts were manually tagged by a language expert using our own tagsets, which were made equivalent to Brill's tagsets. Each of our tagsets is limited to contain up to six Brill tagsets, as shown in Table 1. In this research, all variants of verbs, nouns, pronouns, cardinal numbers and adjectives produced by the CST Tagger are generalized as VB, CN, PN, CD and ADJ respectively. These entities are significant for the later development. Regular words were removed in the Malay corpus leaving only 3466 prominent word tokens. Commercially available Malay taggers are nonexistent. Comparison of the proposed system was, therefore, made against the results of human experts. Several manual tagging rules of thumb were applied to reduce the variation between the human and automated systems.

Table 1. Tagsets

Our Tagsets	Equivalent Brill's Tagsets
Common Noun (CN)	1. Noun, Singular or Mass (NN) 2. Noun, Plural (NNS)
Proper Noun (PN)	1. Proper Noun, Singular (NNP) 2. Proper Noun, Plural (NNPS)
Pronoun (PRN)	1. Personal Pronoun (PRP) 2. Possessive Pronoun (PRP\$)
Verb (VB)	1. Verb in base form (VB) 2. Verb in past tense (VBD) 3. Verb in gerund / present participle (VBG) 4. Verb in past participle (VBN) 5. Verb in non-3rd person singular present (VBP) 6. Verb in 3rd person singular present (VBZ)
Cardinal Number (CD)	1. Cardinal Number (CD)
Adjective (ADJ)	1. Adjective (JJ)

A set of 25 Malay news articles focusing on Indonesian terrorism was tagged by a native Malay speaker using the above tagsets, i.e. CN, PN, PRN, VB, CD and ADJ. The proposed tagsets are used to reduce variation in comparison to using Brill's tagsets, where multiple tagsets are applied to similar categories of entities. Each

⁵ <http://kamus.lamanmini.com/index.php>

Malay word was tagged by the system and by the human. A word tagged by the system in the same way as the human was regarded as a correct system tagging. Wrong tags were recorded when a mismatch was found between the system’s tagging and the human’s tagging, whilst words untagged by the system were considered to be have been missed by the system. The overall results are shown in Table 2.

Table 2. Results Analysis

# Sentences	# Words	# Entities	# Correct	# Wrong	# Missed
263	5413	3466	2515	380	571

Precision and Recall are the evaluation metrics best suited to measuring the performance of text processing systems. Precision is the probability that a projected tag is relevant, while recall is the probability that a tag is relevant. In [31], the precision and recall metrics are defined for word alignment as follows:

$$precision = \frac{|A \cap A_r|}{|A|} \quad \text{and} \quad recall = \frac{|A \cap A_r|}{|A_r|} \tag{2}$$

where A = set of alignments produced by the system and A_r = set of alignments produced by the human. Alternatively, in relation to the data in **Table 2**, the above formulas can be interpreted as follows:

$$Precision = \# \text{ Correct} / (\# \text{ Correct} + \# \text{ Wrong})$$

$$Recall = \# \text{ Correct} / (\# \text{ Correct} + \# \text{ Wrong} + \# \text{ Missed})$$

However, accuracy of most text processing systems is evaluated using the F1-Score. F1-Score is the weighted harmonic means of precision and recall. It is calculated as follows:

$$F1\text{-Score} = (2 \times Precision \times Recall) / (Precision + Recall)$$

In this research, Precision, Recall and F1-Score were calculated for the entire texts in the Malay corpus. The numbers of correct, wrong and missed tags were counted for all words as exemplified in the Fig. 5.

Word	1	2	3	4	...	n
Human	NN	VBG	PRN	JJ		NNS
System	CN	VB	CN			CN
Result	C	C	W	M		C

Fig. 5. Example of test data
(Legend: C = Correct, W = Wrong and M = Missed)

The results demonstrate that the system achieved 86.87% precision, 72.56% recall and 79.07%. These results indicate that the tagger attains slightly better annotation accuracy than in the previous experiment [9]. The reason for this could be that the size of the corpus is 12 times larger than the previous size. Observations have found that the size of the dictionary look-up increases proportionately with the size of the tested

corpus. Consequently, the probability of an English word getting correctly aligned with a Malay word is increased. A total of 1444 pairs of Malay-English words in the dictionary look-up have been collected from this experiment.

As far as we are aware, no previous work has studied the alignment of a Malay corpus with an English corpus by projecting tags using statistical approaches. Thus, the technique is considered promising and novel. The results indicate that this implementation works well when the sentence pair is good, i.e. when there is no data scarcity problem. In linguistics, data scarcity refers to a condition where the data needed for a language model is inadequate. However, some interesting problems have arisen from this experiment, which we now discuss. There were 571 missing words, which is about 16% of the total number of entities. In this experiment, "missed" means skipped or not processed by the system. It was found by observation that a word was missed due to either it not being in the dictionary / lexicon or it being associated with a many-to-one entity. Clearly, system performance can be increased further by adding those missing words to the dictionary / lexicon. On the other hand, ambiguous tagging results returned by Brill's tagger contributed to reducing system performance. Fig. 6 shows an erroneous tagging produced by Brill's tagger. Words given the wrong tag are underlined.

Dulmatin/NNP ./, 39/NNP ./, is/VBZ among/IN three/CD <u>suspected/VBN</u> militants/NNS who/WP were/VBD <u>shot/NN</u> dead/JJ

Fig. 6. Brill's tagging errors

It is interesting to highlight that even though the type of data used, i.e. Indonesian terrorism text, is greatly different to that in Brill's test corpus [23], the system has performed reasonably well, even with very small samples. The most significant difference between the results lies in the type of the language structure. Nevertheless, the proposed statistical technique has shown that bitext mapping between two languages with different structures and grammars is permissible. The integration of grammar rules is expected to increase the performance significantly.

7 Conclusions

Firstly, it has to be pointed out that the results described in section 5 are based on a single test corpus of Indonesian terrorism text consisting of 5413 running words. For a better evaluation on the ability of the system to annotate Malay texts, further tests on a larger terrorism corpus from different countries would be necessary. Secondly, the bigram scoring method combined with the Dice Coefficient similarity function and a dictionary look-up shows promising results for the auto-tagging of Malay with Finally, the bitext alignment method appears to be a powerful unsupervised learning technique mapping two dissimilar languages. Unsupervised techniques heavily reduce the labour required to annotate a Malay corpus, and generate quick results. It is hoped that this idea will encourage more linguistic research on resource-poor languages. Moreover, this work shows that the tagger performed reasonably well (in terms of

accuracy and speed) in the Malay terrorism domain, equivalent to most existing taggers working in less restrictive domains.

There are numerous ways to improve the methods presented here, such as using rules to tie in the many-to-one entities and applying semantic disambiguation rules for Malay, as successfully demonstrated in [22]. Named Entity Recognition of the annotated Malay terrorism corpus is an extension of this research and is currently a work in progress.

References

1. El-Imam, Y.A., Don, Z.M.: Rules and Algorithms for Phonetic Transcription of Standard Malay. *IEICE - Trans. Inf. Syst.* E88-D, 2354–2372 (2005)
2. Hassan, A.: *The Morphology of Malay*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia (1974)
3. Tan, Y.L.: A Minimally-Supervised Malay Affix Learner. In: *Proceedings of the Class of 2003 Senior Conference, Computer Science Department, Swarthmore College* (2003)
4. Abdullah, I.H., Ahmad, Z., Ghani, R.A., Jalaludin, N.H., Aman, I.: *A Practical Grammar of Malay – A Corpus based Approach to the Description of Malay: Extending the Possibilities for Endless and Lifelong Language Learning*. National University of Singapore (2004)
5. Ranaivo, B.: *Methodology for Compiling and Preparing Malay Corpus*. Technical Report. Unit Terjemahan Melalui Komputer. Pusat Pengajian Sains Komputer, Universiti Sains Malaysia (2004)
6. Don, Z.M.: Processing Natural Malay Texts: A Data Driven Approach. *TRAMES* 14(1), 90–103 (2010)
7. Jody, F.: An Overview of Bitext Alignment Algorithm, <http://www.ida.liu.se/~jodfo/gslt/bitext-alignment-jody.pdf> (accessed on March 2012)
8. Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J.: Developing a Robust Part-of-Speech Tagger for Biomedical Text. In: Bozanis, P., Houstis, E.N. (eds.) *PCI 2005. LNCS*, vol. 3746, pp. 382–392. Springer, Heidelberg (2005), doi:10.1007/11573036_36
9. Zamin, N., Oxley, A., Bakar, Z.A., Farhan, S.A.: A Statistical Dictionary-based Word Alignment Algorithm: An Unsupervised Approach. In: *Proceedings of International Conference on Computer and Information Sciences* (2012) (manuscript to be published)
10. Ranaivo-Malanco, B.: Malay Lexical Analysis Through Corpus-based Approach. In: *Proceedings of International Conference of Malay Lexicology and Lexicography (PALMA)*, Kuala Lumpur, Malaysia (2005)
11. Ranaivo-Malancon, B.: Approach for a Malay Morphosyntactic Tagging. In: *Proceedings of the Traitement Automatique des Langues Naturelles*, Dourdan, France (2005)
12. Ranaivo-Malancon, B.: Computational Analysis of Affixed Words in Malay Language. In: *Proceedings of the 8th International Symposium on Malay/Indonesian Linguistics*, Penang, Malaysia (2004)
13. Knowles, G., Don, Z.M.: Tagging a Corpus of Malay Text and Coping with Syntactic Drift. In: *Proceedings of the Corpus Linguistics*. Centre for Computer Corpus Research on Language, pp. 422–428. University of Lancaster (2003)

14. Knowles, G., Don, Z.M.: *World Class in Malay: A Corpus-based Approach*. Dewan Bahasa dan Pustaka (2006)
15. Baldwin, T., Awab, S.: *Open Source Corpus Analysis Tools for Malay*. In: *Proceedings of the International Conference of Language Resources and Evaluation*, Genoa, Italy (2005)
16. Quah, C.K., Bond, F., Yamazaki, T.: *Design and Construction of a Machine-Tractable Malay-English Lexicon*. In: *Proceedings of Asian Association of Lexicography*, Seoul, Korea (2001)
17. Al-Adhaileh, Mosleh, H., Tang, E.K., Melamed, I.: *Malay-English Bitext Mapping and Alignment Using SIMR/GSA Algorithms*. Working Paper, Universiti Sains Malaysia (2009)
18. Mohamed, H., Omar, N., Aziz, A.J.A.: *Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Model Approach*. In: *Proceedings of the International Conference on Semantic Technology and Information Retrieval*, Putrajaya, Malaysia (2011)
19. Hock, O.Y.: *Kamus Dwibahasa Edisi Kedua*. Pearson Longman, Malaysia (2009)
20. Indurkha, N., Damerou, F.J.: *Handbook of Natural Language Processing*, 2nd edn. Chapman & Hall / CRC Press (2010)
21. Toutonova, R., Klein, D., Manning, C.D., Singer, Y.: *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network*. In: *Proceedings Human Language Technology Conference* (2003)
22. Jusoh, S., Fawareh, H.M.A.: *Resolving Ambiguous Semantic in Malay Texts*. In: *Proceedings of International CODATA Conference*, pp. 350–356 (2009)
23. Brill, E.: *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging*. *Journal of Computational Linguistics* (1995)
24. Brill, E.: *A Simple Rule-Based Part of Speech Tagger*. In: *Proceedings of the 3rd Conference on Applied Natural Language Processing* (1992)
25. Christodoulopoulos, C., Goldwater, S., Steedman, M.: *Two Decades of Unsupervised POS Induction: How Far Have We Come*. In: *Proceedings of Empirical Methods in Natural Language Processing* (2010)
26. Dice, L.R.: *Measures of the Amount of Ecologic Association between Species*. *Journal of Ecology* 26, 297–302 (1945)
27. Dien, D.: *Building an English-Vietnamese Bilingual Corpus*. Master Thesis in Comparative Linguistics, University of Social Sciences and Humanity of HCM City, Vietnam (2001)
28. Kondrak, G.: *N-gram Similarity and Distance*. In: *Proceedings of the International Conference on String Processing and Information*, Buenos Aires, Argentina (2005)
29. Dunning, T.: *Statistical Identification of Language*. New Mexico State University, Technical Report MCCA, pp 94-273 (1994)
30. Florian, R., Ngai, G.: *Fast Transformation-based Learning Toolkit*. Technical Report (2001)
31. Ahrenberg, M., Hein, A.S., Tiedemann, J.: *Evaluation of Word Alignment Systems*. In: *Proceedings of International Conference on Linguistic Resources* (2000)

Knowledge Acquisition for Categorization of Legal Case Reports

Filippo Galgani, Paul Compton, and Achim Hoffmann

School of Computer Science and Engineering
The University of New South Wales, Sydney, Australia
{galganif,compton,achim}@cse.unsw.edu.au

Abstract. Natural language processing in complex domains, such as law, requires elaborate features, and their interaction is often difficult to model: thus traditional machine learning approaches might fail to perform satisfactorily. This paper describes our approach to assign categories and generate catchphrases for legal case reports. We describe our knowledge acquisition framework which lets us quickly build classification rules, using a small number of features, to assign general labels to cases. We show how the resulting knowledge base outperforms machine learning models which use both the designed features or a traditional bag of word representation. We also describe how to extend this approach to extract from the full text a list of more specific catchphrases that describe the content of the case.

1 Introduction

In the legal field the problem of information overload is particularly pronounced, with a very large amount of legal material stored in textual form. Due to the importance of precedence in common law, legal research generally is based on searching through the case law of applicable jurisdictions looking for facts that are similar to the facts of the current case. Given the large number of court decisions to be scrutinized, information searching can become a very onerous task for legal professionals [16]. Thus natural language processing applications are potentially very useful in the legal domain. Among the possible applications of language analysis techniques to law, this paper examines the task of generating catchphrases for legal case reports. Case reports often contain a list of catchphrases: phrases that present the important legal points of the case. Catchphrases can be seen as a summary of the case, with an indicative rather than informative function: they present all the legal points considered instead of just summarising the key point(s) of a decision. Automatic generation of catchphrases (both for old documents which do not have any catchphrase as well as to automate the creation of catchphrases for new documents) would improve the performance of retrieval systems and aid research of case law to practising lawyers.

Examples of legal catchphrases for two cases are given in Table 1. It can be seen that there are different types of catchphrases, some are more general

Table 1. Examples of catchphrases list for two cases

COSTS - proper approach to admiralty and commercial litigation - goods transported under bill of lading incorporating Himalaya clause - shipper and consignee sued ship owner and stevedore for damage to cargo - stevedore successful in obtaining consent orders on motion dismissing proceedings against it based on Himalaya clause - stevedore not furnishing critical evidence or information until after motion filed - whether stevedore should have its costs - importance of parties cooperating to identify the real issues in dispute - duty to resolve uncontentious issues at an early stage of litigation - stevedore awarded 75% of its costs of the proceedings
MIGRATION - partner visa - appellant sought to prove domestic violence by the provision of statutory declarations made under State legislation - "statutory declaration" defined by the Migration Regulations 1994 (Cth) to mean a declaration "under" the Statutory Declarations Act 1959 (Cth) in Div 1.5 - contrary intention in reg 1.21 as to the inclusion of State declarations under s 27 of the Acts Interpretation Act - statutory declaration made under State legislation is not a statutory declaration "under" the Commonwealth Act - appeal dismissed

and specify the broad category of the case, such as "*Costs*" and "*Migration*" in the example. Others are much more specific in presenting particular issues examined in the case; for example, "*stevedore not furnishing critical evidence or information until after motion filed*". We propose to tackle different kinds of catchphrases with different methods; in particular we address the more general "high level" catchphrases as a categorization task, where we assign the category of a case by choosing from a list of possible labels. In a separate step we deal with the "second level" of catchphrases. As these seem very specific for each case, and thus re-use is very limited, we propose to identify and extract important fragments from the full text of the case as candidate catchphrases.

Text categorization is often tackled with machine learning approaches which try to automatically learn the weights and interactions between features. However machine learning requires large amounts of training data and the result may still be inferior to manually engineered models; see for example, de Maat et al. [15] for a comparison of machine learning versus knowledge engineering in classification of legal sentences. In this paper we describe our approach based on knowledge acquisition to generate catchphrases for legal text, and show how it outperforms machine learning techniques. In particular, we designed a range of novel features and built a knowledge base composed of rules that assign a specific label (such as *Costs*, *Migration*, etc.) to case reports.

The paper is organized as follows: after describing the corpus in Section 3, we present two simple automatic methods for categorizing cases in Section 4. The main contribution of the paper is the Knowledge Acquisition (KA) framework described in Section 5. We then show that this KA approach performs better than machine learning, trained both with a traditional bag-of-words representation or with the features we designed for this task (Section 6). Finally in Section 7 we describe how to extend this approach for the more challenging problem of extracting specific catchphrases from the full text of the case, in order to cover

lower-level catchphrases such as “*stedore not furnishing critical evidence or information until after motion filed*”. For this more challenging task, we are developing a KA approach which uses an additional number of features, designed to recognize important portions of text, and we have devised an automatic evaluation method to estimate the performance of the knowledge base and thus guide the acquisition of rules.

2 Related Work

Different kinds of language processing have been applied to the legal domain, for example automatic summarization [11, 13], machine translation [4] and citation analysis [8, 20]. Among these tasks, the most relevant to our general catchphrases generation task is extractive summarization, where important fragments of text are identified and extracted; see for example [7].

The specific task of assigning the first level of catchphrases; however, appears more related to text categorization, as in this case we are in fact choosing a general label for the document rather than looking for important content inside it. A variety of methods have been researched for document categorization; see for example [17] for a general overview. In particular, examples of categorization approaches to legal cases include the work of Thompson [18], Goncalves and Quaresma [9] (who showed that some linguistic processing, including stemming and part-of-speech tagging, helped assign general topic categories to Portuguese legal cases) and more recently the work of Ashley and Brüninghaus [1], who attempted to automatically classify textual descriptions of case facts according to a scheme of classification based on fact-specific concepts called Factors.

Regarding the comparison of incremental Knowledge Acquisition (KA) with Machine Learning (ML), some recent works have analysed the advantages of the former in different natural language processing tasks, including email classification [13], open information extraction for the web [12], legal citation classification [8] and POS tagging [19]. These works make use of the Ripple Down Rule methodology (RDR) [2], an error-driven, example based, KA approach in which the incremental refinement of the knowledge base is achieved by patching the errors it makes, using exception rules. While our knowledge base is not organized in a tree structure and does not have exception rules, we take inspiration from RDR, in that the proposed knowledge acquisition is incremental and rules are added to cover cases not yet classified.

3 Corpus of Legal Catchphrases

We use as a source of data the legal database AustLII¹, the Australasian Legal Information Institute [10], one of the largest sources of legal material on the net, which provides free access to reports on court decisions in all major courts in Australia.

¹ <http://www.austlii.edu.au>

We created an initial corpus of 2816 cases accessing case reports from the Federal Court of Australia (FCA), for the years 2007 to 2009, for which author-made catchphrases are provided, and we extracted the full text and the catchphrases of every document. Among the list of catchphrases, we distinguished the first level of catchphrases from the others (an example was given in Table 1).

The first level catchphrases of each case can be recognized as they are listed first, and they are usually in upper case; we call these the **label(s)** of the document. From the 2816 cases, we extracted a total of 3504 labels (2296 cases have one label, 396 have two, 124 have three or more). In total we found 254 different labels in the corpus. Of these, 132 labels occurs only once, 33 twice, the remaining 89 more than twice. The 20 most frequent labels are given in Table 2. Each case has also on average 6.8 lower-level, more specific, catchphrases (10.4 words long on average). In total we collected 19251 of these catchphrases, of which 15303 (79.5%) were unique, appearing in only one document in the corpus.

We also downloaded FCA cases from the year 2006, to be used later as a test set. This test set contains 1073 cases, with a total of 1254 labels (929 cases have one label, 112 two labels, the rest three or more). In total there are 149 different labels in the test set, of which 48 do not occur in the training set at all. These cases were not used to develop our knowledge base, but were used for the final evaluation, described in Section 6.

To add more information to the cases, we also extracted citation data from LawCite, a service provided by AustLII. For each case in our corpus, we downloaded both the catchphrases (where available) and the full text (from AustLII), of both cited (previous) cases and more recent cases that cite the case being considered (citing cases). For citing cases we also searched the full text to extract the location where the citation is explicitly made, and extracted the containing paragraph(s). We then accessed the full text of any law or specific section of the law cited by the judge, and extracted the title of these sections (for example, if section 477 is mentioned in the text, we extract the corresponding title: *CORPORATIONS ACT 2001 - SECT 477 Powers of liquidator*). Each case has on average 10.5 related cases (either cited by the current case or citing it), and 8.3 references to law.

Our corpus thus contains the initial 2816 cases with their given catchphrases, and all cases related to them by incoming or outgoing citations, with catchphrases and citing sentences explicitly identified, and the references to acts and sections of Acts.

4 Automatic Categorization of Cases

We developed two simple baseline methods to assign labels to cases. The first one, called **CITLEG**, assigns labels based on the analysis of cited and citing cases and legislation. The underlying intuition is that a group of cases that cite each other are likely to belong to the same category. For each case, two steps are executed. In the first step the algorithm collects all the labels of all the cited and citing cases. Then it extracts the label that occurs most often (only if it appears

Table 2. The 20 most frequent labels in our corpus

Label	Counts	Label	Counts
PRACTICE AND PROCEDURE	661	INTELLECTUAL PROPERTY	61
MIGRATION	518	EVIDENCE	56
CORPORATIONS	295	CONTRACT	46
ADMINISTRATIVE LAW	235	CORPORATIONS LAW	36
COSTS	170	INCOME TAX	34
TRADE PRACTICES	161	COPYRIGHT	27
INDUSTRIAL LAW	93	PROCEDURE	24
BANKRUPTCY	86	CONTRACTS	24
NATIVE TITLE	79	MIGRATION LAW	24
TAXATION	73	EQUITY	23

in at least two cases). The label that appears second most often is also selected if it occurs in at least three cases.

In the second step we look at references to both cases and legislation, to check if reference(s) to specific Acts, sections of Acts, or previous cases, indicate (i.e. are strongly correlated with) a particular label. We take all references to any case or legislation (including specific sections) in the corpus. Then for each Act, section or case we count how many times it occurs together with each label of the corpus rather than other labels, defining:

- $YES(ref, label)$: how many times ref is cited in cases of the category $label$
- $NO(ref, label)$: how many times ref is cited in case of other categories

To select labels, in the second step of **CITLEG**, we select for each case all the labels for which there is any reference (in the case) with $YES(ref, label) \geq 5$ (the ref occurring at least five time with the corresponding label) and $YES(ref, label) \geq 2 * NO(ref, label)$ (the ref occurring at least with the label the double the times without the label).

The other baseline method that we developed is based on the hypothesis that similar cases are given the same label(s). To test this hypothesis we compute a similarity matrix between each pair of documents, and then check if similar documents have the same label. We tried three different similarity measures:

- Jaccard similarity coefficient: we look at the occurrence of every term in the documents (without counting how many times it occurs). The coefficient is the ratio between the size of the intersection of the words in the two documents, compared to the size of the union of the words in the two documents:

$$Jaccard(d1, d2) = \frac{|d1 \cap d2|}{|d1 \cup d2|} = \frac{\sum_{i=1}^N x_{i,d1} \cdot x_{i,d2}}{\sum_{i=1}^N x_{i,d1} + \sum_{i=1}^N x_{i,d2} - \sum_{i=1}^N x_{i,d1} \cdot x_{i,d2}}$$

where $x_{i,d} = 1$ if word i is in document d , 0 otherwise.

- TF cosine: we consider also the number of occurrences of each term in the documents. We build a vector of frequencies of each term for the two documents and then take as a similarity measure the cosine between the vectors:

$$TFcos(d1, d2) = \frac{\sum_{i=1}^N tf_{i,d1} \cdot tf_{i,d2}}{\sqrt{\sum_{i=1}^N tf_{i,d1}^2} \cdot \sqrt{\sum_{i=1}^N tf_{i,d2}^2}}$$

where N is the total number of terms, and $tf_{i,d}$ the number of times term i occurs in document d .

- TFIDF cosine: compute the cosine similarity using TFIDF score as the weights:

$$TFIDFcos(d1, d2) = \frac{\sum_{i=1}^N tfidf_{i,d1} \cdot tfidf_{i,d2}}{\sqrt{\sum_{i=1}^N tfidf_{i,d1}^2} \cdot \sqrt{\sum_{i=1}^N tfidf_{i,d2}^2}}$$

Where the $tfidf$ is computed as:

$$tfidf_{term,d} = tf_{term,d} \cdot idf(t) = tf_{term,d} \cdot \log\left(\frac{NDocs_{tot}}{NDocs(t)}\right)$$

In order to choose the best similarity measure for our task, we computed similarities between each pair of documents, and then measured how many documents have the same label as their nearest neighbour. Using the Jaccard coefficient, 1678 cases have the same label as their nearest neighbour (over a total of 2816, 59.6%), using TFcos the number goes up to 1756 (62.4%) and using TFIDFcos only 1513 (53.7%) cases.

For this reason we chose the TF cosine as the best way to measure similarity for the **Nearest Neighbour** method. In the task of giving labels to a set of documents, this method cannot be actually used to make predictions as is, because it needs to know the label of the nearest neighbour to assign a conclusion; however, we show later how to use it together with the knowledge base to assign labels to those cases not covered by the rules, taking the label of the most similar labelled case.

5 Building a Knowledge Base for Case Categorization

We developed an efficient knowledge acquisition framework, to build a knowledge base that assigns labels to cases. The knowledge base contains rules that specify a number of conditions, evaluated against each case. If the conditions are satisfied by a case, the label given by the rule conclusion is assigned to it.

5.1 Rule Language

Feature design is often a very critical as well as time-consuming step in developing any natural language application. For categorization tasks, often the simple presence of specific terms has been used, both in machine learning and rule based systems (i.e. [13]). However, when manually examining legal texts, we realized

that more complex features are needed, which also take into account corpus-level information as well as citations to other cases or pieces of legislation.

The proposed rules have the form *condition* \rightarrow *conclusion*; if the condition matches the current case, the conclusion specifies which label should be given to the case.

The **condition** part of each rule consists of the conjunction of constraints on two types of attributes: the first type is document-level attributes which refer to a specific label (the same as the conclusion of the rule). The document level attributes, for a given label, are:

1. **Cit** how many times the label is given in cases cited by the current case or citing it.
2. **CitPerc** how many times the label occurs in related (cited or citing) cases as a percentage of the total (i.e. at least the 40% of the related cases have this label).
3. **MaxRank** the maximum rank of the label in related cases (i.e. it is the first or second most frequent label).
4. **MinNumLeg** minimum number of legislation items or cases cited in the document, which satisfy:
 - **MinYes** minimum number of $YES(ref, label)$.
 - **MinRatio** minimum ratio $YES(ref, label)/NO(ref, label)$.

The second type of attribute refers to a specific term (one or more words), which can be equal to the label or different. To constrain this we need to specify a term (or use the label) and specify minimum and maximum thresholds for the following attributes:

5. Term frequency (**Tf**): number of occurrences of the term in this document.
6. **TFIDF**: computed as the TFIDF rank of the term in the document (i.e. $TFIDF(term)=10$ means that the term has the 10 highest TFIDF value for this document). The TFIDF formula was given in the previous Section.
7. **CitSen**: how many time the term occurs in all the sentences (from other documents) containing a citation to the present case.
8. **CitCp**: how many times the term occurs in all the catchphrases of other documents that cite or are cited by the case.
9. **CitAct**: how many time the term occurs in the titles of any Act cited by the case.

The **conclusion** of a rule is generally a label (i.e. “Costs” or “Migration”) to be assigned to the case. An example of a rule which contains conditions for a particular label is:

Tf(*native title*) ≥ 1.0 and **MaxRank**(*native title*)=0 \rightarrow label=**native title**

if the label *native title* has $Tf \geq 1$ in the document and it is the most common catchphrase among the cited/citing cases, it is assigned to the case. A rule which has condition related to other terms would be, for example:

Tf(discovery) ≥ 8.0 and CitCp(discovery) ≥ 1.0 – > label=practice and procedure

if the word *discovery* occurs at least 8 times in the document and at least once in the catchphrases of cited/citing cases, then the label “practice and procedure” is given to the case.

It is also possible to create a rule without a specific conclusion (we call these *generic* rules): the rule tests the given conditions for all possible labels, and any label which satisfies the conditions is assigned to the document. A *generic* rule, tested for each possible label, is for example:

Cit(AnyLabel) ≥ 8.0 and MaxRank(AnyLabel) = 0 – > label=MatchedLabel

anylabel must occur in at least 8 cited/citing cases and be the most common in cited/citing cases; this condition is tested for each case on all the possible labels and if a case has a label that satisfies both constraints, it is assigned to it.

5.2 Knowledge Acquisition

During the knowledge acquisition phase, rules are acquired from a user looking at examples of cases. We built a tool that allows the user to inspect the cases and create and test rules. The user, for each document, can explore the relevant labels and the other catchphrases of the case. The interface also shows citation information, the catchphrases and relevant sentences of cited/citing cases, and which parts of the relevant legislation are cited with their titles. The user can also see frequency information (according to the attributes previously described) for different terms, including the terms that form the given label.

During knowledge acquisition, the user looks at a case not yet classified, and writes a rule to assign the correct label to it: the user picks relevant attribute(s) that would give the correct classification to the current case and creates one or more conditions. While making the rule, it is tested on all the training corpus; in this way the user can see, after adding each condition, the number of cases correctly and wrongly matched by the rule. The system also presents other matches for manual inspection, and lists which kind of cases were misclassified by the rule, listing the correct labels of the misclassified cases. For a *generic* rule, we can also see which labels were matched. Table 3 gives an example of the kind of information given by the system to assist rule creation. Using this information and looking at specific examples, the user can refine the rule until satisfied and then commit it to the knowledge base.

While adding a condition to a rule we can also compute the probability that the improvement given by the rule/condition is random. As initially described in 5, for a two-class problem (the case has the label/the case does not have the label), we can use a binomial test to calculate the probability that such results could occur randomly. That is, when a condition is added to an existing rule, or added to an empty rule, we compute the probability that the improvement is random. The probability of selecting randomly n cases and getting x or more cases with the specified label is:

$$r = \sum_{k=x}^n \binom{n}{k} p^k (1-p)^{n-k} = \frac{n! p^x (1-p)^{n-x}}{x!(n-x)!}$$

where p is the random probability, i.e. the proportion of cases with that label among all cases selected by the current rule. If we know how many cases the new condition correctly labels (x), we can calculate this probability which can help the user in creating a condition that minimize the value of r .

Despite taking inspiration from the Ripple Down Rule methodology, the rules are not organized in a tree. That is, when a new rule is created it is added to the knowledge base without any particular order. When a new case is classified, all the rules are executed and any rule that fires (i.e. the case satisfies the conditions of the rule) assigns a label to the case. In this way every case can be assigned more than one label if more than one rule fires, but also it will not be assigned a label if no rule fires.

We tried to make the knowledge acquisition process quick and easy for the user: when creating a rule the user is guided both by the particular examples shown by the system, and by the statistics computed on the large training dataset. The user can see all the applicable attributes and pick those which seem most relevant for the case at hand and testing the rule on all the corpus he/she can get an impression of the quality of the rule and refine it (potentially looking also at other examples), until he/she is satisfied. The rule is then committed to the knowledge base, without having to worry about any interaction with previous rules.

Table 3. Example of system information while creating a rule

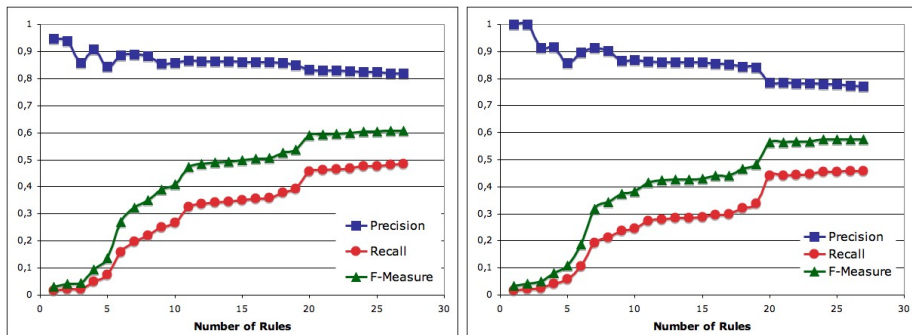
<p>Tf(<i>native title</i>)≥1.0 and MaxRank(<i>native title</i>)=0 - > label=native title 1: Matches= 54/57=0.947 new matches= 27/28=0.964 2: Total 'native title' = 79 matched= 54/79=0.683 3: Errors: 'aborigines': 2, 'aboriginals': 1, 'costs': 1 4: Probability Random improvement= 4.18e-80</p> <hr/> <p>Row1: The rule matches 57 cases, of which 54 (94.7%) are correct (have "native title" label). Of these only 28 (27 correct) were not matched by the rules already in the KB. Row2: The total number of cases with the label native title is 79 (so this rule covers 54/79=68.3% of them). If the conclusion was <i>generic</i> this row would give a list of labels posted by the rule. Row3: The three cases which are labelled incorrectly have labels 'aborigines' (twice), 'aboriginals' (once) and 'cost' (once) (one of the cases has two labels). Row4: The probability that a random rule would be this good is 10e-80.</p>
--

6 Experimental Results

Following the procedure just described we created a knowledge base of 27 rules. Of the 27, 7 rules are of type *generic*, i.e. they can assign different labels to cases, while the other 20 deal with a specific label only. These rules assign label(s) for

1846 cases (out of 2816 cases), giving in total 2051 labels, of which 1681 (81.96%) are correct. 1653 cases are assigned only one label, while 181 cases are given two labels and 12 have three. The remaining 970 cases were not assigned any label by the rules. There are also occasions in which multiple rules fire for the same case giving the same label, thus increasing the confidence that the label is appropriate for the case (for example 1320 labels were assigned at least by two rules, of which 1185, 89.77%, are correct). The rules were inserted by one of the authors, and no legal expert was involved. The time to create the rules was logged, and we measured an average time of 4.5 minutes per rule (121 minutes in total) which includes looking at a case, choosing the attributes, testing and refining the rule and committing it.

We then evaluated the rule base on a unseen test set of 1073 cases (from the Federal Court of Australia, year 2006, as described in Section 3). The rules assigned a label to 663 cases (585 cases get one label, 71 get 2, 7 get three, while the remaining 410 are not assigned a label). Of the 748 labels given in total, 576 are correct (77.00%). The trends for precision and recall as rules are added to the knowledge base are shown in Figure 1a for the training set and in Figure 1b for the test set.



(a) Training set: 2816 documents.

(b) Test set: 1073 documents.

Fig. 1. Precision, Recall and F-measure as the size of the KB increases

To better understand the performance of the knowledge base, we compared the results with those obtained with a simple automatic method **CITLEG** and with some Machine Learning algorithms. For machine learning, we initially used as features a bag of words representation, where each case is described by all the words occurring in it. We removed stopwords from and stemmed all the remaining terms, using the NLTK library² and the included Porter stemmer. We used as features for each document the presence or absence of every term in the corpus. In addition, we also trained the models with an alternative representation, using the same features of the rule system: for each case we encoded the values

² <http://www.nltk.org/>

of the nine features described in Section 5 for every possible label. This representation makes use of citation information which is not available in the bag of words model. We trained two kinds of machine learners, a Naive Bayes model and a Support Vector Machine model, using the implementations in WEKA³ and LibSVM⁴ respectively. Both models have been successful in many natural language processing applications, and we chose one model from the class of discriminative models (SVM) and one from generative models (Naive Bayes). We also experimented with decision trees (in particular J48) but as this did not improve the performance, the results are omitted from the paper.

The performance of the knowledge base is compared with a range of baselines in Table 4 for the training set, and Table 5 for the test set. The methods in the tables are:

- **KB**: the knowledge base of 27 rules.
- **CITLEG**: the method based on citations to cases and legislation, described in Section 4.
- **KB+CITLEG**: for each case, if the rules assigns at least one label to the case, we consider only that label(s); if no rule fires we use CITLEG to get a label for the case.
- **KB+NN**: for each case, if the rules assign at least one label to the case, we consider only that label(s), otherwise we follow the list of nearest neighbours (as defined by TF cosine similarity, see Section 4) until we find a case which has at least one label assigned by the rules, and use the same label for the current case.
- The two machine learners: naive bayes (**NB bow** and **NB myfeat**) and support vector machine (**SVM bow** and **SVM myfeat**), both using the bag of words representation or using the features we designed.

We can see from the two tables that, on the unseen test set, the rules obtain greater precision (77%) than any other method. However, because the rules tend not to cover every case, it is the combination of the rules with the automatic methods that give the highest recall and F-measure. In particular using the knowledge base and the nearest neighbour method as the backup for cases not covered by the rules gives the best performance with an F-measure of 63%.

The machine learning algorithms seem to overfit the training data to a much higher degree than the other methods, with a low performance on unseen test data. The rules on the opposite suffer only a minor loss of performance when tested on new unseen data. Both Naive Bayes and SVM fail in generalizing the features we designed, and we obtained better performance using the more standard bag of words representation; however this also failed to outperform even the quite simple baseline CITLEG. We suggest that since the models were trained with over 10 millions words, it seems unlikely that significant improvement would be obtained by providing more training data for the learners.

³ <http://www.cs.waikato.ac.nz/ml/weka>

⁴ <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

We can conclude that knowledge acquisition is much more successful in encoding the way complex features interact with each other in different cases. Once the feature set was designed, it took a relatively short time to create high precision rules which cover a significant portion of the cases. Another advantage of using our knowledge acquisition framework is that the rule base can easily be extended, inserting additional rules, for example to cover new classes or to use new features, while the machine learners would need to be re-trained and would require new annotated data.

Table 4. Performances measured on the training set (2816 cases)

Method	AvgLabelsPerDoc	Precision	Recall	F-measure
KB	0.73	0.820	0.484	0.609
CITLEG	1.44	0.562	0.631	0.595
KB+CITLEG	1.07	0.694	0.598	0.643
KB+NN	1.14	0.675	0.621	0.647
NB bow	1.00	0.462	0.371	0.411
SVM bow	1.00	0.257	0.207	0.229
NB myfeat	1.00	0.610	0.490	0.543
SVM myfeat	1.00	0.997	0.801	0.889

Table 5. Performances measured on the test set (1073 cases)

Method	AvgLabelsPerDoc	Precision	Recall	F-measure
KB	0.70	0.770	0.459	0.575
CITLEG	1.32	0.555	0.600	0.577
KB+CITLEG	1.03	0.667	0.578	0.620
KB+NN	1.16	0.631	0.629	0.630
NB bow	1.00	0.492	0.421	0.454
SVM bow	1.00	0.314	0.269	0.290
NB myfeat	1.00	0.194	0.166	0.179
SVM myfeat	1.00	0.259	0.222	0.239

7 Extracting Lower-Level Catchphrases

As outlined in the Introduction, categorising the cases according to the top level hierarchy is part of a wider project of generating the full range of catchphrases for legal case reports. As shown in Table 4, together with one or two high level catchphrases, there are several lower level, more specific catchphrases, whose re-use among cases is very limited. To tackle this kind of catchphrase, we propose to use an extractive approach, where important fragments of text are extracted from the case reports as candidate catchphrases.

We are extending our knowledge acquisition framework for this task and plan to build a knowledge base with rules that operate at the sentence level, and

Table 6. Examples of rules for catchphrase extraction

SENTENCE contains at least 2 terms with $Tf > 30$ and $CpOcc > 200$ and $AvgOcc < 2.5$ and $TFIDF < 10$ within a window of 2 terms
SENTENCE contains at least 2 terms with $Tf > 5$ and $CpOcc > 20$ and $FcFound > 0.02$ and $CitCp > 1$ and $TFIDF < 15$ and contains at least 2 terms with $Tf > 5$ and $CpOcc > 2$ and $FcFound > 0.11$ and $AvgOcc < 0.2$ and $TFIDF < 5$
SENTENCE contains at least 10 terms with $CitCp > 10$ and contains at least 6 terms with $CitCp > 20$
SENTENCE contains the term <i>corporations</i> with $Tf > 15$ and $CitCp > 5$

provide conditions to enable a sentence to be recognized as *relevant* and thus extracted and proposed as a catchphrase. Similarly to the approach described above, we plan to acquire a range of rules from a user, and use some simpler automatic methods as a fall-back for the cases not covered by the knowledge base.

In order to extend the approach for relevance identification, we need to extend the feature set. While the rules identify important sentences, looking at one sentence by itself is clearly not enough to decide its importance; we must consider also document-scale information to know what the present case is about, and at the same time we need to look at corpus-wide information to decide what is peculiar to the present case. For this reason we have developed several ways of locating potential catchphrases in legal text, based on different kinds of attributes, such as frequency attributes (how many times a term appear in the document, how many times on average in other documents, in how many documents it appears, tfidf score, etc.), statistical information (i.e. likelihood of appearing in catchphrases), citation information (such as occurrences in citation sentences, in catchphrases of related cases and in the titles of various legislation), and linguistic information such as part of speech tags, the use of specific dictionaries, patterns of words etc.

We hypothesize that such a rich feature set should allow us to create high precision rules to identify relevant sentences in the full text, using a knowledge acquisition procedure similar to the one already described, where a user looks at specific cases to select different attributes and their weights, and is guided by the system both with specific examples and with statistics gathered from the training set. To determine the quality of the results, we developed also an evaluation method based on approximate matching between the sentences and the given catchphrases using Rouge [14], which would allow us to quickly estimate the performance of a set of rules on a large dataset of catchphrases. The evaluation method was described in [7], some preliminary results using knowledge acquisition were presented in [6]. Table 6 gives some examples of possible rules that identify important sentences.

8 Conclusion

This paper presents our knowledge acquisition approach to generate different levels of catchphrases for legal case reports, to assist legal research and provide help to lawyers searching for relevant precedents.

To give high level labels to cases, we first developed a baseline automatic method of selecting labels looking at citation information. We then designed a range of more complex features, including citation and frequency information, and developed an efficient knowledge acquisition framework that allows the quick creation of classification rules using those features. A small knowledge base was built using this tool in approximately two hours. When evaluated on unseen data, the knowledge base, in conjunction with a nearest neighbour back-up method, obtains a precision of 63.1% and a recall of 62.9%, outperforming two different machine learning models trained both using the designed features and a more traditional bag of words representation.

We conclude that our knowledge acquisition approach seems better able to capture the interaction between the complex features required to perform classification in this challenging domain. Machine learning, despite being given the same features and a large amount of training data, fails to generalise to unseen data.

We are now extending this approach for the task of extracting more specific and longer catchphrases from the full text of the cases. For this problem the feature space required to capture relevance in the text is even more complex, and it thus seems unlikely that machine learning could obtain good results on this challenging task. We believe the best approach is to extend our knowledge acquisition framework to handle a larger number of features and to present a more thorough evaluation of rule performance (needed to guide the knowledge acquisition), while keeping the creation of rules as simple as possible for the users.

Acknowledgments. The authors would like to thank Philip Chung, Andrew Mowbray and Graham Greenleaf from AustLII for their valuable help and constructive comments.

References

1. Ashley, K.D., Brüninghaus, S.: Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law* 17(2), 125–165 (2009)
2. Compton, P., Jansen, R.: Knowledge in context: a strategy for expert system maintenance. In: *AI 1988: Proceedings of the Second Australian Joint Conference on Artificial Intelligence*, pp. 292–306. Springer, Adelaide (1990)
3. Farzindar, A., Lapalme, G.: Letsum, an automatic legal text summarizing system. In: *The Seventeenth Annual Conference on Legal Knowledge and Information Systems, JURIX 2004*, p. 11. Ios Pr. Inc. (2004)
4. Farzindar, A., Lapalme, G.: Machine Translation of Legal Information and Its Evaluation. In: Gao, Y., Japkowicz, N. (eds.) *AI 2009. LNCS*, vol. 5549, pp. 64–73. Springer, Heidelberg (2009)

5. Gaines, B.R., Compton, P.: Induction of ripple-down rules applied to modeling large databases. *J. Intell. Inf. Syst.* 5, 211–228 (1995)
6. Galgani, F., Compton, P., Hoffmann, A.: Combining different summarization techniques for legal text. In: *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pp. 115–123. Association for Computational Linguistics, Avignon (2012)
7. Galgani, F., Compton, P., Hoffmann, A.: Towards Automatic Generation of Catchphrases for Legal Case Reports. In: Gelbukh, A. (ed.) *CICLING 2012, Part II*. LNCS, vol. 7182, pp. 414–425. Springer, Heidelberg (2012)
8. Galgani, F., Hoffmann, A.: LEXA: Towards Automatic Legal Citation Classification. In: Li, J. (ed.) *AI 2010*. LNCS, vol. 6464, pp. 445–454. Springer, Heidelberg (2010)
9. Gonçalves, T., Quaresma, P.: Is linguistic information relevant for the text legal classification problem? In: *ICAIL 2005*, pp. 168–176 (2005)
10. Greenleaf, G., Mowbray, A., King, G., Van Dijk, P.: Public Access to Law via Internet: The Australian Legal Information Institute. *Journal of Law and Information Science* 6 49 (1995)
11. Hachey, B., Grover, C.: Extractive summarisation of legal texts. *Artif. Intell. Law* 14(4), 305–345 (2006)
12. Kim, M.H., Compton, P., Kim, Y.S.: Rdr-based open ie for the web document. In: *Proceedings of the Sixth International Conference on Knowledge Capture, K-CAP 2011*, pp. 105–112. ACM, New York (2011)
13. Krzywicki, A., Wobcke, W.: Incremental E-Mail Classification and Rule Suggestion Using Simple Term Statistics. In: Nicholson, A., Li, X. (eds.) *AI 2009*. LNCS, vol. 5866, pp. 250–259. Springer, Heidelberg (2009)
14. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out: Proceedings of the ACL 2004 Workshop*, pp. 74–81. Association for Computational Linguistics, Barcelona (2004)
15. de Maat, E., Krabben, K., Winkels, R.: Machine learning versus knowledge based classification of legal texts. In: *Proceedings of the 2010 Conference on Legal Knowledge and Information Systems*, pp. 87–96. IOS Press, Amsterdam (2010)
16. Moens, M.F.: Summarizing court decisions. *Inf. Process. Manage.* 43(6), 1748–1764 (2007)
17. Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput. Surv.* 34(1), 1–47 (2002)
18. Thompson, P.: Automatic categorization of case law. In: *ICAIL 2001: Proceedings of the 8th International Conference on Artificial Intelligence and Law*, pp. 70–77. ACM, New York (2001)
19. Xu, H., Hoffmann, A.: RDRCE: Combining Machine Learning and Knowledge Acquisition. In: Kang, B.-H., Richards, D. (eds.) *PKAW 2010*. LNCS, vol. 6232, pp. 165–179. Springer, Heidelberg (2010)
20. Zhang, P., Koppaka, L.: Semantics-based legal citation network. In: *ICAIL 2007: Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pp. 123–130. ACM Press, New York (2007)

Extraction of How-to Type Question-Answering Sentences Using Query Sets

Kyohei Ishikawa and Hayato Ohwada

Department of Industrial Administration, Faculty Of Science and Technology,
Tokyo University of Science,
2641 Yamazaki Noda-shi Chiba-ken 278-0022 Japan
kyouhei22@gmail.com, ohwada@ia.noda.tus.ac.jp

Abstract. In this study, we describe how to extract a sentence which serves as a correct answer for a how-to question about actions to be taken on the Web. The distinguishing feature of this study is the extraction of sentences using a set of queries, after collecting queries considered to be effective for answer extraction. By using a query set, it is possible to extract the answer considered to express the contents of the Web page pertinent to the question. In the results of the experiment, the average recall was 77.3% and the average precision was 62.9%.

Keywords: How-to type question answering, Web documents, query set, extraction.

1 Introduction

Due to the rapid spread of the Internet, there are vast amounts of information available. For day-to-day questions, the opportunity to find answers in information found on the Web using a search engine has been increasing. However, to use existing search engines, it is necessary to adequately represent the question in the search query. In addition, there is the problem that we must look for ourselves to locate answer phrases and sentences on the pages that are found.

In order to solve such a problem, question answering system to extract the answer and present it has been studied for natural language question. In the existing study, question answering system is divided into two types: One is the inference type system, another is extraction type system. Since the former are limited areas that can be answered, extraction type has mainly been studied in recent years.

In the extraction type, to conduct study [1] assuming that answers for “WH” questions are nouns and noun phrases is a mainstream approach. In contrast, studies of the non-factoid type, in which the answer is extracted as text corresponding to “why” questions that ask for a reason or “how” questions that ask for actions, have been less prevalent than studies corresponding to WH questions.

The purpose of this study is to extract correct answers for questions without leakage from sentences on the Web, focused on how-to questions. How-to type

question is a question which asks description of means. In this study, we propose a method for extracting question-answering sentences that we have implemented, and evaluate its effectiveness.

2 Related Work

Studies by Asano et al. and Murdoch et al. are related studies of how-type systems [2,3]. These studies specify an approach in which a “method” and “procedure” are described in terms of bulleted lists, table structures, and the structure of the HTML document. In order to measure whether it is a satisfying answer for a how-to question Asano et al. introduced two rate scales in the accuracy of explanation, and the viewpoint of the detail. One is the degree of answer agreement and another is the degree of explanation detailed. And they verified what kind of feature there is in a document with a high rate scale, and whether a rate scale could be predicted from the feature of a document for the document searched by the how-to question.

Yamamoto et al. [4] make use of the property by which a word from the same vocabulary chain found in multiple sentences takes on similar meanings because of its action expression. Thus, that which should be considered as the correct answer is described over multiple sentences. They found that the specification of an appropriate range was difficult. The range over which a similar noun has chained and that which a chain follows is extracted as the answer. Here, the recall was 56.2% and the precision was 56.6%.

3 Proposed Method

In this paper, we propose a method for extracting an answer focusing on the number of occurrences of a specific word that are included in the Web document to realize a how-to question-answering system.

3.1 Summary the Proposed Method

A word appearing the most in a given document is clearly relevant to the document content. We realized that a search question and the word with the highest frequency of appearance had a strong relation in the Web documents returned by the search engine. We considered that whether or not that word is contained in a given sentence might be an indicator of whether that sentence is an appropriate answer. Moreover, we considered that the word with the highest frequency of appearance was a more important word, not only within one document but over two or more documents. Finally, we considered that a sentence containing two or more of these words is more appropriate than a sentence that contains only one of these words. With these considerations in mind, we proposed a method to extract such sentences. Figure 1 gives an outline of the proposed method.

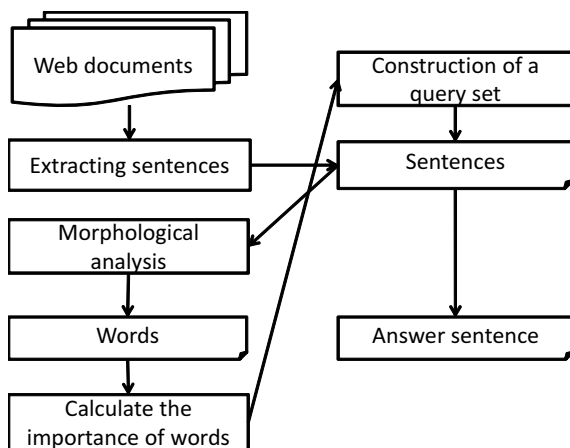


Fig. 1. Flowchart of the proposed method

3.2 Construction of a Query Set

First, we conducted a morphological analysis of sentences extracted from the Web document and decomposed into words. Next, we calculated the importance of each word. The importance of a word was defined by the Eq. (1) below.

$$I_w = \frac{1}{\log \frac{|D|+1}{|d:d \ni t_j|}} \times N \quad (1)$$

In Eq. (1), D is the total number of documents, d is the number of documents containing the word i , N is the number of appearance of the word. The number of appearances of the word refers to the number of appearance of words in all documents. The importance refers to the TF-IDF method. This method assumes that a word used by a few Web documents, rather than a word used by many Web documents, best expresses the subject of the document. In this study, we used the reciprocal because we want words that are used in many Web documents to be considered important words. By using the reciprocal, the importance of words that appear in many Web documents will be raised. A threshold value is provided in this importance measure, and words scoring higher than that threshold value are considered for the query set used for sentence extraction. Figure 2 gives an outline of the proposed method. In Fig.2, Words that threshold is greater than 10 are contained in a query set.

Referring to the query set, we extract the sentences that include two or more words in the query set as answer sentences. Figure 3 presents an example of extraction. In this example, the underlined sentences were extracted as answers. We also set sentences with overlapping contents to one after extracting the answer sentence.

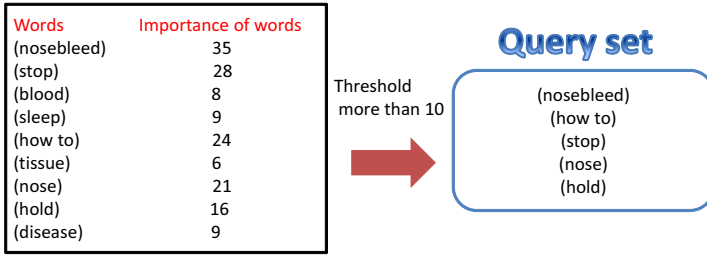


Fig. 2. Example of construction of a query set

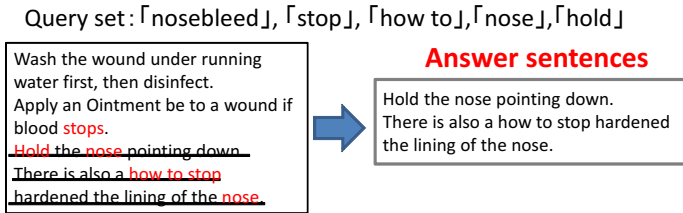


Fig. 3. Example of answer extraction

3.3 Consideration of Part of Speech

When there are many words included in a query set, many sentences will be extracted as answers. Thus, we consider the following factors when extracting an answer.

- (1) Some words are not included in the query set.

Auxiliaries and particles, as well as the verbs “be” and “do” and the nouns “become”, “answer”, and “question” of the noun appear so often that they cannot form a valid query to extract an answer. Therefore these words are not included in a query. Moreover, unknown words and signs are not included in a query.

- (2) A sentence that does not contain a verb is not taken as an answer.

Since a how-to question is a question that asks for a method, a sentence without a verb is not suitable as an answer. Therefore, any sentence that does not contain a verb is not extracted as an answer.

- (3) The part of speech of the sentence ending is considered.

As a result of investigating the part of speech of the sentence ending of a correct answer for a how-to question, we determined that the part of speech of many sentence endings is a verb, a particle, or an auxiliary verb. Therefore, we extracted sentences whose endings matched the parts of speech of the answer.

4 Experiment

We conducted an experiment in order to evaluate whether the method proposed in this study is effective. The proposed method was implemented in Java. We used Sen for the morphological analysis and to track the stored words and the importance of words in MySQL. We constructed query sets by performing searches from the MySQL table which associated words and the importance of words.

4.1 Experiment Method

We obtained sentences from Web pages using existing search engines such as Google. Since Google regards a page that is linked from many other pages as high priority, we can incorporate this measure with partial reliability. We extracted answers from sentences on the top 20 Web pages (a total of 6910 sentences) written in Japanese, with each question sentence turned into a query using the proposed method.

4.2 Evaluation Method

We evaluated the how-to question-answering system using the recall, precision, and F-score given by Eq. (2), Eq. (3), and Eq. (4).

$$recall = \frac{\textit{Num of correct sentence extracted}}{\textit{Num of total correct sentence}} \quad (2)$$

$$precision = \frac{\textit{Num of correct sentence extracted}}{\textit{Num of total sentence}} \quad (3)$$

$$F - score = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} \quad (4)$$

In the experiment, we used the ten questions (how to stop a nosebleed and how to run fast etc.) used by Yamamoto et al. [4].

In addition, “a correct sentence” is defined as a sentence that has been shown to resolve the questions in this experiment. These were evaluated by hand, without considering whether the method was actually effective. We provide an example of the evaluation of answers extracted for “How to stop a nosebleed” in Table 1.

4.3 Experiment Result

We provide a threshold for the importance of a word when constructing a query set, including a word if its importance exceeds the threshold in the query set and thus constructing the query set. Therefore, the words contained in a query

Table 1. Example of evaluation of answers extracted for “How to stop a nosebleed”

Evaluation	Extracted@answer sentence
	Knob to the bottom of the nose with your thumb and forefinger, press 5 to 10 minutes.
	Since its bleeding from a capillary vessel in the case of a nose-bleed, it is most that a life is not in danger, but it will be serious if an allowance is overdue in case of general bleeding.

set are determined by a threshold, and the sentences extracted as answers are determined by a threshold. We extracted answers while changing the threshold for each question, thus evaluating the average of each value in order to determine whether the current threshold is optimal for extracting answers. We present the relationship between the threshold and the evaluated variables in Table 2.

The largest F value occurs when the threshold is 10. Table 3 compares the extraction accuracy of previous studies by Yamamoto et al. [4].

Table 2. Relation between the threshold and the evaluated variables

Threshold	5	10	15	20	25
recall	79.41	77.32	69.58	59.61	48.42
precision	58.28	62.85	58.30	58.19	52.95
F-score	67.22	69.34	63.44	58.89	50.58

Table 3. Comparison of extraction accuracy

	proposed method	Yamamoto et al.
recall	77.3	56.2
precision	62.9	56.6
F-score	69.3	56.4

4.4 Discussion

As the cause by which F value became the maximum when the threshold is 10, the number of the words contained in a query set is considered to be the cause. When a threshold value is small, the number of the words contained in a query set will increase, many incorrect answer sentences to the question will be extracted, and precision will fall. If a threshold value is high, the number of the words contained in a query set will decrease, and correct answer sentences will not be extracted, and recall will fall. For these reasons, we considered F value has become the maximum when a threshold value is 10.

The average recall was 77.3% and the average precision was 62.9% when the threshold is 10. As the cause of a precision fall, since we used the number of appearances of a word in all pages when determining the importance of that

word, a word from a sentence that is not a suitable answer of a how-to question may have been used as a query. Another reason might be that a query used when extracting a sentence is often used in a statement unrelated to resolving the question. In this case, it is assumed that we have extracted an unrelated sentence.

5 Conclusion

This study has proposed an extraction method focused on the importance of the words used in answer extraction to realize a how-to question-answering system. Since the average recall in the evaluation experiment was 77.3%, a sufficiently high value compared with preceding studies to validate the proposal of this study.

Future work will seek to increase both recall and precision by deleting sentences that are unsuitable answers and by specifying and narrowing the queries to extract more appropriate answers. In addition, it is also necessary to extract a range of two or more sentences because some Web pages do not convey the desired meaning in a single sentence. Moreover, it is necessary to conduct experiments using large-scale data and performance scoring to narrow the extracted answer sentences in order to realize a system.

References

1. Kupiec, J.: MURAX: Finding and Organizing Answers from Text Search. In: Strzalkowsky, T. (ed.) *Natural Language Information Retrieval*, pp. 311–332. Kluwer Academic Publishers (1999)
2. Asanoma, N., Furuse, O., Kataoka, R.: Feature Analysis of Explanatory Documents for How-to Type Question Answering. *Information Processing Society of Japan*, 55–60 (2005)
3. Murdock, V., Kelly, D., Croft, W.B., Belkin, N.J., Yuan, X.: Identifying and improving retrieval for procedural questions. *Information Processing and Management* 43(1), 181–203 (2007)
4. Masanori, Y., Kanamori, K., Fukuda, M., Nobesawa, S., Tahara, I.: Extraction of Descriptive Parts Based on Chains of Action Expressions. *Institute of Electronics, Information, and Communication Engineers*, p. 53 (2007)

Image Indexing and Retrieval with Pachinko Allocation Model: Application on Local and Global Features

Ahmed Boulemden¹ and Yamina Tlili²

¹ Badji Mokhtar Annaba University, Algeria
ahmed_boulemden@yahoo.com

² LRI -lab, Badji Mokhtar Annaba University, Algeria
guiyam@yahoo.fr

Abstract. We present in this paper a part of our work in the field of image indexing and retrieval. In this work, we are using a statistical probabilistic model called Pachinko Allocation Model (PAM). Pachinko Allocation Model (PAM) is a probabilistic topic model which uses a Discrete Acyclic Graph (DAG) structure to present and learn possibly correlations of topics which were responsible of generating words in documents, like other topic models such as Latent Dirichlet Allocation (LDA), PAM was originally proposed for text processing, it can be applied for image retrieval since we can assume that image is a text and parts of image (local points, regions, ...) can represent visual words like in text processing field. We propose to apply PAM on local features extracted from images using Difference of Gaussian and Salient Invariant Feature Transform (DoG/SIFT) techniques. In a second part, PAM is applying on global features (color, texture ...), these features are calculated for a set of regions resulting from 4x4 division of images. The proposition is under experimental evaluation.

Keywords: Image Indexing and Retrieval, Pachinko Allocation Model, Scale Invariant Feature Transform, Image Local Features, Color Texture Features, Global Features.

1 Introduction

Managing a huge amount of data including images is a widely occurred problem due to the expansion of computer science and multimedia, retrieval tasks become more and more difficult. Image retrieval is an important field of research in image processing applications and it knows lot of works and propositions, among them the use of "probabilistic topic models" which have been originally developed in the context of text modeling. Topic models try to find and use latent (hidden) semantic spaces that are more accurate to model documents in the context of retrieval tasks and overcome problems such as synonymy and polysemy due to count vector representation of documents, instead, each document is generally assumed to consist of multiple hidden topics that are responsible of generating words in the document. Images can also considered as mixture of topics (a mixture of one or more object/object parts), this fact allow the use of these models in image modeling and retrieval.

Latent semantic analysis LSA [3] was the first model to apply in image indexing and retrieval, in [7] they use LSA to construct an image search engine on the web called ImageRover, LSA uses a Singular Value Decomposition (SVD) for mapping into the semantic space. The pLSA [4] is the probabilistic variant of LSA. Instead of using SVD, it assumes a probabilistic model where each document is represented by a mixture of topics (hidden topics), each topic denotes a distribution over the discrete words (visual words in image retrieval context).

Latent Dirichlet Allocation LDA [1] is a generative probabilistic model which is similar to pLSA. The main difference is that topic probabilities can be easily assigned to new document which not the case for pLSA. Correlated Topic Model CTM [2] is another topic model which is similar to LDA but with some differences concerning topic proportions. Pachinko Allocation Model (PAM) [12] is a topic model which has been used for object recognition tasks [9], and it mainly differs from previous models by the possibility of capturing correlation between topics and not only between words. pLSA, LDA, and CTM have been studied and compared in [5] which represent an important reference in the use of topic models in both object recognition and retrieval tasks, another important point of [5] is that his work focus on the use of topic models in "real-world" noisy databases and, for this purpose she used a database consist of more than 240,000 images which have been downloaded from the public Flickr repository.

In this paper, we propose the application of PAM in two ways, first with local features of images extracted by Difference of Gaussian and Scale Invariant Feature Transform (DoG/SIFT) technique, while in the second one, it is used with a set of color and texture features, note that PAM was absent from [5].

The paper's overview is as follow. We will first mention the technique used for extracting features from images, then we will speak about the appropriate representation of these feature in order to use them in topic models. PAM will be presented in the 3rd section.

2 Image Features Extraction

We work in the first part on local features like in [5], those features are calculated at interest points in images, DoG (Difference of Gaussian) detector is used to detect such interest points and their associated regions. The DoG is scale-invariant region detector which first detects a set of interest points, then it filters this set to preserve only points that are stable under a certain amount of additive noise. First, keypoints (interest points) are identifying by scanning the image over location and scale. It detects localization and scale of keypoints as scale-space extrema of the function $D(x, y, \sigma)$, which is the difference-of-Gaussian function convolved with the input image $I(x, y)$:

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)$$

Where k indicates a constant multiplicative factor and $G(x, y, \sigma_i) = \frac{1}{2\pi\sigma_i^2} e^{-(x^2+y^2)/2\sigma_i^2}$ is a Gaussian kernel.

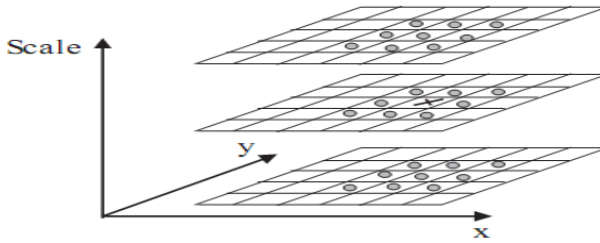


Fig. 1. Detection of extrema in scale-space by comparing a pixel(x) to its neighbors (◦) in the current and adjacent scales [9]

Local 3D extrema of $D(\cdot)$ are detected by comparing each pixel to its 26 neighbors, 8 neighbors in the current scale space level and 9 from both above and below space levels (see Figure 1). A point is selected only if it is larger or smaller than any of these neighbors.

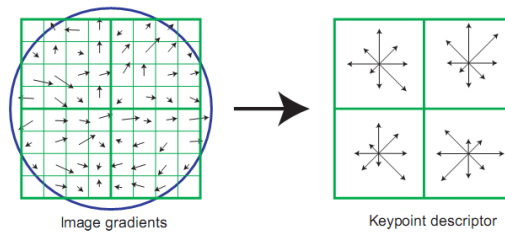


Fig. 2. 2x2 descriptor array computed from an 8x8 set of samples [9]

Computation of features for keypoints detected by DoG detector is realized with the Scale Invariant Feature Transform SIFT [9]. First, an orientation, scale, and location are assigned to keypoints. The scale and location are determined by DoG detector, while one or more orientation are assigned to the keypoint based on the dominant gradient orientation of the local image patch surrounding the interest point. An orientation histogram is used to identify dominant gradient directions by selecting peaks within. This histogram is formed from the gradients' angles of sample points within a region around the keypoint, weighted by each gradients' magnitudes. An interest point is created with that orientation for each dominant orientation (multiple interest points might be created for the same location and scale, but with different orientations)[5]. The descriptor is formed from a vector containing the values of all the orientation histogram entries, a 4x4 array of histograms with 8 orientation bins in each is used, that means $4 \times 4 \times 8 = 128$ element feature vector for each keypoint [9]. The vector is normalized to ensure invariance to illumination conditions. SIFT features are also invariant to small geometric distortions and translations due to location quantization [5]. More details about SIFT technique are available in [5].

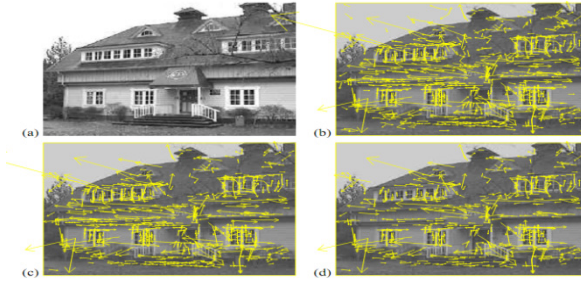


Fig. 3. Stages of keypoints selection [9]

Since local features for images have been calculated, we need to define an appropriate representation to use with topic models. Thus, an equivalent to word as elementary parts in documents has to be found for images known as “visual words”. SIFT descriptors extracted are high dimensional and their entries are continuous, thus a vector quantization is applied to derive discrete visual words. We apply k-means clustering on SIFT features vectors of each image and we keep the mean (cluster’s centroid) of resulting clusters as a visual word and the dimension of its cluster as the term frequency. We need also to cluster again all the centroids extracted by the first clustering in order to construct the corpus vocabulary and thus a bag-of-words model can be derived.

In the second experiment, images are divided into 4x4 regions, then color and texture features are calculated for every region. For color features, we chose to calculate color moments in RGB and HSV spaces, while for texture, we calculate 6 of Haralick texture indices. Each region is considered as a visual word and a kmeans clustering is performed to construct the corpus vocabulary.

We will present next the Pachinko Allocation Model.

3 Pachinko Allocation Model

The Pachinko Allocation Model (PAM) [12] is a probabilistic topic model which uses a Discrete Acyclic Graph (DAG) structure to present and learn possibly sparse topic correlations. In PAM, the concept of topics is extended to be distribution not only over words, but also over other topics. PAM connects words in V and topics in T with an arbitrary DAG, where topic nodes occupy the interior levels and the leaves are words. Graphical model of PAM is presented in figure 3.

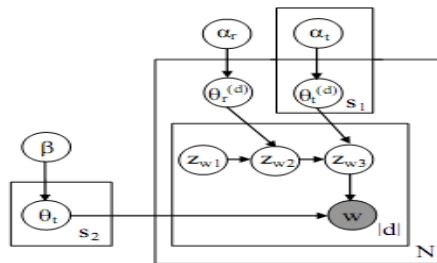


Fig. 4. Graphical Model of PAM [12]

To generate a document in PAM model we follow the process:

- Sample $\theta_{t_1}^{(d)}, \theta_{t_2}^{(d)}, \dots, \theta_{t_s}^{(d)}$ from $g_1(\alpha_1), g_2(\alpha_2), \dots, g_s(\alpha_s)$, where $\theta_{t_i}^{(d)}$ is a multinomial distribution of topic t_i over its children.
- For each word w in the document,
 - Sample a topic path z_w of length $L_w: < z_{w1}, z_{w2}, \dots, z_{wL_w} >$. z_{w1} is always the root and z_{w2} through z_{wL_w} are topic nodes in topics T . z_{wi} is a child of $z_{w(i-1)}$ and it is sampled according to the multinomial distribution $\theta_{z_{w(i-1)}}^{(d)}$.
 - Sample the word w from $\theta_{z_{wL_w}}^{(d)}$.

The joint probability of generating a document d is:

$$P(d, z^{(d)}, \theta^{(d)} | \alpha) = \prod_{i=1}^s P(\theta_{t_i}^{(d)} | \alpha_i) \times \prod_w \left(\prod_{i=2}^{L_w} P(z_{wi} | \theta_{z_{w(i-1)}}^{(d)}) P(w | \theta_{z_{wL_w}}^{(d)}) \right)$$

And a marginal probability of document d is:

$$P(d | \alpha) = \int \prod_{i=1}^s P(\theta_{t_i}^{(d)} | \alpha_i) \times \prod_w \sum_{z_w} \left(\prod_{i=2}^{L_w} P(z_{wi} | \theta_{z_{w(i-1)}}^{(d)}) P(w | \theta_{z_{wL_w}}^{(d)}) \right) d\theta^{(d)}$$

Four level PAM is often used, and it differs from PAM which allows arbitrary DAGs to model topic correlations. PAM has been used in [8] under the context of object recognition task.

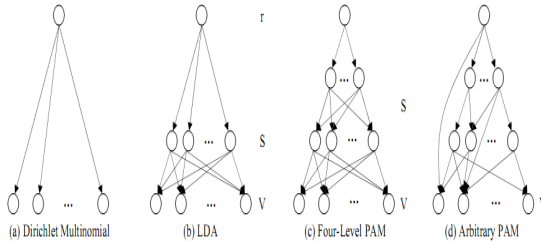


Fig. 5. Model structure for four topic models [12]

We are interested to infer topics and subtopics to use them in indexing and retrieval task.

4 Similarity Measure and Evaluation

The similarity between images can be measured by the similarity between their corresponding distributions, Kullback Leibler (KL) divergence can be used to measure the difference (divergence) between two distributions p and q

$$D(p, q) = \sum_{j=1}^T P_j \log_2 \frac{P_j}{q_j}$$

To evaluate the retrieval performance, we use precision-recall pair [6]:

$$\textit{Precision} = \frac{A}{B} \quad \textit{Recall} = \frac{A}{C}$$

A: number of relevant images retrieved.

B: Total number of image retrieved.

C: Total number of relevant images in the database.

5 Conclusion and Current State

In this paper, we propose to use a probabilistic topic model called Pachinko Allocation Model for image indexing and retrieval. We chose to use this model because of success obtained by similar models. The model is applying on local features and also on a set of color and texture global features. We expect to finish testing this proposition soon.

The extraction of local features using DoG/SIFT method and the construction of visual words representation have been achieved, the package of [10] was used for this phase of work. A subset of Corel image database [11] with 900 training images and 100 test images is used for evaluation, we chose to use 6 visual words per image, and a vocabulary of 52 visual words.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
2. Blei, D., Lafferty, J.: Correlated Topic Models. *Advances in Neural Information Processing Systems* 18 (2006)
3. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science* 41(6), 391–407 (1990)
4. Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning* 42(1-2), 177–196 (2001)
5. Horster, E.: Topic Models for Image Retrieval on Large-Scale Databases. University of Augsburg (2009)
6. Jalab, H.A.: Image Retrieval System Based on Color Layout Descriptor and Gabor Filters. In: *IEEE Conference on Open Systems* (2011)
7. LaCascia, M., Sethi, S., Sclaroff, S.: Combining Textual and Visual Cues for Content-Based Image Retrieval on the World Wide Web. In: *IEEE Workshop on Content-Based Access of Image and Video Libraries*, vol. (6) (1998)
8. Li, Y., Wang, W., Gao, W.: Object Recognition Based on Dependent Pachinko Allocation Model. In: *IEEE ICIP*, pp. 337–340 (2007)

9. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
10. Vedaldi, A.: An implementation of SIFT detector and descriptor,
<http://www.vlfeat.org/~vedaldi/code/sift.html>
11. Wang, J.: Corel Image database,
<http://wang.ist.psu.edu/docs/related.shtml>
12. Wei, L., McCallum, A.: Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations. In: *International Conference on Machine Learning*, Pittsburg (2006)

Detection of CAN by Ensemble Classifiers Based on Ripple Down Rules

Andrei Kelarev¹, Richard Dazeley¹, Andrew Stranieri¹,
John Yearwood¹, and Herbert Jelinek²

¹ Centre for Informatics and Applied Optimization
School of SITE, University of Ballarat
P.O. Box 663, Ballarat, Victoria 3353, Australia

{a.kelarev,r.dazeley,a.stranieri,j.yearwood}@ballarat.edu.au

² Centre for Research in Complex Systems and School of Community Health
Charles Sturt University, P.O. Box 789, Albury, NSW 2640, Australia
hjelinek@csu.edu.au

Abstract. It is well known that classification models produced by the Ripple Down Rules are easier to maintain and update. They are compact and can provide an explanation of their reasoning making them easy to understand for medical practitioners. This article is devoted to an empirical investigation and comparison of several ensemble methods based on Ripple Down Rules in a novel application for the detection of cardiovascular autonomic neuropathy (CAN) from an extensive data set collected by the Diabetes Complications Screening Research Initiative at Charles Sturt University. Our experiments included essential ensemble methods, several more recent state-of-the-art techniques, and a novel consensus function based on graph partitioning. The results show that our novel application of Ripple Down Rules in ensemble classifiers for the detection of CAN achieved better performance parameters compared with the outcomes obtained previously in the literature.

1 Introduction

Ripple Down Rules produce models which are easier to maintain and update than other alternatives, [5], [37]. In addition they have the most compact representations of the models, which can be better explained to and understood by medical practitioners, see Section 4 below for more details. The present article deals with an experimental investigation and comparison of several ensemble methods based on Ripple Down Rules in a novel application for the detection of cardiovascular autonomic neuropathy (CAN) in diabetes patients. Our experiments included several essential ensemble methods, a few more recent state-of-the-art techniques, a novel consensus function based on graph partitioning, and used the Diabetes Screening Complications Research Initiative (DiScRi) data set collected at Charles Sturt University, Albury, Australia.

DiScRi is a very large and unique data set containing a comprehensive collection of tests related to CAN. It has been previously considered in [20], where

decision trees were used. Our new results based on Ripple Down Rules and presented in this paper have achieved substantially higher accuracies compared with the previous outcomes in [20].

The paper is organised as follows. Section 2 deals with cardiovascular autonomic neuropathy. Section 3 describes the Diabetes Complications Screening Research Initiative (DiScRi) organised at Charles Sturt University, and the corresponding data set. Section 4 contains brief background information on Ripple Down Rules (RDR). Section 5 describes ensemble methods investigated in this paper. Section 6 presents the experimental results comparing the efficiencies of several ensemble methods based on RDR for this application domain. These outcomes are discussed in Section 7, where the main conclusions are also provided.

2 Cardiovascular Autonomic Neuropathy

Cardiovascular autonomic neuropathy (CAN) is a condition associated with damage to the autonomic nervous system innervating the heart and highly prevalent in people with diabetes, [13], [14], [29]. It is known as one of the causes of mortality among type 2 diabetes patients. The classification of disease progression associated with CAN is important, because it has implications for planning of timely treatment, which can lead to an improved well-being of the patients and a reduction in morbidity and mortality associated with cardiac arrhythmias in diabetes.

The most important tests required for identification of CAN rely on assessing responses in heart rate and blood pressure to various activities, usually consisting of five tests described in [13] and [14]. It is often difficult for clinicians to collect all test data from patients, since they are likely to be suffering from other illnesses affecting their general fitness and making it hard to follow correct procedures for all tests. More details on various other associated tests for CAN are given in the next section.

3 Diabetes Complications Screening Research Initiative

This paper used the data set of test results and health-related parameters collected at the Diabetes Complications Screening Research Initiative, DiScRi, organised at Charles Sturt University, [8], [20], [34]. There are no other alternative data sets containing comparable collections of test outcomes. The collection and analysis of data in the project has been approved by the Ethics in Human Research Committee of the university before investigations started. People participating in the project were attracted via advertisements in the media. The participants were instructed not to smoke and refrain from consuming caffeine containing drinks and alcohol for 24 hours preceding the tests as well as to fast from midnight of the previous day until tests were complete. The measurements were conducted from 9:00am until 12midday and were recorded in the DiScRi data base along with various other health background data including age, sex

and diabetes status, blood pressure (BP), body-mass index (BMI), blood glucose level (BGL), and cholesterol profile. Reported incidents of a heart attack, atrial fibrillation and palpitations were also recorded. The most important set of features recorded for CAN determination is the *Ewing battery* [13], [14]. There are five Ewing tests in the battery: changes in heart rate associated with lying to standing, deep breathing and valsalva manoeuvre and changes in blood pressure associated with hand grip and lying to standing. In addition features from the ten second samples of 12-lead ECG for all participants were extracted from the data base. These included the QRS, PQ, QTc and QTd intervals, heart rate and QRS axis explained below. The QRS complex reflects the depolarization of the ventricles of the heart. The duration of the QRS complex is called the QRS duration. The time from the beginning of the P wave until the start of the next QRS complex is called the PQ interval. The longest distance from the Q wave to the next T wave is called the QT interval. The period from the beginning of the QRS complex to the end of the T wave is denoted by QT interval, which if corrected for heart rate becomes the QTc. It represents the so-called refractory period of the heart. The difference of the maximum QT interval and the minimum QT interval over all 12 leads is known as the QT dispersion denoted by QTd. It is used as an indicator of repolarisation of ventricular. The deflection of the electrical axis of the heart measured in degrees to the right or left is called the QRS axis.

A preprocessing system has been implemented in Python to automate several expert editing rules that can be used to reduce the number of missing values in the database. These rules were collected during several discussions with the experts maintaining the database. Preprocessing of data using these rules produced 1299 complete rows with complete values of all fields, which were used for the experimental evaluation of the performance of data mining algorithms. The whole data base contained over 200 features. We have created several lists ranking the features in the order of their relevance to CAN classification and used them in consultation with the experts maintaining the data base to select the most essential features, [30], [31].

In this paper discussing of the outcomes of our experiments we use the following acronyms for the features contained in the Ewing battery listed in Figure 1. The same acronyms are used in the original DiScRi data base.

Acronym	Feature
LS HR	Lying to standing heart rate change
DB HR	Deep breathing heart rate change
VA HR	Valsalva manoeuvre heart rate change
HG BP	Hand grip blood pressure change
LS BP	Lying to standing blood pressure change

Fig. 1. Acronyms for the Ewing features

First of all we have investigated the set of all Ewing features. Since it is often difficult to collect all tests and there are many missing values in the data base, we also included the largest subsets of four features chosen in the Ewing battery. These subsets can help clinicians in those situations when one of the tests is missing. Figure 2 explains the notation used for the subsets of features analysed in our experiments.

Notation	Features included in the subset
E_{ALL}	All 5 Ewing features: LS HR, DB HR, VA HR, HG BP, LS BP
E_{LSHR}	4 Ewing features with LS HR excluded: DB HR, VA HR, HG BP, LS BP
E_{DBHR}	4 Ewing features with DB HR excluded: LS HR, VA HR, HG BP, LS BP
E_{VAHR}	4 Ewing features with VA HR excluded: LS HR, DB HR, HG BP, LS BP
$E_{HG BP}$	4 Ewing features with HG BP excluded: LS HR, DB HR, VA HR, LS BP
E_{LSBP}	4 Ewing features with LS BP excluded: LS HR, DB HR, VA HR, HG BP

Fig. 2. Notation for subsets of the Ewing battery

4 Induction of Ripple Down Rules

In medical applications of data mining it is especially important to consider the classifiers producing models that can be expressed in a form understandable by medical practitioners. It is well known that Ripple Down Rules create very compact classifiers that can be expressed by a smaller number of decision rules, compared to other systems. This is why in the present paper we investigate several ensemble classifiers based on Ripple Down Rules.

Ripple Down Rules (RDR) were initially introduced by [5] as an approach facilitating the maintenance problem in knowledge based systems, see also [6]. Their applications in various domains have been actively investigated, for instance, in [24], [26], [36] and [41]. The readers are directed to [37] for background information and a survey of the whole area. Multiple Classification Ripple Down Rules (MCRDR) [25] are of particular interest for medical applications, since they are capable of producing multiple conclusions for each instance, which may correspond to several diagnoses for one patient. Other, more general approaches have also been developed recently, see [3], [11] for further details. Let us also refer to [2], [7], [9], [10], [11], [16], [19], [35], [39], [45], [48] for examples of recent contributions.

As a brief introduction required for understanding the output of algorithms considered in this paper, we follow [11] and recall that RDR uses binary trees, where each node contains a rule, a conclusion, a cornerstone case and two branches labelled as the true (or exception) branch and a false branch. During inferencing, if a rule at the current node is found to be true then the true branch is followed and the same can be said of the ‘false’ branch. The process continues until a node is reached where there are no children along the next appropriate branch. The conclusion returned is the one associated with the last successful rule. A number of variants of RDR models, including advanced methods developed recently and indicated above, use more general trees.

The present paper investigates several ensemble methods based on Induction of Ripple Down Rules in their ability to detect CAN in the DiScRi data set. We used an implementation of a RIpPle-DOWn Rule learner available in Weka and known as Ridor. It uses Induction of Ripple Down Rules originating from [18]. A complete RDR model produced by Ridor for all Ewing features on the DiScRi data set contains only 14 rules and is easy to understand for practitioners. For comparison, let us note that for the same data set Alternating Decision Trees produced 31 rules, J48 classifier produced a model with 55 rules, and Decision Table created 315 rules. Thus, for the DiScRi data set we see that Ripple Down Rules create the most compact models. RDR models are also easier to maintain and update using new contributions from the experts in the area.

5 Ensemble Methods

Experimental investigation of various algorithms applied to particular areas is important, since better performance can be achieved in practical applications by using previous research and experience to match algorithms to the application directions and types of problems rather than applying one and the same algorithm to all problems. This is also confirmed by the so-called “no-free-lunch” theorems, which imply that there does not exist one algorithm which is best for all problems, [43]. We have investigated the performance of the following ensemble techniques: Bagging, Boosting, Dagging, Decorate, Grading, HBGF, MultiBoosting and Stacking.

- *Bagging* (bootstrap aggregating), generates a collection of new sets by re-sampling the given training set at random and with replacement. These sets are called *bootstrap samples*. New classifiers are then trained, one for each of these new training sets. They are amalgamated via a majority vote, [4]. Bagging is probably the most widely used ensemble method, see [32].
- *Boosting* trains several classifiers in succession. Each classifier is trained on the instances that have turned out more difficult for the preceding classifier. To this end all instances are assigned weights, and if an instance turns out difficult to classify, then its weight increases. We used highly successful AdaBoost classifier described in [17].
- *AdaBoost of Bagging* is a novel combined ensemble classifier where AdaBoost is used after Bagging based on Ripple-Down Rules has been applied.

- *MultiBoosting* extends the approach of AdaBoost with the wagging technique, [42]. Wagging is a variant of bagging where the weights of training instances generated during boosting are utilized in selection of the bootstrap samples, [1]. Experiments on a large and diverse collection of UCI data sets have demonstrated that MultiBoosting achieves higher accuracy significantly more often than wagging or AdaBoost, [42].
- *Stacking* can be regarded as a generalization of voting, where meta-learner aggregates the outputs of several base classifiers, [44]. (For our data set StackingC produced the same outcomes and so we did not include it as a separate algorithm.)
- *Decorate* is constructing special artificial training examples to build diverse ensembles of classifiers. A comprehensive collection of tests have established that Decorate consistently creates ensembles more accurate than the base classifier, Bagging, Random Forests, which are also more accurate than Boosting on small training sets, and are comparable to Boosting on larger training sets, [33].
- *Dagging* is useful in situations where the base classifiers are slow. It divides the training set into a collection of disjoint (and therefore smaller) stratified samples, trains copies of the same base classifier and averages their outputs using vote, [40]. Our experiments have shown that for CAN data Ridor does not require this type of assistance provided by Dagging.
- *Grading* trains meta-classifiers, which grade the output of base classifiers as correct or wrong labels, and these graded outcomes are then combined, [38].
- *Consensus functions* can be used as a replacement for voting to combine the outputs of several classifiers. Here we used the HBGF consensus function, following the recommendations of [15] and our previous experience with consensus functions and data sets in [12], [47], and [46]. It is based on a bipartite graph with two sets of vertices: clusters and elements of the data set. A cluster C and an element d are connected by an edge in this bipartite graph if and only if d belongs to C . (The weights associated to these edges may have to be chosen as very large constants if the particular graph partitioning algorithm does not allow zero weights and can handle only complete graphs.) An appropriate graph partitioning algorithm is then applied to the whole bipartite graph, and the final clustering is determined by the way it partitions all elements of the data set. We used the HBGF program implemented in C# in the Centre for Informatics and Applied Optimization of the University of Ballarat. It relies on METIS graph partitioning software described in [27].

6 Experimental Results

We used 10-trial 10-fold cross validation to assess the performance of AdaBoost, Bagging, Dagging, Decorate, Grading, HBGF, MultiBoosting and Stacking for 5 subsets of the Ewing features listed in Figure 2. Our experimental results comparing the performance of these ensemble classifiers for the whole set of 5 Ewing features and for its 5 subsets are presented in Tables 2 through to 4.

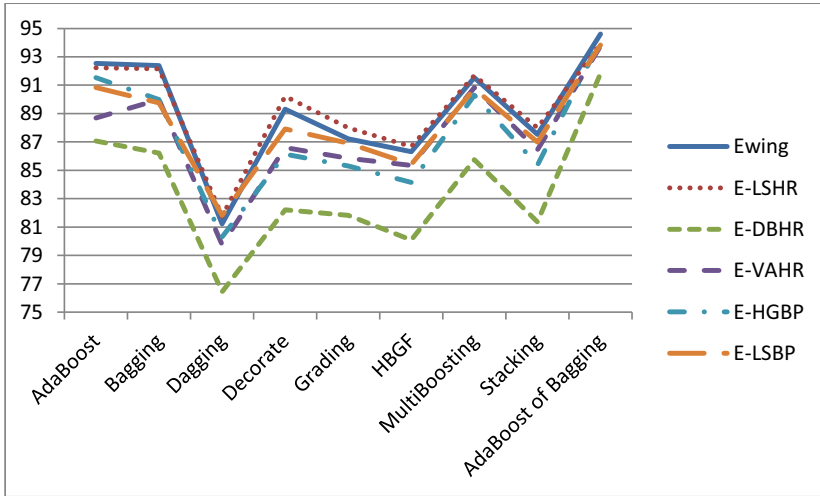


Fig. 3. Accuracies of ensemble classifiers for CAN

These tables include the standard measures of performance: accuracy, precision and recall. Since a number of different associated terms are used in medicine and data mining to discuss these notion, we include a brief overview here.

The accuracy of a classifier is the percentage of all patients classified correctly. It is equal to the probability that a prediction of the classifier for an individual patient is correct. Precision is the ratio of true positives to combined true and false positives. Recall is the ratio of true positives to the number of all positive samples (i.e., to the combined true positives and false negatives). Sensitivity is the proportion of positives (patients with CAN) that are identified correctly. Specificity is the proportion of negatives (patients without CAN) which are identified correctly. Sensitivity is also called True Positive Rate. False Positive Rate is equal to 1 - specificity.

In our tables with outcomes assessing ensemble classifiers, precision and recall refer to their weighted average values. This means that they are calculated for each class separately, and a weighted average is found then.

For the class of patients with CAN, the precision is the ratio of the number of patients correctly identified as having CAN to the number of all patients identified as having CAN. The recall calculated for the class of patients with CAN is equal to sensitivity of the whole classifier. For the cohort of patients without CAN, the precision is the ratio of the number of patients correctly identified as having no CAN to the number of all patients identified as free from CAN. The precision of the classifier as a whole is a weighted average of its precisions for these classes.

Likewise, for the class of patients with CAN, the recall is the ratio of the number of patients correctly identified as having CAN to the number of all

patients with CAN. For the cohort of patients without CAN, the recall is the ratio of the number of patients correctly identified as being free from CAN to the number of all patients without CAN. The recall of the classifier is a weighted average of its recalls for both classes.

For ease of comparison, summary diagrams of the accuracies of all ensemble algorithms for all 5 subsets of Ewing features are given in Figure 3.

To provide complete information of all classifier parameters to the readers, we include Table 1 with command line arguments used in Weka SimpleCLI to run these classifiers in our experiments.

Table 1. SimpleCLI command lines indicating all parameters of ensemble classifiers

Classifier	Command line
AdaBoost	<code>weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0</code>
Bagging	<code>weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0</code>
Dagging	<code>weka.classifiers.meta.Dagging -F 10 -S 1 -W weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0</code>
Decorate	<code>weka.classifiers.meta.Decorate -E 10 -R 1.0 -S 1 -I 10 -W weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0</code>
Grading	<code>weka.classifiers.meta.Grading -X 10 -M "weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0" -S 1 -num-slots 1 -B "weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0"</code>
MultiBoosting	<code>weka.classifiers.meta.MultiBoostAB -C 3 -P 100 -S 1 -I 10 -W weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0</code>
Stacking	<code>weka.classifiers.meta.Stacking -X 10 -M "weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0" -S 1 -num-slots 1 -B "weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0"</code>
AdaBoost of Bagging	<code>weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W weka.classifiers.rules.Ridor -F 3 -S 1 -N 2.0</code>

7 Discussion and Conclusion

It is interesting that, for the Ewing features of the DiScRi data set, bagging and boosting based on Ripple Down Rules have clearly outperformed other ensemble methods; and the best outcomes have been obtained by a novel combined ensemble classifier where AdaBoost is used after Bagging based on Ripple-Down Rules has been applied. Good performance of Adaboost of bagging demonstrates that diversity of the ensemble classifiers used in two levels is crucial for success of the combined multi-level ensemble classifier.

Table 2. Accuracy of ensemble classifiers based on RDR

	Subsets of features					
	E_{ALL}	E_{LSHR}	E_{DBHR}	E_{VAHR}	E_{HGBP}	E_{LSBP}
AdaBoost	92.53	92.22	87.07	88.68	91.53	90.84
Bagging	92.38	92.15	86.22	89.99	89.99	89.76
Dagging	81.22	81.76	76.44	79.75	80.29	81.83
Decorate	89.30	90.22	82.22	86.61	86.14	87.91
Grading	87.22	87.99	81.83	85.84	85.30	86.91
HBGF	86.31	86.67	80.09	85.34	84.15	85.43
MultiBoosting	91.53	91.69	85.76	90.84	90.30	90.76
Stacking	87.53	87.99	81.37	86.45	85.37	86.99
AdaBoost of Bagging	94.61	93.91	91.83	93.76	94.15	93.84

There are several reasons, why other techniques have turned out less effective. First, Dagging uses disjoint stratified training sets to create an ensemble, which benefits mainly classifiers of high complexity. Our outcomes demonstrate that Ripple Down Rules are fast enough and this kind of benefit is not essential. Second, HBGF and other consensus functions can only be used as advanced replacements for majority voting, since they do not discriminate between base classifiers being combined. Finally, stacking and grading use a meta classifier to combine the outcomes of base classifiers. These methods are best applied to combine diverse collections of base classifiers. In this paper we were interested only in those base classifiers, which are easy to maintain and update and which produce models with compact representations that are easier for medical practitioners to interpret, and so we considered only Ripple Down Rules as base classifiers. In this setting stacking clearly performed worse than bagging and boosting.

DiScRi is a very large and unique data set containing a comprehensive collection of tests related to CAN. It has been previously considered in [20], where decision trees were used. Our new results based on Ripple Down Rules and obtained in the present paper have achieved substantially higher accuracies and other performance parameters compared with the previous outcomes in [20]. This improvement is even more significant, because [20] did not use ten-fold cross validation. Overall, the outcomes of the present paper are also appropriate for other areas in general when compared to recent results obtained for other data sets using different methods, for example, in [21], [22], [23], [28] and [47].

The outcomes obtained in this paper also confirm that the sets E_{ALL} and E_{LSHR} are the best subsets of Ewing features to determine CAN. This observation has independent additional significance for clinical practitioners planning and organising tests.

Table 3. Precision of ensemble classifiers based on RDR

	Subsets of features					
	E_{ALL}	E_{LSHR}	E_{DBHR}	E_{VAHR}	E_{HGHP}	E_{LSBP}
AdaBoost	0.926	0.922	0.870	0.886	0.916	0.910
Bagging	0.930	0.924	0.865	0.904	0.905	0.911
Dagging	0.821	0.822	0.765	0.802	0.813	0.828
Decorate	0.898	0.904	0.828	0.868	0.875	0.889
Grading	0.875	0.881	0.822	0.860	0.860	0.879
HBGF	0.864	0.867	0.802	0.854	0.841	0.853
MultiBoosting	0.917	0.918	0.857	0.910	0.908	0.914
Stacking	0.878	0.881	0.818	0.869	0.862	0.880
AdaBoost of Bagging	0.947	0.940	0.919	0.938	0.943	0.940

Table 4. Recall of ensemble classifiers based on RDR

	Subsets of features					
	E_{ALL}	E_{LSHR}	E_{DBHR}	E_{VAHR}	E_{HGHP}	E_{LSBP}
AdaBoost	0.925	0.922	0.871	0.887	0.915	0.908
Bagging	0.924	0.921	0.862	0.900	0.900	0.898
Dagging	0.812	0.818	0.764	0.798	0.803	0.818
Decorate	0.893	0.902	0.822	0.866	0.861	0.879
Grading	0.872	0.880	0.818	0.858	0.853	0.869
HBGF	0.863	0.865	0.799	0.853	0.839	0.851
MultiBoosting	0.915	0.917	0.858	0.908	0.903	0.908
Stacking	0.875	0.880	0.814	0.865	0.854	0.870
AdaBoost of Bagging	0.946	0.939	0.918	0.938	0.941	0.938

As a possible direction for future research, following advice of one of the referees let us mention that it would be also interesting to investigate the performance of expert systems for the detection and classification of CAN created in direct interaction with human experts.

Acknowledgements. Our research on topics of this paper has been supported by several grants from the University of Ballarat.

The authors are grateful to two referees for comments and corrections that have helped us to improve the presentation, and for an indication of a possible direction for future research.

References

1. Bauer, E., Kohavi, R.: An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning* 36, 105–139 (1999)
2. Bindoff, I., Kang, B.H.: Simulated Assessment of Ripple Round Rules. In: Kang, B.-H., Richards, D. (eds.) PKAW 2010. LNCS (LNAI), vol. 6232, pp. 180–194. Springer, Heidelberg (2010)
3. Bindoff, I., Kang, B.H.: Applying Multiple Classification Ripple Round Rules to a Complex Configuration Task. In: Wang, D., Reynolds, M. (eds.) AI 2011. LNCS (LNAI), vol. 7106, pp. 481–490. Springer, Heidelberg (2011)
4. Breiman, L.: Bagging predictors. *Machine Learning* 24, 123–140 (1996)
5. Compton, P., Jansen, R.: Knowledge in Context: a strategy for expert system maintenance. In: Second Australian Joint Artificial Intelligence Conference, AI 1988, pp. 292–306 (1988)
6. Compton, P., Jansen, R.: A philosophical basis for knowledge acquisition. *Knowledge Acquisition* 2, 241–258 (1990)
7. Compton, P., Peters, L., Edwards, G., Lavers, T.: Experience with Ripple-Down Rules. *Knowledge-Based Systems* 19(5), 356–362 (2006)
8. Cornforth, D., Jelinek, H.: Automated classification reveals morphological factors associated with dementia. *Applied Soft Computing* 8, 182–190 (2007)
9. Dazeley, R., Kang, B.: Generalising Symbolic Knowledge in Online Classification and Prediction. In: Richards, D., Kang, B.-H. (eds.) PKAW 2008. LNCS (LNAI), vol. 5465, pp. 91–108. Springer, Heidelberg (2009)
10. Dazeley, R., Park, S., Kang, B.: Online knowledge validation with prudence analysis in a document management application. *Expert Systems with Applications* 38, 10959–10965 (2011)
11. Dazeley, R., Warner, P., Johnson, S., Vamplew, P.: The Ballarat Incremental Knowledge Engine. In: Kang, B.-H., Richards, D. (eds.) PKAW 2010. LNCS (LNAI), vol. 6232, pp. 195–207. Springer, Heidelberg (2010)
12. Dazeley, R., Yearwood, J., Kang, B., Kelarev, A.: Consensus Clustering and Supervised Classification for Profiling Phishing Emails in Internet Commerce Security. In: Kang, B.-H., Richards, D. (eds.) PKAW 2010. LNCS (LNAI), vol. 6232, pp. 235–246. Springer, Heidelberg (2010)
13. Ewing, D., Campbell, J., Clarke, B.: The natural history of diabetic autonomic neuropathy. *Q. J. Med.* 49, 95–100 (1980)
14. Ewing, D., Martyn, C., Young, R., Clarke, B.: The value of cardiovascular autonomic function tests: 10 years experience in diabetes. *Diabetes Care* 8, 491–498 (1985)
15. Fern, X., Brodley, C.: Solving cluster ensemble problems by bipartite graph partitioning. In: 21st International Conference on Machine Learning, ICML 2004, vol. 69, pp. 36–43. ACM, New York (2004)
16. Finlayson, A., Compton, P.: Incremental Knowledge Acquisition Using Generalised RDR for Soccer Simulation. In: Kang, B.-H., Richards, D. (eds.) PKAW 2010. LNCS (LNAI), vol. 6232, pp. 135–149. Springer, Heidelberg (2010)
17. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Proc. 13th Internat. Conf. Machine Learning, pp. 148–156 (1996)
18. Gaines, B., Compton, P.: Induction of Ripple-Down Rules applied to modeling large databases. *J. Intell. Inf. Syst.* 5(3), 211–228 (1995)

19. Ho, V., Compton, P., Benatallah, B., Vayssière, J., Menzel, L., Vogler, H.: An incremental knowledge acquisition method for improving duplicate invoices detection. In: Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, pp. 1415–1418 (2009)
20. Huda, S., Jelinek, H., Ray, B., Stranieri, A., Yearwood, J.: Exploring novel features and decision rules to identify cardiovascular autonomic neuropathy using a hybrid of wrapper-filter based feature selection. In: Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2010, pp. 297–302 (2010)
21. Jelinek, H., Khandoker, A., Palaniswami, M., McDonald, S.: Heart rate variability and QT dispersion in a cohort of diabetes patients. *Computing in Cardiology* 37, 613–616 (2010)
22. Jelinek, H., Rocha, A., Carvalho, T., Goldenstein, S., Wainer, J.: Machine learning and pattern classification in identification of indigenous retinal pathology. In: Proceedings IEEE Conference Eng. Med. Biol. Soc., pp. 5951–5954 (2011)
23. Kang, B., Kelarev, A., Sale, A., Williams, R.: A New Model for Classifying DNA Code Inspired by Neural Networks and FSA. In: Hoffmann, A., Kang, B.-H., Richards, D., Tsumoto, S. (eds.) PKAW 2006. LNCS (LNAI), vol. 4303, pp. 187–198. Springer, Heidelberg (2006)
24. Kang, B., Yoshida, K., Motoda, H., Compton, P.: A help desk system with intelligent interface. *Applied Artificial Intelligence* 11(7-8), 611–631 (1997)
25. Kang, B., Compton, P.: Multiple Classification Ripple Down Rules. In: Third Japanese Knowledge Acquisition for Knowledge-Based Systems Workshop (1994)
26. Kang, B., Gambetta, W., Compton, P.: Verification and validation with ripple-down rules. *International Journal of Human-Computer Studies* 44(2), 257–269 (1996)
27. Karypis, G., Kumar, V.: METIS: A software package for partitioning unstructured graphs, partitioning meshes, and computing fill-reducing orderings of sparse matrices. Technical report, University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Centre, Minneapolis (1998)
28. Kelarev, A., Kang, B., Steane, D.: Clustering Algorithms for ITS Sequence Data with Alignment Metrics. In: Sattar, A., Kang, B.-H. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 1027–1031. Springer, Heidelberg (2006)
29. Khandoker, A., Jelinek, H., Palaniswami, M.: Identifying diabetic patients with cardiac autonomic neuropathy by heart rate complexity analysis. *BioMedical Engineering OnLine* 8 (2009), <http://www.biomedical-engineering-online.com/content/8/1/3>
30. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Boston (1998)
31. Liu, H., Motoda, H., Setiono, R., Zhao, Z.: Feature selection: An ever evolving frontier in data mining. *Journal of Machine Learning Research – Proceedings Track* 10, 4–13 (2010)
32. Mandvikar, A., Liu, H., Motoda, H.: Compact Dual Ensembles for Active Learning. In: Dai, H., Srikant, R., Zhang, C. (eds.) PAKDD 2004. LNCS (LNAI), vol. 3056, pp. 293–297. Springer, Heidelberg (2004)
33. Melville, P., Mooney, R.: Creating diversity in ensembles using artificial data. *Information Fusion* 6, 99–111 (2005)
34. Ng, E., Hambly, B., McLachlan, C., Matthews, S., Jelinek, H.: WEKA machine learning classification in identifying autonomic dysfunction parameters associated with ACE insertion/deletion genotypes. In: Proceedings of the IASTED International Conference Biomedical Engineering, BioMed 2012, pp. 161–166 (2012)

35. Richards, D.: A social software/Web 2.0 approach to collaborative knowledge engineering. *Information Sciences* 179(15), 2515–2523 (2009)
36. Richards, D., Compton, P.: Taking up the situated cognition challenge with ripple down rules. *International Journal of Human-Computer Studies* 49(6), 895–926 (1998)
37. Richards, D.: Two decades of Ripple Down Rules research. *Knowledge Eng. Review* 24(2), 159–184 (2009)
38. Seewald, A.K., Fürnkranz, J.: An Evaluation of Grading Classifiers. In: Hoffmann, F., Adams, N., Fisher, D., Guimarães, G., Hand, D.J. (eds.) *IDA 2001. LNCS*, vol. 2189, pp. 115–124. Springer, Heidelberg (2001)
39. Taylor, M., Richards, D.: Discovering Areas of Expertise from Publication Data. In: Richards, D., Kang, B.-H. (eds.) *PKAW 2008. LNCS (LNAI)*, vol. 5465, pp. 218–230. Springer, Heidelberg (2009)
40. Ting, K., Witten, I.: Stacking bagged and dagged models. In: *Fourteenth International Conference on Machine Learning*, pp. 367–375 (1997)
41. Wada, T., Horiuchi, T., Motoda, H., Washio, T.: A description length-based decision criterion for default knowledge in the ripple down rules method. *Knowledge and Information Systems* 3(2), 146–167 (2001)
42. Webb, G.: Multiboosting: A technique for combining boosting and wagging. *Machine Learning* 40, 159–196 (2000)
43. Wolpert, D.: The lack of a priori distinctions between learning algorithms. *Neural Computation* 8, 1341–1390 (1996)
44. Wolpert, D.: Stacked generalization. *Neural Networks* 5, 241–259 (1992)
45. Xu, H., Hoffmann, A.: RDRCE: Combining Machine Learning and Knowledge Acquisition. In: Kang, B.-H., Richards, D. (eds.) *PKAW 2010. LNCS (LNAI)*, vol. 6232, pp. 165–179. Springer, Heidelberg (2010)
46. Yearwood, J., Webb, D., Ma, L., Vamplew, P., Ofoghi, B., Kelarev, A.: Applying clustering and ensemble clustering approaches to phishing profiling. In: *Data Mining and Analytics 2009, Proc. 8th Australasian Data Mining Conference: AusDM 2009, CRPIT*, vol. 101, pp. 25–34 (2009)
47. Yearwood, J., Kang, B., Kelarev, A.: Experimental investigation of classification algorithms for ITS dataset. In: *Pacific Rim Knowledge Acquisition Workshop, PKAW 2008, Hanoi, Vietnam, December 15-16*, pp. 262–272 (2008)
48. Yoshida, T., Wada, T., Motoda, H., Washio, T.: Adaptive Ripple Down Rules method based on minimum description length principle. *Intell. Data Anal.* 8(3), 239–265 (2004)

Improving Open Information Extraction for Informal Web Documents with Ripple-Down Rules

Myung Hee Kim and Paul Compton

The University of New South Wales, Sydney, NSW, Australia
{mkim978, compton}@cse.unsw.edu.au

Abstract. The World Wide Web contains a massive amount of information in unstructured natural language and obtaining valuable information from informally written Web documents is a major research challenge. One research focus is Open Information Extraction (OIE) aimed at developing relation-independent information extraction. Open Information Extraction systems seek to extract all potential relations from the text rather than extracting a few pre-defined relations. Existing Open Information Extraction systems have mainly focused on Web's heterogeneity rather than the Web's informality. The performance of the REVERB system, a state-of-the-art OIE system, drops dramatically as informality increases in Web documents.

This paper proposes a Hybrid Ripple-Down Rules based Open Information Extraction (Hybrid RDROIE) system, which uses RDR on top of a conventional OIE system. The Hybrid RDROIE system applies RDR's incremental learning technique as an add-on to the state-of-the-art REVERB OIE system to correct the performance degradation of REVERB due to the Web's informality in a domain of interest. With this wrapper approach, the baseline performance is that of the REVERB system with RDR correcting errors in a domain of interest. The Hybrid RDROIE system doubled REVERB's performance in a domain of interest after two hours training.

Keywords: Ripple-Down Rules, Open Information Extraction.

1 Introduction

The Web contains a large amount of information mainly in unstructured text and its quantity keeps increasing exponentially to an almost unlimited size. Web information extraction (WIE) systems analyze unstructured web documents and identify valuable information, such as particular named entities or semantic relations between entities. WIE systems enable effective retrieval of Web information to support various applications such as Automatic Text Summarization (ATS), Information Retrieval (IR) and Question-Answering (QA) systems.

The Web IE task has a number of significant differences compared to the traditional IE task of extracting particular instances from a small range of well-written documents. Most Web documents are not written under strict supervision and tend to be written informally. The followings are some characteristics of Web documents which affect extraction:

Informal Writing Styles. Huge amounts of Web documents are written informally and do not following strict writing styles like journalistic text [1]. Many NER techniques as part of a WIE rely on title and trigger words. As these markers are often absent in Web documents, there can be significant errors.

Spelling Mistakes and Incomplete Sentences. Web documents often include spelling mistakes and incomplete sentences, which hinder the syntactic analysis and cause extraction errors, since most of the existing systems are trained with formal texts with an assumption that the content of texts follows strict writing guidelines.

Large Amount of Newly and Informally Generated Vocabulary. Web documents contain a large number of newly generated unknown words, informal slang and short abbreviations which cannot be found in the formal dictionaries that are often utilized.

Web IE seeks to extract a large number of facts from heterogeneous Web documents while traditional IE has focused on extracting pre-defined relationships from smaller numbers of domain-specific documents. Open IE differs from previous IE in that its goal is to avoid using pre-defined target relations and extraction models for individual target relation. The OIE approach is intended to reduce the amount of time necessary to find the desired information. The open IE paradigm was proposed as ‘preemptive IE’ [2]. TextRunner [3] is an example of Open IE applied to Web IE.

Most OIE systems are developed using Machine Learning (ML) approaches and require a large amount of training data. They use self-supervised learning which generates a labeled training dataset automatically with some heuristics. For example, TextRunner uses an NLP tool to label entities and a parser to identify positive/negative examples with a small set of hand-written heuristic rules. A limit with this approach is that it cannot handle NLP errors since it relies on prior automatic labeling from NLP tools. This seriously affects the system performance as mentioned in [4], for example when a verb is incorrectly tagged as noun. Current OIE systems tend to use well-written journalistic documents as training data, probably to minimize errors from the NLP tools they depend on. It is likely that such training data is not the most appropriate for Web IE.

We have recently demonstrated how we can build an RDR-based OIE system that outperformed a previous machine-learning OIE system, TEXTRUNNER on Web data in a narrow range of interest [5]. Although the RDROIE system has not been tested on data outside the range of interest, necessarily it will perform worse a general OIE system in general domain. Therefore, we suggest that if we build an RDR-based OIE system to correct the errors of a more general system then overall it should produce better results because the minimum performance should be that of the general system performance.

Our contributions are summarized as follows:

- We propose the Hybrid RDROIE system that employs Ripple-Down Rules’ incremental learning technique as an add-on to the state-of-the-art REVERB system in order to handle any performance degradation of REVERB due to the Web’s informality.
- We evaluate the state-of-the-art REVERB system on a Web dataset with a fair level of Web informality and analysed errors that critically degrade performance.

- We demonstrate how the Hybrid RDROIE system handles informally written Web documents and doubles the performance of the REVERB system in a domain of interest after two hours training.

The remainder of this paper is structured as follows. Section 2 presents related work and section 3 presents an error analysis of the REVERB system on Web data. Section 4 explains our Hybrid RDROIE system in detail, section 5 presents the experimental setting and results and section 6 discusses the results and future work.

2 Related Work

2.1 Open Information Extraction

Sekine [6] introduced a new paradigm “On-Demand Information Extraction (ODIE)” which aims to eliminate high customization cost from target domain change. The ODIE system automatically discovers patterns and extracts information on new topics the user is interested in, using pattern discovery, paraphrase discovery, and extended named entity tagging. Shinyama et al. [7] developed the ‘preemptive IE’ framework with the idea of avoiding relation specificity. They clustered documents using pairwise vector-space clustering, and then they re-clustered documents based on named entity types in each document cluster. The system was tested on limited size corpora, because the two clustering steps made it difficult to scale the system for Web IE. TextRunner is the first open IE system for Web IE [3]. Two versions have been developed. The first is called O-NB which treated OIE task as a classification problem using a Naïve Bayes classifier [3]. The more recent system is O-CRF, which treated the task as a sequential labeling problem using ‘Conditional Random Fields (CRF)’ [4]. O-CRF outperforms O-NB almost doubling recall. StatSnowball [8] performs both relation-specific IE and open IE with a bootstrapping technique which iteratively generates weighted extraction patterns. It employs shallow features only such as part-of-speech tags. In StatSnowball, two different pattern selection methods are introduced: l_1 -norm regularized pattern selection and heuristic-based pattern selection. Wu et al. [9] introduced a Wikipedia-based Open Extractor (WOE) which used heuristic matches between Wikipedia infobox attribute values and corresponding sentences in the document for self-supervised learning. WOE applied two types of lexical features: POS tag features and dependency parser features. Although with dependency parser features the system ran more slowly, it outperformed the system with POS tag features. Fader et al. [10] presented the problems of state-of-the-art OIE systems such as the TEXTRUNNER system [4] and the WOE system [9] where system outputs often contain uninformative and incoherent extractions. To address these problems, they proposed two simple syntactic and lexical constraints on binary relations expressed by verbs. Furthermore, the REVERB system proposed by Fader et al. is a ‘relation first’ rather than an ‘arguments first’ system, to try to avoid the errors of previous systems. REVERB achieved an AUC¹ that is 30% higher than WOE_{parse} and more than double the AUC of WOE_{pos} or TEXTRUNNER [10].

¹ Area Under the Curve computed by a precision-recall curve by varying confidence threshold.

2.2 Ripple-Down Rules (RDR)

The basic idea of RDR is that cases are processed by the knowledge based system and when the output is not correct or missing one or more new rules are created to provide the correct output for that case. The knowledge engineering task in adding rules is simply selecting conditions for the rule which is automatically located in the knowledge base with new rules placed under the default rule node for newly seen cases, and exception rules located under the fired rules. The system also stores cornerstone cases, cases that triggered the creation of new rules. If a new rule is fired by any cornerstone cases, the cornerstones are presented to the expert to select further differentiating features for the rule or to accept that the new conclusions should apply to the cornerstone. Experience suggests this whole process takes at most a few minutes. A recent study of a large number of RDR knowledge bases used for interpreting diagnostic data in chemical pathology, showed from logs that the median time to add a rule was less than 2 minutes across 57,626 rules [11].

The RDR approach has also been applied to a range of NLP applications. For example, Pham et al. developed KAFTIE, an incremental knowledge acquisition framework to extract positive attributions from scientific papers [15] and temporal relations that outperformed machine learning [16]. Relevant to the work here, RDR Case Explorer (RDRCE) [17] combined Machine Learning and manual Knowledge Acquisition. It generated an initial RDR tree using transformation-based learning, but then allowed for corrections to be made. They applied RDRCE to POS tagging and achieved a slight improvement over state-of-the-art POS tagging after 60 hours of KA. The idea of using an RDR system as a wrapper around a more general system was suggested by work on detecting duplicate invoices where the RDR system was used to clean up false positive duplicates from the general system [12].

3 Error Analysis of REVERB on a Web Dataset

In this section, we analyse the performance of the REVERB system on a Web dataset, Sent500 and categorise the types of errors. The experiment is conducted on the Web dataset referred to as ‘Sent500’ and the detail of it is explained in section 5.1. Originally, in this dataset, each sentence has one pair of entities manually identified for the relation extraction task, but tags are removed for this evaluation. That is, there are no pre-defined tags in the Sent500 dataset used here.

Extractions are judged by the following: Entities should be proper nouns; pronouns such as he/she/it etc. are not treated as appropriate entities. In a tuple, entity1, relation and entity2 should be located in the appropriate section. For example, if entity1 and relation are both in entity1 section and the relation section is filled by noise then, it is treated as an incorrect extraction. On the other hand, if entity1, relation and entity2 are properly located, then some extra tokens or noise are allowed as long as they do not affect the meaning of extraction. For example, the tuple extraction (**Another example of a statutory merger , is , software maker Adobe Systems acquisition of Macromedia**) is incorrect but the tuple extraction (**Adobe , has announced the acquisition of , Macromedia**) is correct. N-ary relations such as (**Google , has officially acquired YouTube for , \$ 1.65 bil**) are treated as a correct extraction.

Table 1. The performance of the REVERB system on the Sent500

	Total	VERB	NOUN+PREP	VERB+PREP	INFINITIVE
P	41.32%	69.72%	42.03%	69.86%	50.00%
R	45.25%	55.62%	26.13%	54.26%	20.45%
F1	43.20%	61.88%	32.22%	61.08%	29.03%

Table 1 shows the performance of the REVERB system overall and on four different classes. The overall result is evaluated based on all extractions from Sent500 using REVERB, while the four category results are evaluated based on extraction of the pre-tagged entities and relations in Sent500. The results show that overall REVERB performance on Sent500 is quite poor at around 40%. The VERB and VERB+PREP categories show higher precision than the NOUN+PREP and INFINITIVE categories. Especially, the recall of NOUN+PREP and INFINITIVE categories is very low, 26.13% and 20.45%, respectively. This is because that the REVERB system aims to extraction binary relations expressed by verbs.

Table 2. Incorrect extraction errors analysis on each category

	VERB	NOUN+PREP	VERB+PREP	INFINITIVE
Correct relation but incorrect entities	84%	18%	91%	33%
Correct relation and entities but incorrect position with noise	4%	27%	0%	0%
Incorrect relation and entities	12%	55%	9%	67%

Table 2 summaries the types of incorrect extraction errors on four categories. For VERB and VERB+PREP categories, most of false positive errors, 84% and 91% respectively, are due to incorrect entity detection while relation detection is correct. As REVERB extracts entities using noun phrases, which are located nearest to the detected relation, it often recognizes an inappropriate noun phrase as an entity.

For example, in a sentence '*Google has acquired the video sharing website YouTube for \$ 1.65billion (883million) in shares after a large amount of speculation over whether __ was talking about a deal with __ .*', (**Google, has acquired, the Video**) is extracted instead of (**Google, has acquired, YouTube**).

Some of entities have boundary detection errors due to noise or symbols used within an entity. For instance, REVERB only extracted 'Lee' for an entity 'Tim Berners – Lee'. On the other hand, in the NOUN+PREP and INFINITIVE categories, most of false positive errors, 55% and 67% respectively, are due to incorrect detection of both relations and entities.

Table 3. Missed extraction errors analysis on each category

	VERB	NOUN+PREP	VERB+PREP	INFINITIVE
NLP error	72%	7%	14%	0%
Non-verb-based relation	11%	93%	5%	100%
Noise	11%	0%	0%	0%
Unusual expression	6%	0%	81%	0%

Table 3 presents the types of missed extraction errors on the four categories. In the VERB category, 72% of errors are caused by NLP errors. For example, in ‘*Google Buys YouTube.*’, REVERB misses an extraction because ‘Buys’ is tagged as a noun.

Due to the Web’s informality such as informally used capital letters, NLP tools often incorrectly annotate Web datasets. In the VERB+PREP category, 81% of the errors are due to unusual expressions. REVERB includes approximately 1.7 million distinct normalized relation phrases, which are derived from 500 million Web sentences. As REVERB uses this set of relation phrases to detect relations, it tends to miss relations not expressed in the system. For example, in the sentence ‘*Kafka born in Prague*’, the relation ‘born in’ is not detected while in the sentence ‘*Kafka was born in Prague*’, the relation ‘was born in’ is correctly detected. Moreover, in the sentence ‘*Google acquire YouTube*’, the relation ‘acquire’ is not detected while in the sentence ‘*Google acquires YouTube*’, the relation ‘acquires’ is correctly detected.

In the NOUN+PREP and INFINITIVE categories errors are mostly due to non-verb-based relation extraction. As REVERB aims to extract binary relations expressed by verbs, it only can extract NOUN+PREP and INFINITIVE type relations when there is verb before a NOUN+PREP and INFINITIVE relation phrase. That is, when there exist tuples like (entity1, verb NOUN+PREP, entity2) and (entity1, verb TO VB, entity2), REVERB can extract NOUN+PREP and INFINITIVE type relations in the Sent500 dataset. For example, a tuple (**Novartis** , **completes acquisition of 98% of** , **Eon Labs**) is successfully extracted from the sentence ‘*Novartis completes acquisition of 98 % of Eon Labs , substantially strengthening the leading position of its Sandoz generics unit (Basel , July 21 , 2005)*’ while no tuple is extracted from the sentence ‘*Here is the video of the two _founders talking about the Google acquisition in their YouTube Way !*’ because there is no verb between two entities ‘Google’ and ‘YouTube’. In the INFINITIVE case, for instance, a tuple (**Paramount Pictures** , **agreed to buy** , **DreamWorks SKG**) is correctly extracted from the sentence ‘*Viacom s Paramount Pictures agreed to buy DreamWorks SKG for \$ 1.6 billion in cash and debt , wresting the movie studio away from NBC Universal and securing the talents of Steven Spielberg .*’, while no tuple is extracted from ‘*Adobe About to Buy Macromedia .*’ because there is no verb between ‘Adobe’ and ‘Macromedia’.

The REVERB system has shown very poor recall on Sent500. 89% and 95% of the false negative errors (which affects recall) on VERB and VERB+PERP are due to the Web’s informality (NLP error, noise and unusual expression). Also, 93% and 100% of false negative errors on NOUN+PERP and INFINITIVE are due to non-verb relations. The aim of Hybrid RDROIE is to correct REVERB’s.

4 Hybrid RDROIE System Architecture

The Hybrid RDR-based Open Information Extraction (Hybrid RDROIE) system shown in Fig. 1 consists of four main components: preprocessor, NLPRDR KB learner, REVERB system and TupleRDR KB learner. We considered that it was more efficient to clean up NLP errors before using REVERB rather than just fixing errors after. This is because as shown above, REVERB's recall is very poor and one of main reasons is NLP error. If we use REVERB first before NLPRDR KB, then we cannot improve REVERB's recall. In section 4.1, the implementation details of the three components are explained; the RDR rule syntax is described in section 4.2 and RDR KB construction demonstrated in section 4.3 and finally the user interface is shown in section 4.4.

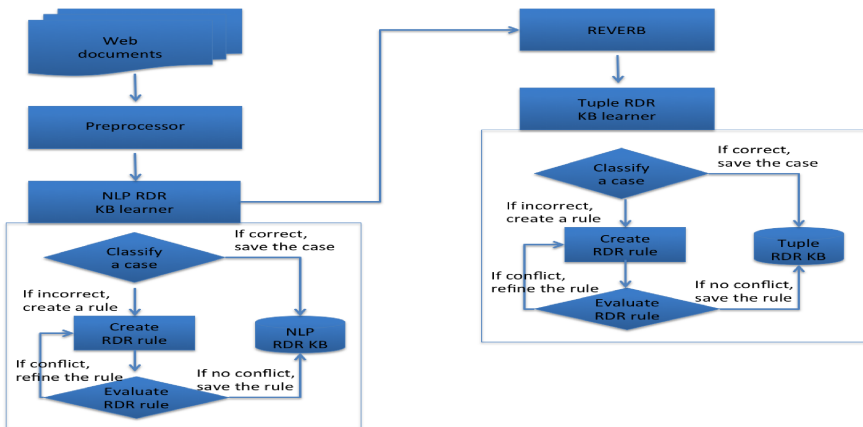


Fig. 1. Architecture of the Hybrid RDROIE system

4.1 Implementation

Preprocessor. The preprocessor converts raw Web documents into a sequence of sentences, and annotates each token for part of speech (POS) and noun and verb phrase chunk using the OpenNLP system. It also annotates named entity (NE) tags using the Stanford NER system. Annotated NLP features are used when creating rules.

NLPRDR KB Learner. The NLPRDR KB is built incrementally while the Hybrid RDROIE system is in use. The system takes a preprocessed sentence as a case and the NLPRDR KB returns the NLP classification result. When the NLP classification result is not correct, the user adds exception rules to correct it. There are three steps:

Step1: NLP Classification. The NLPRDR KB takes each preprocessed sentence from the preprocessor and returns the classification results. If RDR rules are fired and the fired rules deliver correct the classification results, then the system saves the case (a sentence) under the fired rules. The system also saves the refined sentence based on

the fired rule's conclusion action and sets the current case sentence as the refined sentence and passes it to the REVERB system for tuple extraction. If the root rule is fired and the sentence is correct, then the current case sentence is kept as is.

Step2: Create RDR Rule. Whenever the NLPRDR KB gives incorrect classification results, the user adds rules to correct the classification results.

Step3: Evaluate and Refine RDR Rule. Once the new rule is created, the system automatically checks whether the new rule affects KB consistency by evaluating all the previously stored cornerstone cases that may fire the new rule. To assist the expert, the user interface displays not only the rule conditions of previously stored cases but also the features differentiating the current case and any previously stored cases, which also satisfy the new rule condition but have a different conclusion. The expert must select at least one differentiating feature, unless they decide that the new conclusion should apply to the previous case.

As the NLPRDR KB corrects NLP errors on the sentence, more tuples can be extracted from the REVERB system.

TupleRDR KB Learner. The TupleRDR KB is used to correct errors on REVERB's tuple extractions, whereas the NLPRDR KB described above was used to tidy up NLP errors on the given sentence before using REVERB.

The TupleRDR KB is built incrementally while the system is in use. In the Hybrid RDROIE system, the user gets the tuple extractions in the form of binary relation (entity1, relation, entity2) from the REVERB system. The TupleRDR KB returns the tuple classification result and if the tuple classification result is incorrect, the user adds exception rules to correct it. There are following three steps:

Step1: Tuple Classification. The TupleRDR KB takes each tuple extraction from the REVERB system and returns the classification results. If the RDR rules fire and the fired rules deliver the correct classification results, then the system saves the case (a tuple extraction) under the fired rules and also saves the corrected tuple based on the fired rules' conclusion action. If the root rule is fired and the tuple is correct, then only action is to save the correct extraction in the database.

Step2: Create RDR Rule. Whenever incorrect classifications results are given (by the REVERB system or the TupleRDR KB add-on), the user adds rules to correct the classification results.

Step3: Evaluate and Refine RDR Rule. Same as step3 in NLPRDR KB.

4.2 Hybrid RDROIE's Rule Description

An RDR rule has a condition part and a conclusion part: 'IF (*condition*) THEN (*conclusion*)' where *condition* may indicate more than one condition. A *condition* consists of three components: (ATTRIBUTE, OPERATOR, VALUE). In the NLPRDR KB, the ATTRIBUTE refers to the given sentence and in the TupleRDR KB, ATTRIBUTE refers to the given sentence and each element of the given tuple, ENTITY1, RELATION and ENTITY2. Both the NLPRDR KB and the TupleRDR KB provide 9 types of OPERATOR as follows:

- hasToken: whether a certain token matches
- hasPOS: whether a certain part of speech matches
- hasChunk: whether a certain chunk matches
- hasNE: whether a certain named entity matches
- hasGap: skip a certain number of tokens or spaces to match the pattern
- notHasPOS: whether a certain part of speech does not match
- notHasNE: whether a certain named entity does not match
- beforeWD(+a): checks tokens located before the given attribute token by +a
- afterWD(+a): checks tokens located after the given attribute token by +a

VALUE is derived automatically from the given sentence corresponding to the ATTRIBUTE and OPERATOR chosen by the user in the user interface.

In both NLPRDR KB and TupleRDR KB, conditions are connected with an ‘AND’ operator. A sequence of conditions begins with the ‘SEQ’ keyword and it is used to identify a group of words in sequence order, so patterns can be detected. For instance, the sequence condition: ‘SEQ((RELATION hasToken ‘born’) & (RELATION hasToken ‘in’))’ detects ‘born in’ in the RELATION element of the tuple.

In NLPRDR KB, a rule’s CONCLUSION part has the following form:

```
(fixTarget,    --- target element
fixType,      --- refinement type, default is token
fixFrom,      --- classification result before refinement
fixTo)        --- classification result after refinement
```

In TupleRDR KB, a rule’s CONCLUSION part has the following form:

```
(relDetection, --- relation existence detection
fixTarget,     --- target element
fixFrom,       --- classification result before refinement
fixTo)         --- classification result after refinement
```

4.3 Examples of Hybrid RDROIE Rules

The Hybrid RDROIE system is based on Multiple Classification RDR (MCRDR) [13]. Fig. 2 demonstrates MCRDR-based KB construction as the NLPRDR KB system processes the following three cases starting with an empty KB (with a default rule R1 which is always true and returns the NULL classification).

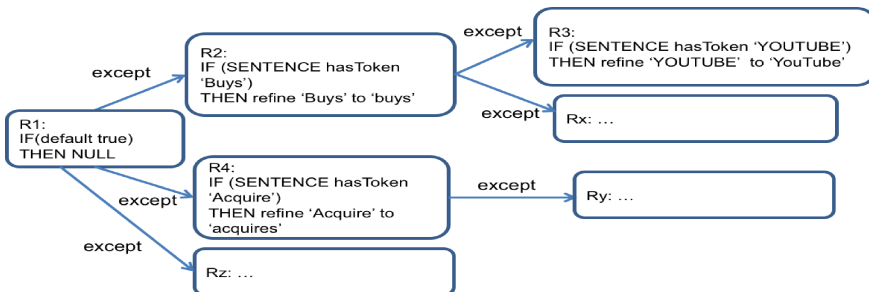


Fig. 2. MCRDR structure of the NLPRDR KB system

Case1: A sentence ‘*Google Buys YouTube.*’

→ The default rule R1 is fired and the KB system returns a NULL classification and the user considers this is an incorrect classification result because ‘Buys’ should be refined as ‘buys’.

→ A user adds a new rule R2 under the default rule R1.

Case2: A sentence ‘*Google Buys YOUTUBE.*’

→ Rule R2 is fired but the user considers the result is incorrect because ‘YOUTUBE’ should be refined to ‘YouTube’ to be correctly tagged by NLP tools and extracted by REVERB.

→ A user adds an exception rule R3 under the parent rule R2.

Case3: A sentence ‘*Adobe system Acquire Macromedia.*’

→ Default rule R1 fires and the KB system returns a NULL classification, which the user considers as an incorrect result because ‘Acquire’ should be refined to ‘acquires’ to be extracted correctly by the REVERB system.

→ A user adds a new rule R4 under the default rule R1.

Fig. 3 shows an MCRDR based KB construction as the TupleRDR KB system processes the following three cases (tuples) starting with an empty KB.

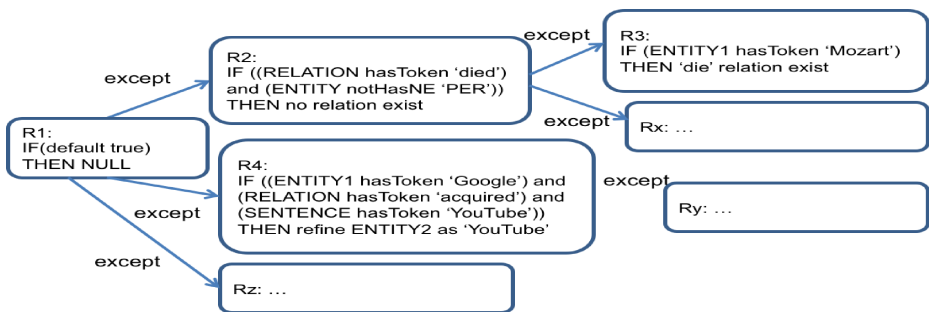


Fig. 3. MCRDR Structure of the TupleRDR KB System

Case1: Tuple (**Prague July 3, 1883** , **died near** , **Vienna June 3, 1924**) from the given sentence ‘*Franz Kafka (born Prague July , 1883 died near Vienna June 3, 1924) was a famous Czech - born , German - speaking writer .*’

→ The default rule R1 is fired and the KB system returns a NULL classification, which the user considers as an incorrect classification result because ENTITY1 contains ‘Prague July 3, 1883’ instead of ‘Franz Kafka’.

→ A user adds a new rule R2 under the default rule R1.

Case2: Tuple (**Wolfgang Amadeus Mozart** , **died** , **5 December 1791**) from the given sentence ‘*Wolfgang Amadeus Mozart died 5 December 1791.*’

→ Rule R2 fires and classifies the given tuple as ‘no relation tuple’ and the user considers it as an incorrect result because the tuple contains a correct relation. This happens since the NE tagger has not tagged the token ‘Mozart’ as PERSON NE.

→ The user adds an exception rule R3 under the parent rule R2.

Case3: Tuple (**Google** , **has acquired** , **the Video**) from the given sentence ‘*Google has acquired the Video sharing website YouTube for \$ 1.65billion (883million).*’

→ The default rule R1 fires and the KB system returns a NULL classification, which the user considers as an incorrect result because ENTITY2 contains ‘the Video’ instead of ‘YouTube’.

→ A user adds new rule R4 under the default rule R1.

4.4 Hybrid RDROIE User Interface

The Hybrid RDROIE system provides a graphic interface that aids in creating and adding RDR rules and maintaining the KB system by end-users. Because most of the relevant values are displayed automatically and the system is built based on the normal human process of identifying distinguishing features when justifying a different conclusion, a user should be able to manage the system after few hours training. Industrial experience in complex domains supports this [11]. Fig. 4 presents the Hybrid RDROIE user interface. The Hybrid RDROIE system is written in Java (Java 1.6) and adopted the OpenNLP system (version 1.5), the Stanford NER system (version 1.5) and the REVERB OIE system (version 1.1).

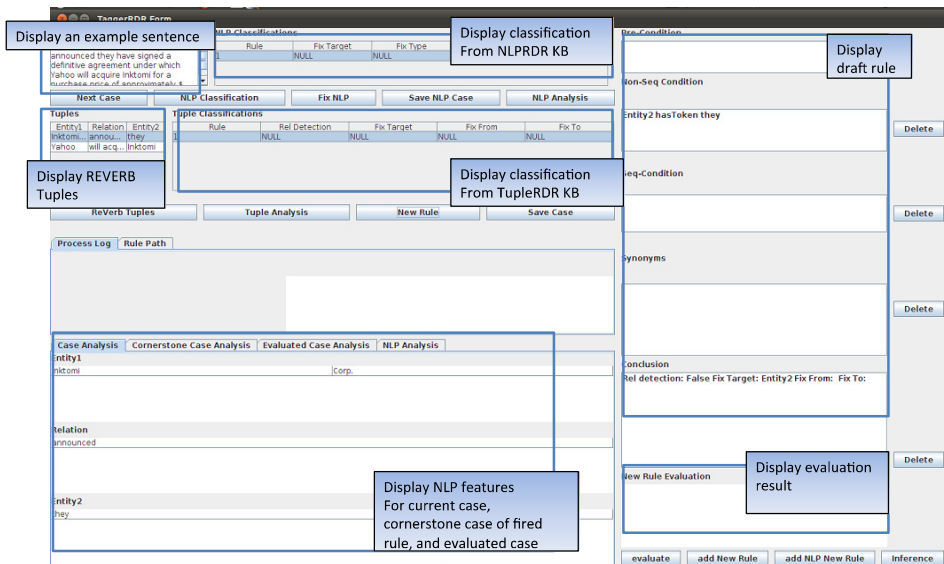


Fig. 4. User Interface of the Hybrid RDROIE system

5 Experiments

Section 5.1 describes the Web dataset used. Section 5.2 shows the initial knowledge base construction of the Hybrid RDROIE system and the section 5.3 presents the

results achieved by the Hybrid RDROIE system and discusses how our system improved the existing performance of the REVERB system on the Web data

5.1 Web Datasets

The experiments were conducted on the two Web datasets, Sent500 and Sent300, which were also used in experiments for the RDROIE system [5]. Sent300 was used as training dataset to construct the RDR KB and Sent500 was used as test dataset to test the performance of the overall Hybrid RDROIE system. Sent300 is derived from the MIL dataset developed by Bunescu and Mooney [14]. The MIL dataset contains a bag of sentences from the Google search engine by submitting a query string ‘a1 ***** a2’ containing seven wildcard symbols between the given pair of arguments. Sent500 was developed by Banko and Etzioni [4]. It contains some randomly selected sentences from the MIL dataset and some more sentences for ‘inventors of product’ and ‘award winners’ relations using the same technique as used for MIL datasets. In Sent300 and Sent500, each sentence has one pair of entities manually tagged for the relation extraction task, but those entity tags were removed in this experiment. That is, there are no pre-defined tags in our training and test dataset.

5.2 RDR Initial KB Constructions

This section presents the analysis of the initial KB construction using the Hybrid RDROIE system. In processing the Sent300 training set, 119 NLP errors were identified and rules were added as each error occurred. For the NLPRDR KB, 28 new rules were added under the default rule R1 and 6 exception rules were added for the cases which received incorrect classification results from earlier rules. Secondly, 98 tuples extracted from the REVERB system, which could not be corrected by fixing NLP errors with the NLPRDR KB were identified as incorrect relation extractions and rules were added for each incorrect tuple extraction. For the TupleRDR KB, in total, 14 new rules were added under the default rule R1 and 5 exception rules were added for the cases which received incorrect classification results from earlier rules.

As the Hybrid RDROIE system handles for both NLP error and tuple error, all rules are used together within a single process flow. In total 53 rules were added within two hours. KB construction time covers from when a case is called up until a rule is accepted as complete. This time is logged automatically.

5.3 Hybrid RDROIE Performance

The Hybrid RDROIE system was tested on the Sent500 dataset. Table 4 presents the performance of the Hybrid RDROIE system on total extractions and on four category extractions. The REVERB system extracts multiple tuples from a sentence without using pre-defined entity tags. The performance on total extractions is evaluated on all tuple extractions of the REVERB system. The performance on four extraction types is calculated based on the explicit tuples when the pre-defined entity tags exist.

Table 4. The performance of the Hybrid RDROIE system on total extraction and on four categories of extraction on the Sent500 dataset

	Total	VERB	NOUN+PREP	VERB+PREP	INFINITIVE
P	90.00%	90.24%	74.00%	90.00%	85.00%
R	81.45%	83.15%	66.67%	86.17%	77.27%
F1	85.51%	86.55%	70.14%	88.04%	80.95%

On total extractions, overall the Hybrid RDROIE system achieved 90% precision, 81.45% recall and an F1 score of 85.51%, while the REVERB system by itself achieved 41.32% precision, 45.25% recall and a 43.20% F1 score on the same dataset (see table 1). That is, the Hybrid RDROIE system improved the performance of the REVERB system by almost double. Precision improved as the TupleRDR KB reduced false positive errors by filtering incorrect extractions and recall improved as the NLPDR KB reduced false negative errors by amending informal sentences.

Across the four category extractions, on average the Hybrid RDROIE system improved around 30% on precision, recall and F1 score over all four categories. VERB and VERB+PREP categories achieved high precision and NOUN+PREP and INFINITIVE categories also achieved reasonably good precision. In particular the recall of NOUN+PREP and INFINITIVE categories improved dramatically from 26.13% and 20.45% to 66.67% and 77.27%, respectively. This improvement suggests that the Hybrid RDROIE system supports relation extractions on non-verb expression while the REVERB system mainly extracts relation expressed by verbs.

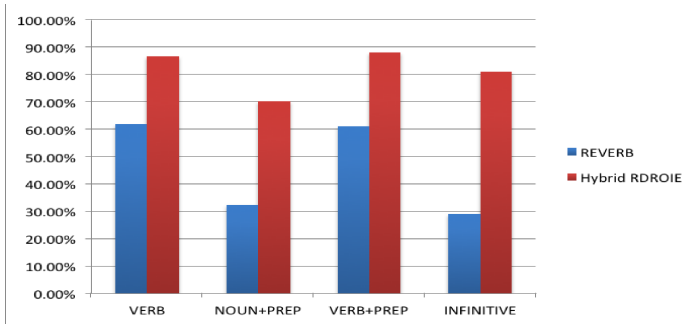
**Fig. 5.** Performance improvement of the Hybrid RDROIE system from the REVERB system on F1 score over four categories

Fig. 5 presents the performance improvement of the Hybrid RDROIE system over the REVERB system on F1 score over four categories. For all categories the Hybrid RDROIE system improved REVERB performance. In particular, NOUN+PREP and INFINITIVE category had the biggest improvement.

6 Discussion

Table 4 shows that the Hybrid RDROIE system achieves high precision and recall after only two hours initial KB construction by a user on a small training dataset. Given the very rapid training time we suggest that rather than simply having to accept an insufficient level of performance delivered by the REVERB system in a particular application area, it is a worthwhile and practical alternative to very rapidly add rules to specifically cover the REVERB system's performance drop in that application area.

In the Hybrid RDROIE system, lexical features are mainly utilized when creating RDR rules. Because it is difficult to handle the Web's informality using NLP features such as part-of-speech, chunk phrase and named entity most of errors for the REVERB system occurred because of NLP errors on the Web dataset.

The advantage of utilising lexical features directly was demonstrated by the REVERB system's performance compared to previous OIE systems such as the TEXTRUNNER and WOE systems [4, 9], but we note that this was for data without a high level of informality. The REVERB system primarily utilises direct lexical feature matching techniques using the relation phrase dictionary, collected from 500 million Web sentences. Previous OIE systems such as the TEXTRUNNER system and the WOE system utilised more NLP features such as part-of-speech and chunk phrase and used machine learning techniques on a large volume of heuristically labelled training data (e.g. 200,000 sentences used for TEXTRUNNER and 300,000 sentences for WOE). The Hybrid RDROIE system, similarly utilises lexical features to handle the Web's informality and improve the REVERB system's performance further. In consequence, as shown in table 4, the Hybrid RDROIE system outperformed the REVERB system and achieved a good balanced overall result compared to other OIE systems. Section 4.3 showed examples of the Hybrid RDROIE system using lexical features in rule creation. We note that other systems focusing more on NLP issues outperformed REVERB on this data set, but we also note that as shown in [5] a pure RDR approach did even better.

The Hybrid RDROIE system is designed to be trained on a specific domain of interest. One might also comment that the rules added are simple fixes of lexical errors, and to produce a large system would need a large number of rules. This is really the same type of approach as REVERB with its vast relation phrase dictionary. If the Hybrid RDROIE system is to be used for a particular domain, which we believe would be the normal real world application, we see little problem in adding the rules required and keeping on doing this as new errors are identified and we note that in pathology people have developed systems with over 10,000 rules [11]. The Hybrid RDROIE system required very little effort and the study here it took about two minutes on average to build a rule. Experience suggests that knowledge acquisition with RDR remains very rapid even for large knowledge bases [11]. On the other hand, if the aim was a very broad system, it would also be interesting to see if it was possible to extend domain coverage by some type of crowd sourcing, with large numbers of people on the web contributing rules.

References

1. Collot, M., Belmore, N.: Electronic Language: A New Variety of English. In: *Computer-Mediated Communications: Linguistic, Social and Cross-Cultural Perspectives* (1996)
2. Shinyama, Y., Sekine, S.: Preemptive information extraction using unrestricted relation discovery. In: *Proceedings of the HLT/NAACL* (2006)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (2007)
4. Banko, M., Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. Paper Presented at the *Proceedings of ACL 2008: HLT* (2008)
5. Kim, M.H., Compton, P., Kim, Y.-s.: RDR-based Open IE for the Web Document. In: *6th International Conference on Knowledge Capture, Banff, Alberta, Canada* (2011)
6. Sekine, S.: On-demand information extraction. In: *Proceedings of the COLING/ACL* (2006)
7. Shinyama, Y., Sekine, S.: Preemptive information extraction using unrestricted relation discovery. In: *Proceedings of the HLT/NAACL* (2006)
8. Zhu, J., Nie, Z., Liu, X., Zhang, B., Wen, J.-R.: StatSnowball: a statistical approach to extracting entity relationships. In: *Proceedings of the 18th WWW* (2009)
9. Wu, F., Weld, D.S.: Open Information Extraction using Wikipedia. In: *The 48th Annual Meeting of the Association for Computational Linguistics, Uppsala, Sweden* (2010)
10. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: *EMNLP, Scotland, UK* (2011)
11. Compton, P., Peters, L., Lavers, T., Kim, Y.-S.: Experience with long-term knowledge acquisition. In: *6th International Conference on Knowledge Capture*, pp. 49–56. ACM, Banff (2011)
12. Ho, V.H., Compton, P., Benatallah, B., Vayssiere, J., Menzel, L., Vogler, H.: An incremental knowledge acquisition method for improving duplicate invoices detection. In: *Proceedings of the International Conference on Data Engineering* (2009)
13. Kang, B., Compton, P., Preston, P.: Multiple classification ripple down rules: evaluation and possibilities. In: *Proceedings of the 9th Banff Knowledge Acquisition for Knowledge Based Systems Workshop, Banff, February 26-March 3, vol. 1, pp. 17.1 – 17.20* (1995)
14. Bunescu, R.C., Mooney, R.J.: Learning to Extract Relations from the Web using Minimal Supervision. In: *Proceedings of the 45th ACL* (2007)
15. Pham, S.B., Hoffmann, A.: Extracting Positive Attributions from Scientific Papers. In: *Discovery Science Conference* (2004)
16. Pham, S.B., Hoffmann, A.: Efficient Knowledge Acquisition for Extracting Temporal Relations. In: *17th European Conference on Artificial Intelligence, Italy* (2006)
17. Xu, H., Hoffmann, A.: RDRCE: Combining Machine Learning and Knowledge Acquisition. In: Kang, B.-H., Richards, D. (eds.) *PKAW 2010. LNCS, vol. 6232*, pp. 165–179. Springer, Heidelberg (2010)

Ripple-Down Rules with Censored Production Rules

Yang Sok Kim¹, Paul Compton¹, and Byeong Ho Kang²

¹ University of New South Wales, Sydney,
New South Wales, Australia

{yskim,compton}c@cse.unsw.edu.au

² University of Tasmania, Sandy Bay,
Tasmania, Australia

bhkang@utas.edu.au

Abstract. Ripple-Down Rules (RDR) has been successfully used to implement incremental knowledge acquisition systems. Its success largely depends on the organisation of rules, and less attention has been paid to its knowledge representation scheme. Most RDR used standard production rules and exception rules. With sequential processing, RDR acquires exception rules for a particular rule only after the rule wrongly classifies cases. We propose censored production rules (CPR), to be used for acquiring exceptions when a new rule is created using censor conditions. This approach is useful when we have a large number of validation cases at hand. We discuss inference and knowledge acquisition algorithms and related issues. The approach can be combined with machine learning techniques to acquire censor conditions.

1 Introduction

Ripple-Down Rules (RDR) has been successfully applied in many practical knowledge-based system developments for the last 20 years [1]. It has been extended to cover a range of problems, such as single classification (SCRDR) [2], multiple classifications (MCRDR)[3], configuration[4], etc. It is notable that the success of RDR largely depends on its distinctive operational semantics on standard production rules (SPR), rather than representational differences.

Conventional production systems, such as OPS5[5], use an **unordered** collection of standard production rules. Each rule has a basic form,

IF [condition] THEN [action].

The data operated on by the system is held in a global database called the **working memory**. Unordered production systems operate as follows:

1. Match. Evaluate the [condition] part of the production rules to determine whether they satisfy the current data of working memory.

2. Conflict resolution. Select one production rule which satisfied the [condition] part based on conflict resolution strategy;
3. Act. Perform [actions] of the selected production rule.

This sequence is repeated until no further rules can be matched, or an explicit end state is reached through an action [6].

This approach is based on the claim that declarative knowledge should be separated from procedural knowledge. Bench-Capon [7] described the requirements of ‘declarativeness’ as follows: *“If our aim is to separate the knowledge represented in a program from the techniques used to manipulate this knowledge, for we can make no assumptions as to what computation will have taken place before a particular statement in the knowledge base is used. Ideally then, we would want our representation to be referentially transparent, and for the meaning of statements in the representation to be independent of other statements in the knowledge base and of the means of manipulating the knowledgebase.”* It was claimed that well written production rules may improve knowledge maintenance and understanding [8]. Furthermore, this enables domain experts and researchers or engineers to put their expertises into the tasks they might perform better. That is, the researchers or engineers focus on the development of the inference engines, and the domain expert should only be concerned with declarative knowledge [9].

However, Bench-Capon conceded that this principle is usually compromised in practice. *“In practice, this ideal of declarativeness must often be compromised, but declarativeness is a property to which the representation should aspire.”* [7] Experience in expert system developments reveal that knowledge acquisition and maintenance are still significant problems even though production rules were created by following the knowledge separation principle [10]. In addition, experience shows pattern matching is not efficient in conventional production systems. To improve pattern matching efficiency, research such as RETE [6], implicitly creates orders between production rules.

RDR took the opposite direction from conventional production systems by making explicit use of an **ordered** collection of standard production rules. An RDR system executes production rules without re-evaluation of rule ordering or without using conflict resolution strategy, since the rule execution order is already determined when the rules are acquired. Ordered production systems operate as follows:

1. Match. Evaluate [condition] of the production rules from the root node to determine whether or not they satisfy the given current data in the working memory. If a fired production rule specifies the next production rule to be evaluated, the inference engine evaluates it; otherwise stop the inference process.
2. Act. Perform [actions] of the selected production rule.

In particular, RDR has a special rule structure to resolve knowledge maintenance problems. RDR has an explicit structure - an assertion rule + exception rule/rules. Production rules in an RDR knowledge base have the following structure:

[Rule 1] Assert Rule: IF [condition] THEN [action] [CC 1]

[Rule 1.1] Exception Rule: IF [condition] THEN [action] [CC 2]
 [Rule 1.1.1] Exception Rule: IF [condition] THEN [action] [CC 3]
 [Rule 1.2] Exception Rule: IF [condition] THEN [action] [CC 4]

...

Exceptions are used in different ways, depending on the chosen RDR approach. With multiple classification RDR a rule will fire as long as none of its exception rules fires, but with single classification a rule will fire as long as none of its exception rules fires and as long as no older sibling rule fires.

Even though RDR took a significantly different approach for organising production rules, and was obviously successful in incremental knowledge base system developments, it does not differentiate its production rule representation from the conventional representation. RDR systems in general process cases sequentially and new rules are progressively added whenever the current knowledge base suggests wrong conclusions. Whenever a new rule is created, it is necessary to validate the rule normally by checking whether or not the future cases are given the correct classifications.

The problem changes when a large volume of cases are available for validation. This situation arise 1) when the RDR system acquires production rules in batch mode with large volumes of labelled or unlabelled cases, and 2) when the RDR system already is used routinely and has many cases already processed. The batch mode means when a rule is acquired all cases that satisfy it are processed at the same time. In this situation, a new rule that is acquired usually covers a number of inappropriate cases. To remove these incorrect cases, we can choose one of the following strategies:

1. **Direct Refinement strategy:** The expert directly refines the new rule adding conditions until all incorrect cases are removed, or
2. **Deferred Refinement strategy:** The expert creates the new rule with the current condition and adds exception rules later to correct errors whenever errors are detected.

The ‘deferred refinement strategy’ is practical since it does not need too much effort to refine the rule, but it does allow errors. The ‘direct refinement strategy’ is desirable since it does not allow incorrect classifications or minimally allows them, but it is not easy to construct this kind of rule with resource constraints such as limited time and information. However, if we have enough case data to acquire exceptions and we can get help from other learning techniques, it would be better that the experts directly add exceptions to refine the new rule.

For this purpose, we adopted an alternative production rule representation, called *censored production rules* (CPR) [11], for RDR. It was originally proposed as a mechanism for reasoning with incomplete information and resource constraints and the certainty of an inference could be varied to conform to cost constraints [12]. However, if we have enough data, the exceptions can be extracted directly using CPR [13]. Previously CPR have been used with unordered rules in conventional expert systems, but it could be combined with an RDR approach[14]. This paper presents how CPR can be combined with the existing RDR methods. We will review SPR and its extensions, and then we discuss how CPR can be combined with RDR.

2 Related Work

2.1 Standard Production Rule

Most expert systems employ the standard production rules and each rule is defined as

IF [condition] THEN [action].

For example, we can write admission rules as follows:

R1: IF [<student, GRE score, greater than 1350>]
THEN [<student, admission status, yes>]

R2: IF [<student, GRE score, less than 1350>]
THEN [<student, admission status, no>]

The elements of [condition] and [action] are object-attribute-value (OAV) triplets or entity-attribute-value (EAV) [15]. For example, in rule R1, [condition] has ‘student’ as object, ‘GRE score’ as attribute, and ‘greater than 1350’ as value. Similarly, [action] has ‘student’ as object, ‘admission status’ as attribute, and ‘yes’ as value. Even though [condition] and [action] triplets are similar in format, they represent different semantics. While the value in [condition] is *matched* with any value provided in a data case, the value in [action] assigns (or produces) value/values to an attribute. This process is sometimes referred to *binding*, since the attribute does not have a value, but once it is matched, a value is bound. This format is surprisingly flexible, since it is in effect a notation for binary relations. Multiple [condition] and [action] elements can be combined by conjunctive [AND] or disjunctive [OR] clauses [7]. One of the main criticisms of the standard production rule approach is that “*it severely fragments the knowledge that exists in the data, thereby resulting in a large number of rules. The fragmentation also makes the discovered rules hard to understand and to use.*” [16]

2.2 Extension of Standard Production Rule with an ELSE Statement

One extension of standard production rules is to add an ELSE statement as follows:

IF [condition] THEN [action] ELSE [alternative action].

For example, R1 and R2 can be combined as follows:

R3: IF [<student, GRE score, greater than 1350>]
THEN [<student, admission status, yes>]
ELSE [<student, admission status, no>]

Rule R3 is equivalent to both R1 and R2 and is more compact than the basic rule representation format. However, it is generally advisable to avoid the use of ELSE

statements in expert systems [15]. On the one hand, validation of such rules is more difficult than those of their basic IF-THEN equivalents. On the other hand, when encountered in the inference process, such rules will tend to always reach a conclusion. This can result in some unexpected results. Knowledge bases with many rules containing ELSE may behave more like conventional programs.

2.3 Extension of Basic Production Rule with Censor Statement

Another extension of standard production rules is censored production rules (CPR), which have the exceptions, called censors, to the standard production rules [11-12]. Each rule in CPR takes the form

IF [condition] THEN [action] UNLESS [censor-condition].

For example, the following rule represents a censored production rule:

R4: IF [<student, GRE score, greater than 1350>
 THEN [<student, admission status, yes>
 UNLESS [<student, GRE score, older than 5 years>]

CPR was proposed in a different context. Firstly, it was proposed as the idea of a logic in which the certainty of an inference could be varied to conform to incomplete information and resource constraints [11]. This Variable Precision Logic (VPL) used CPR to encode both domain and control information. Hierarchical Censored Production Rules (HCPR) extend VPL by explicitly considering specificity as well as certainty. HCPR were 'Hierarchical' since "it is possible for related HCPRs with different levels of specificity to be treated in a tree structure" [17-19]. Secondly, the rule-plus-exception model (RULEX) was proposed for modelling human classification learning to investigate the psychological plausibility of a "rule + exception" learning strategy [20-21]. The RULEX representation assumes that people tend to form simple logical rules and memorize occasional exceptions to those rules when learning to classify objects, so it learns a decision tree on a trial-by-trial basis using induction over examples. The model accounts for many fundamental classification phenomena. Furthermore, research showed individuals vary greatly in their choices of rules and exceptions to the rules, which leads to different patterns of generalization[20]. Yiyu et al. [22] suggested a RULEX based information analysis. Thirdly, the censored production rule approach was used in summarising or transforming machine learning models. Delgado et al. [23] surveys various exception rule mining in the association rule mining context, but in most methods rules only have exceptions, but no relations to each other. Liu and his colleagues proposed a method, called general rules, summaries and exceptions (GSE) to organize, summarize and present discovered association rules[24] or decision trees [16]. Dejean [25] proposed an exception rule learning method for learning linguistic structures, where he tried to use exceptions to remove noise in linguistic data. Lastly, Boicu et al. [26] suggest a knowledge acquisition system that uses censored production rules in the knowledge

acquisition context, where the system (agent) helps the experts find possible general rules and their exception rules.

We wish to use censored production rules in RDR. Our approach differs from the previous research. RDR differs from VPL and HCPR since it explicitly structured rules based on the assert rule and its exceptions relation. VPL does not consider the order of rules and HCPR only considers relations based on ‘specificity’. Though Liu and his colleagues [16, 24] structured the rules as a tree hierarchy, they did not consider domain knowledge as part of knowledge acquisition. Boicu et al. [26] used censored production rule as well as domain experts, but it is not clear how they organised the rules. In a previous study of possible knowledge representations for RDR, censored production rules, but then called composite rules, were investigated and in simulation studies were shown to learn a single classification domain more rapidly than other RDR [14]. This study used the conventional deferred refinement strategy for RDR whereas here we examine the utility of CPR for direct refinement when many validation cases are available.

3 RDR with Censored Production Rule

In this section, we will describe how to combine CPR with RDR. Our use of censored production rules is based on multiple classifications RDR (MCRDR), although the approach should be able to be used for any RDR variant. The RDR process is described **Fig. 1**.

Algorithm Process ()

Process a case with current knowledge base

```

-----
create a queue of Case object named as case-pool;
while case-pool is not empty do
  current-case ← dequeue from case-pool; // current-case is a Case object
  fired-rule-set ← Inference (current-case, root-rule);
  if fired-rule-set is not empty then
    while fired-rule-set is not empty do
      fire-rule ← dequeue from fired-rule-set; // fired-rule is a Rule object

      if fired-rule is wrong then
        create an empty Rule object named as new-rule;
        AcquireRule(current-case, fired-rule, new-rule);
    Else
      create an empty Rule object named as new-rule;
      AcquireRule(current-case, fired-rule);
    end if
  end while
end while

```

Fig. 1. Algorithm Process ()

RDR retrieves all cases to be processed as a *case-pool*, where each element is a Case object and a Case holds an identifier, attributes and a class, if this has been

specified. For each case, RDR obtains the inference result using the Inference (*current-case*, *root-rule*, *temp-conclusion*, *final-conclusion*) algorithm. The result of the inference is a set of Rule objects and each object contains a rule identifier, parent rule identifier, condition, and a conclusion class. If there is no inference result or a wrong inference result, the experts acquire new rules using AcquireRule (*current-case*, *fired-rule*, *new-rule*) algorithm. *new-rule* is an object of Rule which has a null value.

Algorithm Inference (Case *case*, Rule *current-rule*, Queue *temp-conclusion*, Queue *final-conclusion*)

Get fired rules with a case and current knowledge base

```

-----
get a queue of Rule objects named as child-rule-set
  that have current-rule as parent from knowledge base;

if child-rule-set is empty then
  enqueue current-rule into final-conclusion;
else
  set a Boolean named as isChildFired as false;
  while child-rule-set is not empty do
    child-rule ← dequeue from child-rule-set;
    if condition of child-rule is matched to case then
      if sensor-condition of child-rule is not matched to case then
        push child-rule to temp-conclusion;
        isChildFired=true;
      end if
    end if
  end while
end if

if isChildFired equals true then
  enqueue current-rule into final-conclusion;
end if

if temp-conclusion is not empty then
  current-rule ← dequeue from temp-conclusion;
  Inference (case, current-rule, temp-conclusion, conclusion);
else
  return final-conclusion;
end if

```

Fig. 2. Algorithm Inference ()

The detailed inference process is described in **Fig. 2**. The inference process starts with four parameters – *case* (an object of Case), *current-rule* (an object of Rule), *temp-conclusion* (a queue of Rule objects temporarily fired), and *final-conclusion* (a queue of Rule objects finally fired). Initially the process starts with a null value for *temp-conclusion* and *final-conclusion* and root rule for *current-rule*. With these parameters RDR retrieves the child rules of *current-rule*. If there is no child rule, the

current-rule becomes an element of *final-conclusion*. If there are child rules, RDR sets *isChildFire* as false and evaluates all child rules whether or not any rule is fired. If any rule is fired the fired rule adds an element of *temp-conclusion* and *isChildFire* is set as true. After evaluating all child rules, if *isChildFire* is false, current-rule is added as *final-conclusion*. Finally if *temp-conclusion* is null, RDR returns *final-conclusion*; otherwise, it picks a temporary fired rule from *temp-conclusion* and performs the inference process with this rule. The inference process with the CPR based RDR is the same as standard production rule based RDR. The only difference is that the CPR based RDR needs to evaluate whether or not the censor conditions are matched to the given case. If any censor-condition is matched, the rule is not fired; otherwise it is fired. The censors is logically interpreted as

IF [condition] & $\bar{}$ [censor-condition] THEN [action].

The steps for evaluating [condition] and [censor-condition] can be combined together and there is no critical difference in inference process between the SPR based RDR and the CPR based RDR. However, censors are used in a different way to negating conditions. This will be discussed in the next section.

The rule acquisition process with CPR-based RDR is summarized in **Fig. 3**. The process is initiated with three parameters – current case (*current-case*), current fired rule/s (*fired-rule*), and a new rule (*new-rule*). At first, *new-rule* is null and its value changes over the rule acquisition process. The expert changes the condition of *new-rule* using *current-case* and the cornerstone case/s of *fired-rule*. Once the expert finalises the condition composition, the system retrieves a set of cases that satisfy the condition of *new-rule* as well as those of *fired-rule*. As discussed above the expert attempts to define the most appropriate condition for *new-rule*. The rule is then tested on the validation cases. As a rule will nearly always be over-generalised, *censor conditions* are added until all the incorrect validation cases are all removed. Finally the new rule is added to the knowledge base as a child of the current fired rule.

Algorithm AcquireRule (Case *current-case*, Rule *fired-rule*, Rule *new-rule*)
Acquire a new rule to handle the current case

 modify *condition* of *new-rule*;
 retrieve a set of cases that satisfy the condition of *new-rule*;
if *condition* of *new-rule* is sufficient **then**
 while incorrect cases is empty **do**
 add *censor-condition* to *new-rule*;
 end while
else
 AcquireRule (*current-case*, *fired-rule*, *new-rule*);
end if
 register *new-rule* as a child of *fired-rule*;

Fig. 3. Algorithm AcquireRule ()

4 Discussion

4.1 Cornerstone Cases

One of the unique features of RDR is that it keeps cases that were used to create rules, called cornerstone cases, and they are used in consequent knowledge acquisition [2]. Introducing CPR to RDR makes changes in cornerstone case management. While the conventional RDR approach only maintains conforming cases for a new rule, the CPR based approach maintains non-conforming cases as well as conforming cases. The non-conforming cases are used to create censor conditions and the conforming cases are used to create the main rule condition. This feature allows the CPR based RDR to use more information when the expert creates a rule.

Let us assume that CPR is used to classify documents. Words in a document are used as conditions and topics are used as conclusions. Documents are classified in batch mode, such that when a rule is created, all documents that satisfy the condition are also classified into the conclusion. Let us assume that the expert creates the following rule with a document, doc-1.

```
IF document contains word-1 THEN class=topic-1 [doc-1]
```

With this rule, documents which have word-1 will be classified into topic-1. However, it is possible some of the classified documents should not be classified into topic-1. In this case, the expert can directly add censor rules to exclude them. For example, the following censors can be added into the above rules.

```
IF document contains word-1 THEN class=topic-1
UNLESS document contains word-2 [doc-2]
UNLESS document contains word-3 [doc-3]
```

The censor rule conditions (word-2 and word-3) are not in doc-1 but in doc-2 and in doc-3. Therefore, in this case doc-1 is used as a conforming cornerstone case and doc-2 and doc-3 are used as non-conforming cornerstone cases. If an exception rule is created for this rule, it is necessary to consider both cornerstone cases.

4.2 Combining with Machine Learning

Machine learning techniques can be combined with RDR as follows: Firstly, machine learning techniques can be used to create knowledge bases automatically with a data set. For example, Gaines and Compton proposed a Binomial theorem based RDR induction method [27] and Wada et al. [28-29] proposed a Minimum Description Length (MDL) based induction method. Secondly, machine learning methods can be used to provide simulated experts in evaluating RDR [14, 30]. Lastly, machine learning techniques can be used as a part of knowledge acquisition. The MDL method proposed by Wada et al. provided for both induction and knowledge acquisition on

the same knowledge base [27]. We suggest that they can also be combined by using machine learning techniques to identify sensors.

To exemplify this, we used the ‘adult’ data set, obtained from the UCI machine learning repository (<http://archive.ics.uci.edu/ml/datasets/Adult>). This data set consists of 32561 training and 16281 test data with about 7% missing value records in the training data. The data has 6 continuous and 8 nominal features and the response class is a binary nominal feature. Each case describes whether case person has more or less than \$50K income. The data set also provides reference performance for different machine learning methods, whose error rates ranges from 14% to 21%. Let us assume that the expert acquires the following rule without sensors:

```
IF (person, marital-status, Never-married)
THEN (person, income, <50K)
(491/10,683)
```

This rule covers 10,683 cases, but a total of 491 cases (4.6%) of these have the wrong class ($\geq 50K$). To learn sensor condition with machine learning, we used See5 (<http://www.rulequest.com/see5-info.html>), a commercial version of the well known C4.5 decision tree learner. **Fig. 4** shows a decision tree generated by C5 using 10,683 cases covered by the above rule.

Decision tree:

```
capital-gain > 7443:
...age > 22: >50K (135)
: age <= 22:
: ...age <= 20: <=50K (4)
:   age > 20: >50K (6)
capital-gain <= 7443:
...capital-loss > 2206:
...capital-loss <= 2339: <=50K (23/7)
:   capital-loss > 2339: >50K (19)
capital-loss <= 2206:
...education-num <= 12: <=50K (8246/88)
education-num > 12:
...education-num <= 14: <=50K (2107/181)
education-num > 14:
...age <= 32: <=50K (62/6)
age > 32: >50K (81/32)
```

Classification result (10683 cases) :

```
(a) (b) <-classified as
---- ----
209 282 (a): class >50K
32 10160 (b): class <=50K
```

Fig. 4. C5 Decision Tree Result

The decision tree shows that if a person has a capital gain of more than \$7,443 and their age is greater than 20, his/her income is greater than 50K. This condition covers 141 cases out of the wrongly classified 491 cases (28.7%).

[1] UNLESS (person, capital-gain, >7443) AND
(person, age, >20) [141/0];

Similarly the following censor conditions can be added based on the decision tree results:

[2] UNLESS (person, capital-gain, <=7443) AND
(person, capital-loss, >2339) [19/0];
[3] UNLESS (person, capital-loss, <=2339) AND
(person, education-num, >14) AND
(person, age, >32) [81/32];

However, this approach has problems since the subset of cases covered by the rule are likely to be unbalanced. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore small classes, resulting in very poor performance for the minor class. Therefore, the censors acquired by using a decision tree may perform badly. On the one hand, the censors may not cover all incorrect cases. In our example, the decision tree rules correctly classify a total 209 cases out of 491 cases, which have '>50K' income, but also wrongly classify 32 cases, which have '<=50K' as the class, and classify them into the '>50K' class. The censors obtained by a decision tree may be incomplete (not cover all incorrect cases) or incorrect (censor 3 classifies 32 cases wrongly). It would be useful to use machine learning techniques that manage this kind of problem, such as OcVFDT [31].

5 Conclusions

This paper proposes a new RDR approach that uses censored production rules instead of the standard production rules. This suggested method can be integrated with any existing RDR approaches. We discussed the impact of censored production rules on the inference and knowledge acquisition processes. The introduction of censored production rule should not affect on the inference process, but changes the knowledge acquisition process. The censored production rule based RDR can preserve richer context information than the standard production rule based RDR. We expect our approach would be appropriate when there are large volumes of data that have class flags or batch mode knowledge acquisition (e.g. document classification) is appropriate. Our approach should be beneficial in knowledge acquisition for validating cases. This paper covers only preliminary suggestions, and in future work the supposed advantages will be investigated experimentally.

Acknowledgement. This work was funded by Australian Research Council Discovery Project Grant and Korea Association of Industry Academy and Research Institute.

References

- [1] Richards, D.: Two decades of ripple down rules research. *The Knowledge Engineering Review* 24(2), 159–184 (2009)
- [2] Compton, P., Jansen, R.: A philosophical basis for knowledge acquisition. *Knowledge Acquisition* 2(3), 241–258 (1990)
- [3] Kang, B., Compton, P., Preston, P.: Multiple Classification Ripple Down Rules: Evaluation and Possibilities. In: 9th AAI Sponsored Banff Knowledge Acquisition for Knowledge-Based Systems Workshop, University of Calgary, Banff (1995)
- [4] Mulholland, M., Preston, P., Sammut, C., Hibbert, B., Compton, P.: An expert system for ion chromatography developed using machine learning and knowledge in context. In: *Proceedings of the 6th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 258–267. Gordon & Breach Science Publishers, Edinburgh (1993)
- [5] Forgy, C., McDermott, J.P.: OPS, A Domain-Independent Production System Language. In: 5th International Joint Conference on Artificial Intelligence, pp. 933–939. William Kaufmann, Cambridge (1977)
- [6] Forgy, C.L.: Rete: A fast algorithm for the many pattern/many object pattern match problem. *Artificial Intelligence* 19(1), 17–37 (1982)
- [7] Bench-Capon, T.J.M.: *Knowledge Representation - An Approach to Artificial Intelligence*. The APIC Series, vol. 32. Academic Press (1990)
- [8] Pedersen, K.: Well-structured knowledge bases. *AI Expert* 4(4), 44–55 (1989)
- [9] Melle, W.v.: A domain-independent production-rule system for consultation programs. In: *Proceedings of the 6th International Joint Conference on Artificial Intelligence*, vol. 2, pp. 923–925. Morgan Kaufmann Publishers Inc., Tokyo (1979)
- [10] McDermott, J.: RI: the formative years. *Readings from the AI Magazine*, 93–101 (1988)
- [11] Michalski, R.S., Winston, P.H.: Variable precision logic. *Artificial Intelligence* 29(2), 121–146 (1986)
- [12] Haddawy, P.: Implementation of and Experiments with a Variable Precision Logic Inference System. In: *AAAI 1986*, pp. 238–242 (1986)
- [13] Prati, R.C., Monard, M.C., de Carvalho, A.C.P.L.F.: A Method for Refining Knowledge Rules Using Exceptions. In: *ASAI 2003 Simposio Argentino de Inteligencia Artificial*, Buenos Aires, Argentina (2003)
- [14] Cao, T.M., Compton, P.: A simulation framework for knowledge acquisition evaluation. In: *Proceedings of the Twenty-Eighth Australasian Conference on Computer Science*, vol. 38, pp. 353–360. Australian Computer Society, Inc., Newcastle (2005)
- [15] Ignizio, J.P.: *Introduction to expert systems: the development and implementation of rule-based expert systems* (1991)
- [16] Liu, B., Hu, M., Hsu, W.: Intuitive representation of decision trees using general rules and exceptions. In: 17th National Conference on Artificial Intelligence, pp. 615–620 (2000)
- [17] Jain, N.K., Bharadwaj, K.K.: Some learning techniques in hierarchical censored production rules (HCPRs) system. *International Journal of Intelligent Systems* 13(4), 319–344 (1998)
- [18] Jain, S., Jain, N.K.: A generalized knowledge representation system for context sensitive reasoning: Generalized HCPRs System. *Artificial Intelligence Review* 30(1-4), 39–52 (2008)
- [19] Bharadwaj, K.K., Jain, N.K.: Hierarchical Censored Production Rules (HCPRs) system. *Data & Knowledge Engineering* 8(1), 19–34 (1992)

- [20] Navarro, D.J.: Analyzing the RULEX model of category learning. *Journal of Mathematical Psychology* 49(4), 259–275 (2005)
- [21] Nosofsky, R.M., Palmeri, T.J., McKiley, S.C.: Rule-plus-exception model of classification learning. *Psychological Review* 101, 53–79 (1994)
- [22] Yiyu, Y., Fei-Yue, W., Zeng, D., Jue, W.: Rule+exception strategies for security information analysis. *IEEE Intelligent Systems* 20(5), 52–57 (2005)
- [23] Delgado, M., Ruiz, M.D., Sánchez, D.: Mining Exception Rules. In: Bouchon-Meunier, B., Magdalena, L., Ojeda-Aciego, M., Verdegay, J.-L., Yager, R.R. (eds.) *Foundations of Reasoning under Uncertainty*. STUDEFUZZ, vol. 249, pp. 43–63. Springer, Heidelberg (2010)
- [24] Liu, B., Hu, M., Hsu, W.: Multi-level organization and summarization of the discovered rules. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 208–217. ACM, Boston (2000)
- [25] Dejean, H.: Learning rules and their exceptions. *The Journal of Machine Learning Research* 2, 669–693 (2002)
- [26] Boicu, C., Tecuci, G., Boicu, M., Marcu, D.: Improving the Representation Space through Exception-Based Learning. In: *Sixteenth International Flairs Conference*, pp. 336–340. AAAI Press (2003)
- [27] Gaines, B.R., Compton, P.: Induction of ripple-down rules applied to modeling large databases. *Journal of Intelligent Information Systems* 5(3), 211–228 (1995)
- [28] Wada, T., Horiuchi, T., Motoda, H., Washio, T.: Characterization of Default Knowledge in Ripple Down Rules Method. In: Zhong, N., Zhou, L. (eds.) *PAKDD 1999*. LNCS (LNAI), vol. 1574, pp. 284–295. Springer, Heidelberg (1999)
- [29] Wada, T., Horiuchi, T., Motoda, H., Washio, T.: Integrating Inductive Learning and Knowledge Acquisition in the Ripple Down Rules Method. In: *6th Pacific Knowledge Acquisition Workshop*, Sydney, Australia, pp. 325–340 (2000)
- [30] Compton, P.: Simulating Expertise. In: *PKAW 2000: The 2000 Pacific Rim Knowledge Acquisition Workshop*, Sydney, Australia (2000)
- [31] Li, C., Zhang, Y., Li, X.: OcVFDT: one-class very fast decision tree for one-class classification of data streams. In: *Proceedings of the Third International Workshop on Knowledge Discovery from Sensor Data*, pp. 79–86. ACM, Paris (2009)

RM and RDM, a Preliminary Evaluation of Two Prudent RDR Techniques

Omaru Maruatona, Peter Vamplew, and Richard Dazeley

Internet Commerce Security Laboratory
University of Ballarat Ballarat, Australia
o.maruatona@icsl.com.au

Abstract. Rated Multiple Classification Ripple Down Rules (RM) and Ripple Down Models (RDM) are two of the successful prudent RDR approaches published. To date, there has not been a published, dedicated comparison of the two. This paper presents a systematic preliminary evaluation and analysis of the two techniques. The tests and results reported in this paper are the first phase of direct evaluations of RM and RDM against each other.

Keywords: Prudence Analysis, RDR, MCRDR, KB brittleness, RM, RDM.

1 Introduction

Traditional knowledge based systems (KBS) have been often criticized for ignoring Knowledge Acquisition (KA) and maintenance innovations [1], [2]. Consequently, Ripple Down Rules (RDR) was introduced as an incremental KA technique whereby KA and maintenance are essentially integrated and usually not requiring the additional services of a knowledge engineer. RDR has since been used in commercial applications including in the Pathology Interpretative Expert Reporting System (PIERS) system, which has been described as user maintained and not requiring knowledge engineering expertise [3]. Due to RDR's inability to provide more than a single classification, Multiple Classification RDR (MCRDR) was introduced with the ability to generate multiple classifications [4]. A further advancement in RDR technologies was the idea of Prudence Analysis (PA). Prudence was introduced to address KBS brittleness, which occurs when a KBS does not realise when its knowledge is inadequate for a particular case [5]. A prudent KBS is one with a mechanism to issue warnings or alerts whenever a current case is beyond the system's expertise. This paper reports on a methodical comparison of two PA techniques: RM and RDM. These two methods had been independently evaluated before but have never been directly compared. Another contribution of this paper is the introduction of a Multiple Classification version of RDM.

2 Rated MCRDR (RM)

RM is a hybrid approach combining MCRDR with an Artificial Neural Network (ANN) [6]. RM is based on [7]'s premise that if captured, a pattern of fired MCRDR

rules can provide an additional context about a given domain. A grouping of this pattern can be given a value representing its contribution to a particular task [7]. RM has a MCRDR output simplifying mechanism which indexes MCRDR conclusions into a set of binary inputs for the ANN. These inputs are assigned a 0 or 1 value depending on whether the particular rule was fired for the current case. The following diagram illustrates the basic composition of RM.

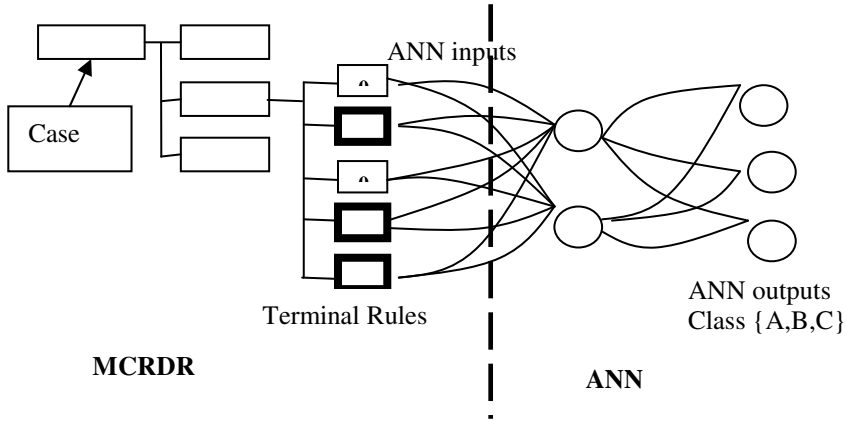


Fig. 1. A basic RM schematic. The bolded MCRDR outputs represent 1 and 0 for the non bolded outputs.

The indexed binary set is fed into a standard 3 layer perceptron ANN such that each firing terminating rule will produce a 1 input for the ANN, and a 0 input for non firing terminal rules. For example, in the RM diagram, the terminating rules are indexed into a binary word 01011 which is the input for the ANN. The ANN uses two main learning approaches. If there are no new rules added to MCRDR, a standard back-propagation algorithm with a sigmoid thresholding function is engaged. If a new rule is added in MCRDR, an additional input is created for the ANN. This may be problematic to the ANN in terms of erasing the previously learned information. To counteract this threat, new shortcut connections are introduced from the newly created input to each output node. The shortcut weights are calculated using the single step initialization formula [6] (see equation 1 below).

$$w = \left(\log \left(\frac{fnet + \partial + 0.5}{0.5 - (fnet + \partial)} \right) \right) - ((\Sigma A) + (\Sigma B)) / m \tag{1}$$

where A and B are the weighted sums at the hidden and output nodes respectively, \mathbf{z} is the step distance modifier in the range of 0 to 1. It is the rate of adjustment of for the new features and determines how quickly the shortcut weights adjust to the correct output. m is the number of newly added inputs and ∂ is the sum of differences between the network calculated outputs and the target outputs (or error sum value) at an

output neuron. As the MCRDR produces different classifications, the ANN learns the patterns of the fired rules for each classification. A warning is then given whenever the MCRDR and the ANN produce different classifications.

3 Ripple Down Models (RDM)

RDM, like RM has two main components, the RDR part and a complementary outlier detection mechanism. As in RM, RDM first engages an RDR engine and passes the output to the complementary outlier detection component. In RDM, the RDR output passed to the outlier detector is a model (hence the acronym RDM) [8]. A model is made up of situated profiles. Each situated profile consists of a number of profiles corresponding to the number of attributes in a case. RDM has two outlier detection functions: the Outlier Estimation with Backward Adaptation (OEBA) for continuous attributes and the Outlier Detection for Categorical Attributes (OECA) for discrete attributes [9].

For OEBA, profiles of each attribute in a case are grouped as a Situated Profile and organised according to the conclusions generated by RDR. For example, an OEBA Situated Profile may contain minimum and maximum values for each attribute for the corresponding RDR classification. For each classification produced by RDR, a Model comprising the Situated Profile(s) is returned to the outlier detection component. Ideally, OEBA should flag an anomaly for incorrect classifications by RDR. If an outlier was flagged incorrectly, then Backward Adaptability adjusts the appropriate profiles' minimum and maximum values. In OECA, each profile keeps a set of an attribute's values, a corresponding M value and a New Value Ratio (NVR. The NVR is the ratio of the current attribute's M value and the M value for the last updated value in the profile [8]. An anomaly is flagged when the NVR of a case is greater than a set threshold.

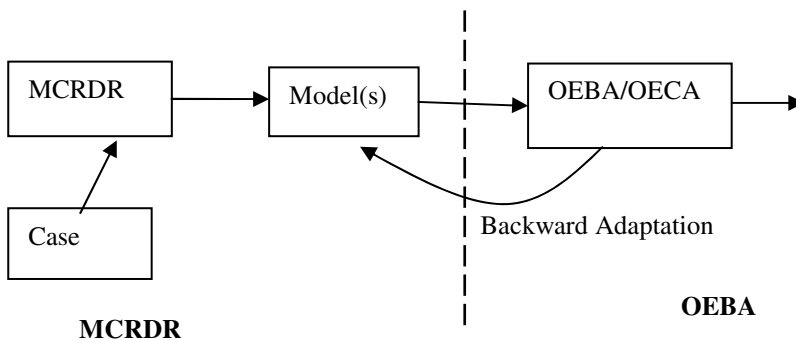


Fig. 2. RDM schematic

Originally, RDM was designed with models passed from a single class RDR engine [8]. This research developed a multiple classification version of RDM where models were passed from a MCRDR rule base. The primary difference between the two versions is that the MCRDR alternative has the ability to generate multiple models from a single case if need be. As in RM, the prudence of RDM is in how well the warning system works. Figure 2 shows the general architecture of RDM.

4 Evaluation Methodology

4.1 Simulated Expert

Evaluating KA methods is an important but difficult task mainly because it is hard and expensive to get a readily available expert for controlled tests [10]. A common solution to this problem has been the use of simulated experts. Simulated experts have been used extensively in testing RDR methodologies [7], [6], [10]. For this research, the simulated expert uses a ruleset file (for each dataset) generated from the See5 tool. For each dataset, only cases that could be matched to a rule (or condition) were used such that the resultant simulated expert was faultless and missed no cases.

4.2 Test Data

Three simple UCI datasets were used for these tests mainly because developing perfect simulated experts for such data is less time consuming. The tests reported in this paper are a preliminary part of a wider research project. Table 1 describes the three datasets used. The last column of the table shows the ratio of each dataset's rules to the total number of cases.

Table 1. Description of datasets

Name	Type	Instances	Rules in SE	Rules Ratio
Iris Plants	Numerical	146	5	3%
Car Evaluation	Categorical	288	15	5%
Physical Action	Numerical	250	60	24%

4.3 Evaluation Metrics

The comparison of the two PA systems, RM and RDM was based on two metrics: Balanced accuracy and prudence. Balanced accuracy is based on the system's True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP). The TP in this case is when a warning was issued correctly. TN is when a warning was not issued correctly. FN includes instances when a warning was not issued but should have. When a warning was issued incorrectly, then this is a FP [8]. Balanced accuracy combines these metrics to prevent a scenario where a system can warn on every case and still get an accuracy of 100%. The following formula is used for balanced accuracy:

$$bA = TP(\alpha/T)/\beta + TN(\beta/T)/\alpha \tag{2}$$

where $\alpha = TN + FP$, $\beta = TP + FN$ and $T = TN + TP + FN + FP$.

The prudence of a system will be determined by the effectiveness of its warning mechanism. This will be the rate of correct warnings minus the rate of incorrect warnings. For example, given a dataset of 40 cases in which 10 are TP’s and 30 are TN’s, a system with 8 TP’s and 3 FP’s will have a prudence measure of $(8/10)\% - (3/40)\%$ which is 72.5%. Formula 3 is used to calculate the prudence measure.

$$p = (TPs/TP)\% - (FPs/T)\% \tag{3}$$

where TPs is the number of correct warnings issued by the system, TP is the total number of warnings that should have been issued. In this study, the total number of warnings expected is the number of rules in the simulated expert. FPs is the incorrect warnings issued by the system and T is the total number of cases in a dataset.

Incorrect warnings and the proportion of the TP in the data also affect the overall prudence measure of a system. In cases where a dataset has a large proportion of TP’s, it may be better to lower the warning threshold so that the system issues more warnings. As the dataset grows and fewer rules are added to the system, there might be a need to raise the warning threshold effectively increasing the system’s prudence. This is because as the system acquires knowledge and sees fewer new cases, the frequency of warnings is likely to decline.

5 Results and Analysis

Table 2 displays the two PA systems’ corresponding TP, TN, FP and FN metrics on the three datasets. For the Physical Action dataset, RM was tested with two different z values, 0.01 and 0.9. In the other datasets, RM’s z value was set at 0.5. Table 2 shows the two systems’ balance accuracy (calculated using formula 2) and prudence measures computed from formula 3.

Table 2. RM and RDM’s confusion metrics, Balanced Accuracy (BA) and Prudence (P)

Dataset	System	TP	TN	FP	FN	BA (%)	Pr (%)
Iris	RM	4	142	0	0	100	100
	RDM	4	135	7	0	99.86	95
Car Evaluation	RM	115	48	73	52	52	40
	RDM	160	102	23	3	89	42
Physical Action	RM (= 0.01)	146	8	89	7	42	30
	RM (= 0.9)	119	47	50	34	60	56
	RDM	160	29	40	21	54	44

5.1 Analysis

The RM method seems to have a slightly better balanced accuracy and prudence over RDM in the Iris dataset. In the Car Evaluation dataset, RDM outperformed RM by a vast margin in terms of BA but was just slightly better in Pr. However, [7] advises that RM's accuracy and prudence is not preset and can be controlled by altering the \mathbf{z} value. For the dataset that RM was tested with different \mathbf{z} values, it is clear that a high \mathbf{z} value (0.9) produces a higher BA and Pr than RDM and a low \mathbf{z} value conversely resulted in a BA and Pr much less than RDM's. Based on the results in Table 2, there does not seem to be an obvious correlation between balanced accuracy and prudence. However there seems to be a consistency in that the system with the higher BA also had a higher PA. The ratio of rules to the total number of cases a dataset has does not seem to affect the prudence of either system. The prudence results were expected to be lower for the Physical Action dataset since each rule covers very few cases. The likelihood of a misclassification in such a setting is compounded by the fact that the differences between the rules may be minute. So when a system has proportionally many, almost similar rules, it is likely that some rule may overlap with another, resulting in a lot more misclassifications. For these tests however, this claim was not affirmed.

6 Conclusion

RDM and RM are two PA systems whose accuracy and viability have been demonstrated in different domains [8,11]. These two approaches have not been directly compared previously. This paper presented a preliminary comparison of the two systems using three relatively small datasets. For the smallest and simplest dataset, RM appears to have a higher accuracy and prudence, albeit by a small margin. RDM outperformed RM in the categorical dataset and in the Physical Action dataset, RDM's performance was almost midpoint between RM's optimal setting ($\mathbf{z} = 0.9$) and worst setting ($\mathbf{z} = 0.01$). The tests conducted for this paper are part of a bigger research project whose aim is to integrate RM and RDM into a single, prudent anomaly detection system. Future tests will use bigger, complex datasets and will use optimal configurations for the two systems.

References

- [1] Richards, D.: Two decades of Ripple Down Rules research. *The Knowledge Engineering Review* 24(2), 159–184 (2009)
- [2] Compton, P., Peters, L., Edwards, G., Lavers, T.G.: Experience with Ripple-Down Rules. In: *AI 2005*, Cambridge (2005)
- [3] Edwards, G., Compton, P., Malor, R., Srinivasan, A., Lazarus, L.: PEIRS: a pathologist maintained expert system for the interpretation of chemical pathology reports. *Pathology* 25(1), 27–34 (1993)

- [4] Kang, B., Compton, P., Preston, P.: Multiple Classification Ripple Down Rules: Evaluation and Possibilities. In: 9th Banff Knowledge Acquisition for Knowledge Based Systems Workshop, Banff (1995)
- [5] Compton, P., Preston, P., Edwards, G., Kang, B.: Knowledge based systems that have some idea of their limits. *CIO* 15, 57–63 (1996)
- [6] Dazeley, R., Kang, B.: Detecting the Knowledge Boundary with Prudence Analysis. In: Wobcke, W., Zhang, M. (eds.) *AI 2008. LNCS (LNAI)*, vol. 5360, pp. 482–488. Springer, Heidelberg (2008)
- [7] Dazeley, R.: To the Knowledge Frontier and Beyond: A Hybrid System for Incremental Contextual-Learning and Prudence Analysis. University of Tasmania, PhD Thesis (2007)
- [8] Prayote, A.: Knowledge Based Anomaly Detection, University of New South Wales, PhD Thesis (2007)
- [9] Prayote, A., Compton, P.: Detecting Anomalies and Intruders. In: Sattar, A., Kang, B.-H. (eds.) *AI 2006. LNCS (LNAI)*, vol. 4304, pp. 1084–1088. Springer, Heidelberg (2006)
- [10] Compton, P., Preston, P., Kang, B.: The Use of Simulated Experts in Evaluating Knowledge Acquisition. In: Knowledge Acquisition for Knowledge Based Systems Workshop, Banff (1995)
- [11] Dazeley, R., Kang, B.: The Viability of Prudence Analysis. In: The Pacific Rim Knowledge Acquisition Workshop, Hanoi (2008)
- [12] Dazeley, R., Kang, B.: Detecting the Knowledge Frontier: An Error Predicting Knowledge Based System. In: Knowledge Acquisition Workshop, Auckland (2004)

Planning Children's Stories Using Agent Models

Karen Ang and Ethel Ong

Center for Language Technologies, De La Salle University, Manila, Philippines
karenang0903@yahoo.com, ethel.ong@delasalle.ph

Abstract. Automatic story generation systems must consider both consistency and coherency issues in the stories that they produce. This paper discusses the different agent models adapted to enhance the planning process of Picture Books 2 to produce stories that contain a logical flow of events. Along with the agent models, the ontology structure utilized as the source of storytelling knowledge must be efficiently designed to support the tasks of the agents. Preliminary results show that the interaction among the agents enabled the story planner to perform better than the earlier version of Picture Books 2 in generating logical stories.

Keywords: Automatic story generation, Agents, Ontology.

1 Introduction

Picture Books 2 [1] was developed to produce fable-like stories for children age 6-8 years old. It follows from its predecessor, Picture Books 1 [2], where the series of events focused on a child's behavioral development from negative to positive, commencing with his disobedience, experiencing the consequences of his actions, and acquiring a positive behavior at the end of the story. Though the system is able to generate good stories as evaluated by human experts, the reasoning engine used by its planner to reason over the storytelling knowledge is inadequate. The planner simply retrieves the causal chain of events between the initial state (the setting of the story) and the target goal (or moral lesson) of the story to be generated. By solely relying on a causal chain of events, the scoring function used to select among candidate sets of story events is random, resulting in stories with unbelievable or illogical flow.

There are two main factors for this. First, events are represented as a chain of cause-effect using the binary relation *EffectOf*. This representation is insufficient to support the planner in inferencing over the events chain. The events must be augmented with pre-conditions and post-conditions to define the criteria to be satisfied for an event to occur, and to be satisfied after the event has taken place, respectively.

The second factor is the lack of representation of the world state. The story world state comprises the first level among five levels of story knowledge representation defined by [3]. As characters perform actions and interact with one another, their physical and emotional states, as well as the states of the different objects (that they manipulate or use for interaction) in the environment, also change. Succeeding events and actions may be influenced by these changes. Thus, a mechanism must be in place to track the world state for use by the story generator during its planning.

This paper presents our work in developing PB Planning Agents (PA), a story planner for Picture Books 2 that utilizes three agents in the generation of children's stories. These three agents are adapted from the multi-agent framework of the Virtual Storyteller [4] that models three types of knowledge, namely the plot agent for narrative structures, the character agent for representing story characters, and the world agent to represent the world state.

2 Related Systems

PB-PA is a story planner that utilizes agents to model the various knowledge sources needed in generating a story. The work is motivated by the current limitations in the implementation of the Picture Books 2 system along two aspects, namely on event representation and modeling of the world state.

This section gives a brief overview of the issues in the Picture Books 2 system. A more detailed discussion of the planning process and knowledge representation can be found elsewhere ([1], [5]). A review of the approaches utilized by existing story generators to model the story world state is also provided. This section ends with a discussion of the knowledge representation employed in PB-PA.

2.1 Picture Books 2

In Picture Books 2 [5], the theme of the story was determined based on an input picture comprising of multiple scenes. Each scene contains story elements that were specified by the child from a predefined repository of backgrounds, characters and objects. A theme-based cause-effect planner then utilized a semantic ontology that has been manually populated with knowledge relevant to the target themes. This knowledge includes explicit events represented as a series of cause and effect chain. The task of the story planner involved organizing these events into a logical flow to depict a child's daily activities as he explores and learns about his environment while undergoing behavioral development.

Like its predecessor, Picture Books 1 [2], Picture Books 2 adopted an author-centric approach to its story generation process. In an author-centric approach, the system "computationally models the creative process of a human author" [6]. A plan library containing predefined author goals and character goals is utilized to control the flow and outcome of the story, resulting in stories with consistent plot and structure.

However, whereas Picture Books 1 uses pre-scripted action sequences, the story planner of Picture Books 2 reasons through the semantic ontology by retrieving a causal chain of events between the initial state to the target goal of the story [7]. This led to the generation of stories that lack coherency, as evaluated by human experts.

Picture Books 2 defined coherency in terms of the story making sense and is easy to understand [1]. Because coherency can be enhanced through the use of discourse markers [8], the system achieved coherency by focusing on the generation of appropriate discourse markers to connect two events together in a single sentence, as shown in Listing 1.

Listing 1. Danny the dog went to camp.

- #1 The camp is very crowded.
- #2 He feels dizzy, because he sees the marshmallow.
- #3 Danny the dog fixes a bedsheet.

Another story that lacked coherency is shown in Listing 2, wherein both the objects found in the input scene, *marshmallow* and *flashlight*, were correctly introduced in the story. However, the *marshmallow* did not affect the story plot and was never mentioned again.

Listing 2. Danny the dog went to camp (again).

- #1 One dark evening, Danny the dog was in the camp for camping. He gets a white *marshmallow*.
- #2 The camp is very far. He feels tired, since he walks.
- #3 Danny the dog feels thirsty. He feels refreshed, because he drinks a water.
- #4 He walks. Danny the dog sees a shadow. He feels scared. He does not know what to do.
- #5 He turns on a flashlight. Danny the dog searches the shadow.
- #6 He is not scared anymore. He learns that when He is scared, He should search shadow.
- #7 Since then, Danny the dog learns to be brave.

A story is coherent if it contains logical events, in terms of believable character actions based on his trait, history of past events, and the current state of the story world. Incoherent stories, such as those in Listings 1 and 2, were generated due to the insufficiency of the representation of events in the semantic ontology of the system. Patterned after ConceptNet [9], Picture Books 2 made use of binary relations, such as *EffectOf* to model the causal relation between two events, *CapableOf* to model the actions a character can perform, and its complement *ReceivesAction* to model actions that can be applied to an object [5]. No representation of the world state is maintained as the story progresses. Consequently, the events that happen in the story were not checked to determine if the next action is possible given the current world state.

2.2 World State

The story world state [3] is an instantiation of the current properties and relationships of a concept (e.g., characters, objects, locations) in a certain point of story time. Different terms and representations of world states are implemented for various systems. Tale-Spin [10] made use of conceptual dependency expressions. Minstrel [11] used a schema-based representation to represent its goals, actions, and states of the world. It implemented a case-base problem-solving approach in constructing stories revolving around King Arthur and his knights. Similarly, Mexica [12] also stored its previous stories to construct a new one. However, it made use of the linguistic representation of actions to store knowledge about valid sets of actions and world states. In contrast, the Virtual Storyteller [4] had three sources of knowledge which are the plot agent,

the character agent and the world agent. The latter modeled the world state. Finally, Fabulist [6] made use of the intent-driven partial order causal link tuples to track and control the coherent flow of the causal chain of events that happen in the story.

2.3 Knowledge Representation

Picture Books 2 provides two models of the knowledge used by its story planner – the storytelling knowledge and the ontology. The storytelling knowledge contains the narratological information necessary to produce a coherent story, and includes the character traits that drive the possible themes of the story to be generated.

On the other hand, the ontology contains commonsense knowledge about concepts and their relationships that are familiar to children. The ontology follows the binary structure of ConceptNet [9], wherein a semantic relation is used to define the relationship between two concepts, such as those illustrated in Table 1.

Table 1. Sample Concepts and Semantic Relations in Picture Books 2

Category	Concept1	Relation	Concept2
Agent	Character	CapableOf	Look
Things	Morning	HasProperty	Sunny
Functional	Telescope	UsedFor	Gaze
Transition	Rest	EffectOf	Dizzy
Transition	Cheat	HasResolution	Apologize
Conflict	Helpful	CausesConflictOf	Ignore

PB-PA adopted the semantic ontology model of Picture Books 2 and categorized this as the domain knowledge or the static knowledge of the system. However, modifications were made to the relations supported by the planner. Table 2 shows some of the new relations used by PB-PA.

PB-PA also modeled world states in the form of fabula elements, parameters, conditions, primitives, and links. Links represent the causal relationships that exist between fabula elements. This model of the world state has been adapted from Trabasso's General Transition Network (GTN) model [13] that was modified in the Virtual Storyteller [4] to comprise the story elements and causal relationships for use in story generation. In PB-PA, the fabula elements *goal*, *event*, *action*, *perception*, and *internal element* are linked together through four kinds of causal relations, which are the *physical causality*, *psychological causality*, *motivation*, and *enablement*.

A sample fabula element is illustrated in Table 3 for the concept *lose*, which is an event. It requires the parameters #agent and #patient, and an optional parameter *location. The symbol '?' represents a variable which may refer to a state or element in the world. States include ?know, ?hasProperty, and ?holds, and world element such as ?mainchar.

Primitives [14] are composed of the most basic actions, such as *ATRANS* (transfer of abstract relationship, i.e., *give*) and *PTRANS* (transfer of physical location of object, i.e., *grasp*). It was utilized to reduce the need for appending similar conditions for each fabula element.

Table 2. Sample Relations in PB-PA

Category	Relation	Description & Example (conceptX-rel-conceptY)
Object	containedIn	Defines object <i>y</i> where object <i>x</i> is usually found in, i.e., <i>Water-containedIn-waterjug</i>
	ProducedBy	Defines what the entity <i>y</i> can produce, i.e., <i>Shadow-producedby-tangible</i>
Character	onlyDoneBy	Defines the action <i>x</i> that can be done by agent <i>y</i> only, i.e., <i>Fly-onlyDoneBy-bird</i>
	canBe	Defines the profession a character can have, i.e., <i>Child-canbe-student</i>
Generic	NegativelyEqualTo	Relates two opposite concepts, i.e., <i>Love-negativelyEqualTo-hate</i>
	kindOf	Concept <i>x</i> is a possible value of the attribute <i>y</i> , i.e., <i>Thirsty-kindOf-bodilyneeds</i>
Emo_inference	Follows	Denotes emotion <i>y</i> can be experienced after experiencing emotion <i>x</i> , i.e., <i>Satisfaction-follows-hope</i>
Spatial	DoneAt	Indicates a location <i>y</i> where the action <i>x</i> can be done, i.e., <i>Fly-doneAt-outdoor</i>
	FoundAt	Indicates where the entity <i>x</i> can be found, i.e., <i>Bird-foundAt-outdoor</i>
Time	HappensAt	Defines when (day <i>y</i>) the event <i>x</i> can take place, i.e., <i>Class-happensAt-weekday</i>
	StartsAt	Defines the time <i>y</i> when the event <i>x</i> commences, i.e., <i>Class startsAt-morning</i>

Table 3. Structure of a Fabula Element

Category	event
Concept	<i>lose</i> (#agent, #patient, *location)
Primitive	GRASP
Pre-condition	?inCareOf(#agent, #patient) ^ ?know(#agent, ?location(#patient)) ^ ?isA(#agent, character) ^ ?isA(#patient, object)
Post-condition	¬?holds(#agent, #patient) ^ ¬?hasProperty(#patient, lost) ^ ¬?know(#agent, ?location(#patient))
Recommendation	?is(?trait(#agent), irresponsible)

A set of conditions is also defined for each fabula element. Pre-conditions must be satisfied before a fabula element can be executed, to ensure that the execution of an event is logical based on the current world state. Post-conditions are the resulting states the world must possess after the fabula element is executed. Recommendations are conditions that may or may not be satisfied. Character agents use these to determine if they are acting according to their current states. Recommendations are used to

aid in the reduction of the necessary relationships. For instance, a recommendation for a fabula element, such as `?trait(#agent)-is-irresponsible`, may be used to substitute for the binary relation `irresponsible-tendsTo-lose`.

3 Agents in Planning

The design of the PB-PA planner utilizes three different agents to generate a story plan that corresponds to the input scenes and the selected theme. As illustrated in Figure 1, these are the plot agent, character agent, and world agent.

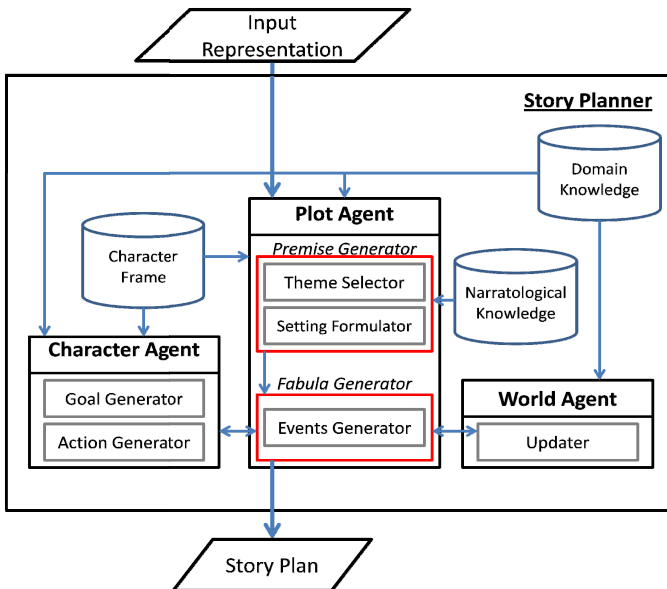


Fig. 1. Architecture of the Story Planner

The plot agent serves as the director and ensures that the story path will lead to the identified theme of the story. It supervises the actions of the characters in order to maintain an overall plot and structure [15].

The character agent generates the plan of action of a character given his goals, to ensure consistency of the character actions to his personality and thus leading to character believability as described in [16].

The plot agent and the character agent work hand-in-hand with the world agent, which represents the story world and tracks the changes to the environment as a result of character actions. The world agent checks if an action can occur by verifying if the current world state corresponds to the pre-conditions of an action or event.

3.1 Plot Agent

[17] stated that a guiding force must be present during story generation to prevent the emerging story to go to a meaningless direction. This premise is comprised of three

essential parts, which are the character and his environment, the conflict, and the conclusion. The conflict is the action and counter action it evokes, while the conclusion is described as the final state of the world.

To achieve the premise of the story, the plot agent selects the best-fit theme given the input scenes and supervises the events based on this theme. Following the approach in Picture Books 2, the theme is selected based on the negative traits of the main character [1]. A scoring function based on the conditions of a theme and the elements of the input scenes is used. The conditions include existence (in the input scenes) and positioning (on a specified location in the selected background) of the necessary characters and objects needed in the theme.

After trimming the candidate themes down based on the negative traits, it proceeds to checking the adherence of the input scenes to the set of post-conditions of both the conflict action and resolution. These are based on the required existence and positioning of the characters and objects in the scene. The following shows a sample post-condition of the action *lose*:

```
Lose :=  ¬?holds(#agent, #patient) ^
        ¬?hasProperty(#patient, lost) ^
        ¬?know(#agent, ?location(#patient))
```

The conflict *lose* has a post-condition which is for the character (*the agent*) not to hold the object (*the patient*) and requires that object to be lost (property). “*Lost*” in this case is defined to be the absence of the object in the scene, which should be depicted in the middle scene of the input picture. The third component of the post-condition requires the character to be unaware of the location of the object.

A point is given for every condition satisfied by the input scene. These points are used to calculate the score to derive the best-fit theme. If the input scenes match all the conditions of the conflict and resolution, then that theme is the best-fit theme. This criterion focuses on the possible character and objects that the user can place in the scenes. The highest scoring theme would then be selected to formulate the setting, which introduces the characters, objects, and location of the story. It is the starting point from which the story progresses.

To support its task, the plot agent makes use of a collection of narratological knowledge, comprising of themes and author goals that drive the flow of the story. This narratological knowledge is synonymous to the storytelling knowledge of Picture Books 2. The author goals [2], first introduced in Picture Books 1 but modified to work with the planning strategy of PB-PA, represent the different parts of the premise for the story. Its structure is depicted in Table 4.

Since stories revolve primarily around the main character, the events that happen in the story are composed mainly of actions that the character performs to satisfy his goal. A mechanism to ensure coherence in the story based on the selected theme is to verify if the actions to be performed by the character could lead him first to the conflict, then to experience its consequences, and finally to the resolution of the conflict. In cases of deviations, the plot agent formulates possible events to force the character to change his plan of actions towards the achievement of the story theme. This may include, among others, the performance of an action that necessitates the explicit use of objects and the possible introduction of new characters into the story. The semantic ontology, presented in Section 2, is used to identify the set of events, in consultation

with the world agent. If a story plan cannot be generated, the plot agent selects the next highest scoring theme and performs the same process again.

Table 4. Sample theme structure, ‘*Danny learns to be responsible*’

Theme Title	?MAINCHAR LEARNS TO BE RESPONSIBLE		
Target Trait	Responsible		
Conflict Action	Author Goal: <i>lose</i>		
	Fabula Element	Category	Event
		Concept	<i>lose</i> (#agent, #patient, *location)
		Primitive:	GRASP
		Pre-condition:	?inCareOf(#agent, #patient) ^ ?know(#agent, ?location(#patient)) ^ ?isA(#agent, character) ^ ?isA(#patient, object)
		Post-condition:	¬?holds(#agent, #patient) ^ ¬?hasProperty(#patient, lost) ^ ¬?know(#agent, ?location(#patient))
		Recommendation:	?is(?trait(#agent), irresponsible)
Pre-condition	?love(#agent, ?owner(#patient)) ^ ¬?owns(#agent, #patient) ^ ?love(?owner(#patient), #patient)		
Post-condition	NULL		
Conflict Counter-Action	Author Goal: <i>search</i> ...		
Resolution	Author Goal: <i>find</i> ...		

3.2 Character Agent

In a character-centric story generation system, characters are emphasized in the story and their cognitive processes are simulated as they act and react to the environment of the story world they are situated in [16]. These actions and reactions must be believable depending on their personalities.

The character agent maintains a character frame, as illustrated in Table 5, to provide a static model of a character’s predefined basic information, such as its name, gender, traits, roles and desires.

The personality of a character is one of the primary factors considered by the plot agent to select the theme. It contains both positive and negative traits of a character. (Note that a negative trait means the absence of the particular trait and can be used as a theme where the character undergoes behavioral changes as he learns this new trait.)

Desires define the general goal that a character wants. This includes both the desires for an action and for an object (a noun), for example, *children love play* and *children love sweets*.

Roles describe the relationship and the emotion of one character to another. For instance, *Peggy is Peter's sister* and *Peter loves Peggy*. Likewise, *Peter is the brother of Peggy* and also *Peggy loves Peter*. While the roles define the initial emotion of a character towards another, the emotions may change over time within the context of a story. In case *Peter* does something to make *Peggy* mad, then the emotion of *Peggy* towards *Peter* may change.

Table 5. Character frame representation

Name	Peter	
Gender	Male	
Traits	Negative	responsible, helpful, obedient
	Positive	honest, brave, persevering
Roles	Sister (Peggy)	
Desire	Play, Sweets	
Bodily needs	Food: false Water: false Rest: false Excitement: false Comfort: false	
Physical state	Mind state	Normal
	Holds	Water jug
Emotion	Love(Peggy)	
Perception	?location(water jug, camp)	

The character frame also models the dynamic attributes of a character, namely its bodily needs, physical state, emotion and perception, which are all affected based on the post-conditions of an action. Every action that is performed triggers the world state to change, including the states of the character.

Bodily needs define the basic needs of a character – food, water, excitement, comfort, and rest. At the start of the story, all these attributes are set to *false*. As the story progresses, one or more of these attributes may be set based on the action performed by the character. For example, if *Danny* played in the *camp*, he will feel *tired* based on the post-condition of the action *play*. The post-condition of the action *play* would trigger an update in *Danny's* character frame, setting his need to take a *rest* to *true*.

Emotion has a significant role in modeling characters. It influences the next actions that the characters would perform. As there are infinite emotions present, the OCC model [18] was adapted to be able to represent these in a manageable degree. The initial emotional state of a character is set to *null*. As roles are populated, these emotions change to adhere to the predefined emotion of a character towards another. Character emotions are also modified as a character performs an action, based on the defined effects or post-condition of the stated action.

Perception is what the character perceives to have happened in the world. It stores the current world state that the character is aware of. Updating the perception of the character is essential because not all facts in the story world are what the character perceives to be. For instance, if *Danny* lost the *water jug* while playing but he was at that time *distracted*, then he may not notice that the *water jug* was *lost*.

A character may have multiple goals. The character agent selects which goal to pursue first based on the character's current state. Goals are prioritized based on their intensity; goals triggered by emotions are prioritized first, followed by the bodily needs, and last by the desires. At the start of the story, a character would pursue a goal based on its desires. After every execution of an action, the character reevaluates its goal by checking the world state. If no other important issues that must be addressed are found then the character would keep on executing its plan of action to accomplish its current goal. When a circumstance happens, for instance, while *Danny* was playing he suddenly got *hurt*. In this case, *Danny* would abandon his current goal and focus on his new goal which is to *attend to his wound*. After the accomplishment of this goal and no other goals that need immediate concern are present, then the abandoned goal will carry on if it still holds.

Whenever a goal is triggered, the character agent generates the plan of action for the character to accomplish its goal. The plan of action is a causal chain of events that is retrieved from the domain knowledge. A goal is considered "accomplished" if the post-conditions of that goal have been satisfied. These conditions are also taken from the domain knowledge.

A goal and its corresponding plan of action are stored in the agent's character frame. Each is represented as a story plan element linked to another. Shown below is a sample plan of action for the goal *find*.

$$\text{Goal}(\text{find}) := \text{action}(\text{search}) \wedge \text{event}(\text{found})$$

3.3 World Agent

A story world is comprised of the characters and objects, as well as their locations and properties. Without a model of the world, Picture Books 2 [1] is unable to track the location, the position and the state of a character or an object at a particular point in the story.

The world agent holds the state of the story world at a particular point in time. This state, combined with the domain knowledge, is used by the plot agent to determine if the pre-conditions of an action or an event have been satisfied, thus allowing its execution in the story to take place. The state is updated after the completion of an action or event, and the new values are determined from the post-conditions of the completed action.

4 Preliminary Results

Though Picture Books 2 was able to generate stories, it is important to consider the quality of the resulting stories. Instead of a simple ontology structure comprising of binary *EffectOf* relations between two events and a story planner that utilizes a random strategy to select the events to be included in the story plan, PB-PA adapted different agents to reason over an ontology that has been enhanced with more knowledge to guide the planning process.

Preliminary experiments conducted on PB-PA yielded an important question – *How would the plot agent ensure that the consistent actions of the character (in relation to its predefined character traits) would not hinder the progression of the story to*

the selected theme? The traits assigned to the character may hinder the plot agent from generating stories that would adhere to the story theme. For instance, the target trait to be developed in the story is *honesty*, given the premise *break => lie => apologize*. However, if the main character also possesses the *responsible* trait, then the character agent would not permit the character to break an object, as breaking an object is an act attributed to an irresponsible character (given the conceptual relation *break-onlyDoneBy-irresponsible*). In order to select the best-fit story for the theme, PB-PA first generates all possible stories for the theme without validating the consistency of the actions to the character's traits. After all the possible plans are generated, PB-PA then performs a re-scan to score each of the story plan based on the consistency of character actions. Thus, the heuristics for the story plan is based on the number of actions done consistently in proportion to all the actions done by the character, which is retrieved by checking the recommendation of each fabula element.

The use of world states captured the state of the story world whenever an event or an action has taken place. This is very helpful in assuring that the conditions for the occurrence of an event or the performance of an action are satisfied. An example story plan (modified to be readable) is shown in Listing 3.

Listing 3. Plan elements for the occurrence of a shadow

```
#1 Bird fly near bonfire.
#2 Danny the dog saw shadow using the <INSTRUMENT> in the
    camp.
```

While Listing 2 (line #4) shows that Picture Books 2 is able to produce the same story in which the character *Danny* felt *scared* because he saw a *shadow*, it was not able to narrate how the shadow came about because the planner simply followed a causal chain of events that does not take into consideration the possibility of that event happening or how it happened. On the other hand, PB-PA's use of world states allowed it to detect that for a character to see a shadow, a source must be present, which is why the *bird* was introduced.

Listing 4 is a sample partial story plan that is generated by PB-PA. Story lines #1 to #3 are part of the story setting that brought the character to the input location. In line #4, the character agent creates a plan for the character's initial goal (i.e., *to play*). The character performs *play* to satisfy its goal. The world agent can be seen working in line #6 – an effect of *play* is that the character would become *tired*. However, the plot agent discovers that if it lets the character continue to address its new concern (which is to get some rest), the story will not lead to its intended conflict. The plot agent has to intervene; it does so by checking if it can insert an event with the previous action to advance the story. For instance, while playing, a character can be distracted and lead him to lose something, as narrated in line #5. Furthermore, because of this distraction, the character was not able to perceive that the *waterjug* was lost. As the story progresses to line #9, the character needed the *waterjug* to have a drink, which triggers his perception to lead him to a realization that the *waterjug* is missing. This realization has an emotional effect on the character, as presented in line #10. The fear of losing an object led to a goal of the need to find the missing object, as seen in the subsequent text in the story.

Listing 4. Partial story plan, 'Peter learns to be responsible'

```
#1 One evening, Peter wanted to play in the camp.
#2 Peter borrowed Peggy's waterjug.
#3 Peter went to the camp.
#4 Peter played in the camp.
#5 Peter lost the waterjug in the camp.
#6 Peter felt tired.
#7 Peter felt thirsty.
#8 Peter wanted to drink the water from the waterjug.
#9 Peter realized that he lost the waterjug in the camp.
#10 THUS Peter felt scared.
#11 Peter wanted to find the waterjug in the camp.
#12 Peter searched the waterjug in the camp.
#13 Peter found the waterjug in the camp.
#14 From then on, Peter learned to be responsible.
```

The evaluation process for PB-PA should follow the method employed by its predecessors, Picture Books 1 and 2, in order to provide a baseline for comparing the quality of the generated stories. Manual evaluation by three experts (literary writers and/or child educators) will be conducted. Because this research focused on the story planning process, the evaluation metrics will give importance on the content of the story plan rather than the surface text. The stories will be evaluated in terms of *coherence and cohesion; character, objects and background; and content*. A discussion of these criteria can be found in [1].

5 Comparative Analysis with Related Systems

The use of agents to guide the planning process is not new. The idea has been adapted from previous works, specifically the Virtual Storyteller [4]. Both systems provide their characters the ability to act believably in a story world under the guidance of the plot agent to achieve a well-structured plot.

PB-PA employed a more simplistic approach in handling its knowledge and its characters. Virtual Storyteller provided a more precise representation of its characters that is the basis for its corresponding goals and actions. Its characters may possess mixed emotions, such as *being happy* and *slightly sad*. This is modeled through a range of intensities, -100 to 100. On the other hand, PB-PA only models its emotion in bipolar forms, such as the character being sad or being happy.

Furthermore, since Virtual Storyteller is character-centric, the system prioritizes character believability in generating its stories [15]. Although PB-PA uses character agents, it prioritizes the generation of a story that adheres to the given theme. While characters have the choice to select their respective next actions, the plot agent often intervenes if the story is not progressing towards the fulfillment of the story theme.

The use of recommendations allows a PB-PA character to perform an action that may not be according to its current state when it is the last available option. Recommendations [6] provide a way for the planner to be able to generate stories that does

not strictly adhere to a character's state. Consequently, the recommendation plays an important role in computing for the heuristics of the generated stories.

Virtual Storyteller also utilizes episodic scripts to handle its global plot structure [19]. It has a set of goals and constraints that each episode should satisfy. The equivalent of the episodic scripts in PB-PA is the use of author goals. Each story theme has three author goals, which represents the premise of the story. An author goal represents a fabula element that must be performed by a specified character to proceed to the next author goal.

Compared to its predecessors, Picture Books 1 and 2, PB-PA provides a more flexible and logical approach in generating stories. Picture Books 1 [2] generated its stories using a predefined set of planning rules (author goals and character goals), thus hindering the system to generate story variants on the fly. Picture Books 2 [1] addressed this limitation by eliminating the need for a predefined set of author goals and instead relied solely on the causal chain of events as defined in its ontology. However, evaluation showed that this was not enough to generate logical stories. PB-PA addressed this through the adaptation of the agents that organize the process of generating the story plan. While all agents should be satisfied for an event to occur, the plot agent supervises the story progression to ensure that the story flows to the direction of the selected theme.

6 Conclusion

The task of the story planner involves making a myriad of decisions at each point of the story, from identifying a character action that is consistent with its trait to ensuring that the flow of the story progresses towards a coherent plot that attains the selected theme. Previous events that have occurred in the story world and have changed the state of the world must also be considered. The work presented in this paper showed that the interaction of three agents, namely the character agent, the plot agent, and the world agent, is essential to aid the story planner in its task, by managing the character representation, maintaining coherency in the story flow, and tracking the world state, respectively.

However, the preliminary results detailed in this paper only illustrate the successful execution of some of the functions of each agent. Given that the ontology is barely populated with sufficient storytelling knowledge, the candidate actions that can be performed are limited. Because of this, the tasks of the agents are yet to be evaluated.

References

1. Ang, K., Yu, S., Ong, E.: Theme-Based Cause-Effect Planning for Multiple-Scene Story Generation. In: Proceedings of the 2nd International Conference on Computational Creativity, Mexico City, Mexico, pp. 48–53 (2011)
2. Hong, A.J., Solis, C., Siy, J.T., Tabirao, E., Ong, E.: Planning Author and Character Goals for Story Generation. In: Proceedings of the NAACL Human Language Technology 2009 Workshop on Computational Approaches to Linguistic Creativity, pp. 63–70. ACL, Colorado (2009)
3. Oinonen, K., Theune, M., Nijholt, A., Uijlings, J.: Designing a Story Database for Use in Automatic Story Generation. In: Proceedings of the 5th International Conference Entertainment Computing, Cambridge, UK, pp. 298–301 (2006)

4. Theune, M., Faas, S., Nijholt, A., Heylen, D.: The Virtual Storyteller – Story Creation by Intelligent Agents. In: Göbel, S., Braun, N., Spierling, U., Dechau, J., Diener, H. (eds.) *Proceedings TIDSE 2003: Technologies for Interactive Digital Storytelling and Entertainment*, pp. 204–215. Fraunhofer IRB Verlag (2003)
5. Ang, K., Antonio, J., Sanchez, D., Yu, S., Ong, E.: Generating Stories for a Multi-Scene Input Picture. In: *Proceedings of the 7th National Natural Language Processing Research Symposium*, pp. 21–26. De La Salle University, Manila (2010)
6. Riedl, M.O., Young, R.M.: An Intent-Driven Planner for Multi-Agent Story Generation. In: *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-agent Systems*, pp. 186–193. IEEE Computer Society, Washington, DC (2004)
7. Ong, E.: The Art of Computer-Generated Stories. In: *Proceedings of the 2012 DLSU Arts Congress*. De La Salle University, Manila (2012)
8. Mann, W.C., Thompson, S.A.: *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report No. ISI/RS-87-190, Information Sciences Institute (1987)
9. Liu, H., Singh, P.: ConceptNet - A Practical Commonsense Reasoning Tool-kit. *BT Technology Journal* 22(4), 211–226 (2004)
10. Meehan, J.: Tale-Spin, An Interactive Program that Writes Stories. In: *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, Cambridge, M.A, pp. 91–98 (1997)
11. Turner, S.: *The Creative Process - A Computer Model of Storytelling and Creativity*. Cambridge University Press, New York (1994)
12. Perez, R., Perez, Y., Sharples, M.: Mexica: A Computer Model of a Cognitive Account of Creative Writing. *Journal of Experimental and Theoretical Artificial Intelligence* 13(2), 119–139 (2001)
13. Trabasso, T., Van den Broek, P., Suh, S.: Logical Necessity and Transitivity of Causal Relations in Stories. *Discourse Processes* 12(1), 1–25 (1989)
14. Schank, R.C.: The Primitive Acts of Conceptual Dependency. In: *Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, pp. 34–37. Association for Computational Linguistics, Stroudsburg (1975)
15. Kruizinga, E.E.: *Planning for Character Agents in Automated Storytelling*. MSc thesis, University of Twente, Netherlands (2007)
16. Riedl, M.: *Narrative Generation - Balancing Plot and Character*. Ph.D. Dissertation, Department of Computer Science, North Carolina State University (2004)
17. Egri, L.: *The Art of Dramatic Writing – Its Basis in the Creative Interpretation of Human Motives*. Wildside Press (2007)
18. Ortony, A., Clore, G., Collins, A.: *The Cognitive Structure of Emotions*. Cambridge University Press (1990)
19. Theune, M., Rensen, S., den op Akker, R., Heylen, D., Nijholt, A.: Emotional Characters for Automatic Plot Creation. In: Göbel, S., Spierling, U., Hoffmann, A., Iurgel, I., Schneider, O., Dechau, J., Feix, A. (eds.) *TIDSE 2004*. LNCS, vol. 3105, pp. 95–100. Springer, Heidelberg (2004)

A Framework for a Multi-agent Collaborative Virtual Learning Environment (MACVILLE) Based on Activity Theory

Nader Hanna and Deborah Richards

Department of Computing, Macquarie University, NSW 2109, Australia
{nader.hanna,deborah.richards}@mq.edu.au

Abstract. Virtual Learning Environments are increasingly used in education to encourage collaborative learning and engagement with the learning material. However, collaborative activity is often minimal and superficial. We propose a multi-agent collaborative virtual learning environment in which the agents support communication skills and collaboration among and with learners. Our framework is underpinned by Activity theory which is a theoretical model of human activity that reflects its collaborative nature.

Keywords: multi-agent, virtual environment, collaborative learning, activity theory.

1 Introduction

Collaborative learning is an educational approach based on the idea that learning is a social behavior that involves groups of learners working together as a team to find a solution to a problem or complete a required task. Johnson and Johnson [1] found that classroom learning improves significantly when students participate socially, interacting in face-to-face collaborative learning. Collaboration is broadly defined as the interaction among two or more individuals and can encompass a variety of behaviors, including communication, information sharing, coordination, cooperation, problem solving, and negotiation. Collaborative learning is much more efficient in learning situations that need discovery by allowing the learner to be actively involved through presenting ideas and defending them, being ready to be questioned about his beliefs and accepting the opinion of others; what is important in collaborative learning is that there should not be a right answer.

Computer supported collaborative learning (CSCL) may be defined as the educational use of computer technology to facilitate group learning. CSCL can increase student responsibility, initiative, participation, learning and higher grades, as well as increase communication with peers through discussion of course concepts [2]. 3D Virtual Learning Environments (VLE) can provide an engaging and immersive experiential learning experience. While the use of 3D VLE for collaborative learning is

compelling, collaborative activity is often minimal or superficial. We propose that collaboration can be facilitated via the inclusion of multi-agent technology where the agents become learning partners. Using Activity Theory (AT) as a foundation, we have designed the Multi-Agent Collaborative Virtual Learning Environment (MACVILLE). AT provides a number of useful concepts that can be used to analyze collaborative learning activities and to create a conceptual framework for collaboration between learners and agents.

In the next section we briefly present Activity Theory, followed by a review of related literature (section 3). In section 4 we introduce MACVILLE. Section 5 provides an instantiation of MACVILLE. Conclusions and future work appear in section 6.

2 Activity Theory

Activity Theory (AT) is a theoretical framework for analyzing human practices in a given context. According to AT, people are embedded actors (not processors or system components) with both individual and social levels interlinked at the same time. The origin of activity theory can be found in the early writings of Vygotsky (1896-1934), who suggests that social activity, the basic unit of analysis, may serve as an explanatory principle in regard to human consciousness.

Kuuti [3] sees three levels in an activity: activity, action, operation. Fig. 1 depicts the Hierarchical Levels of an Activity that describe the short-term processes that take place during the course of that activity. The activities level consists of actions or chains of actions. The first condition for any activity is the presence of a need. Needs stimulate but do not direct the activity.

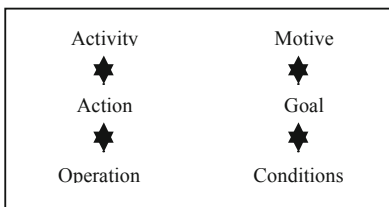


Fig. 1. Hierarchical Levels of an Activity

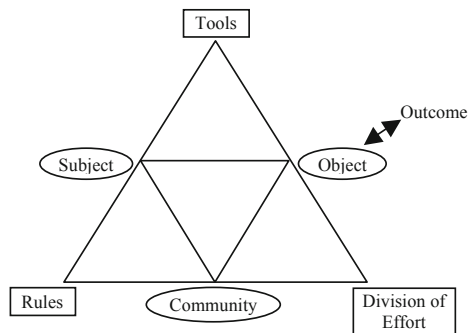


Fig. 2. Structure of Activity theory

In his general model, Engeström [4] asserts that human activity is “object-oriented, collective, and culturally mediated” [5] and composed of subject, object, actions, and operation, as shown in Fig. 2. The subject, object, and community interact through tools, rules, and a division of labor [6]. A subject is a person or a group engaged in an

activity. An object aids the subject and motivates activity [7]. The Subject uses Tools to interact with Object. Tools are used to mediate the activity. In the same way, Community uses Division of Effort to interact with Object, and uses Rules to interact with Subject. Internalization is a key concept where the subject uses tools seamlessly and automatically to execute actions which require conscious thought or planning. The Subject will make a plan according to his mental model of the real world and then s/he will use actions to achieve this plan. If actions are not performed according to the subject's plan, s/he will adjust them and retry to execute actions [3]. In summary, key AT concepts include: object-orientedness; hierarchical structure of activity; internalization/externalization; mediation and development.

3 Related Work

3D VLE have been used for collaborative learning. An early example is the Virtual European Schools (VES) project [8] that simulates a classroom where several students are allowed to navigate within the environment simultaneously; a text-chat facility enables students to communicate with each other. More recently, C-VISions [9] is a multi-user collaborative science education environment in which users interact with the environment to make scientific experiments. Monahan et al. [10] present a collaborative multi-user virtual reality web-based system called CLEV-R which provides communication tools to support collaboration among students. Liu et al. [11] present an agent-based approach to design and implement a virtual e-learning. The system includes student agent, teacher agent and instructor agent. The learner can log in the system and select learning material and discuss it with other learners or teacher.

CSCL research using AT includes the work by Gifford and Enyedy [12]. They propose a framework called Activity Centered Design (ACD). Their proposed framework is based on three main concepts of AT: a) that activity is mediated by cultural artifacts; b) that activity must be analyzed at various levels; and c) that internal activity (thinking) first occurs in the social plane (contextualized activity). In ACD, learners progress through activities as partial participants to full participants. Jonassen and Rohrer-Murphy [13] argue that AT provides a powerful framework for analyzing needs, tasks, and outcomes for designing constructivist learning environments. Zurita and Nussbaum [14] identified six steps to propose a conceptual framework for mobile Computer Supported Collaborative Learning (MCSCCL) activities. Liang et al. [15] further build on the six steps to define components and their relationships for collaborative network learning. Norris and Wong [16] use AT to identify any difficulties that users may have when navigating through QuickTime Virtual Reality Environments (QTVR). The authors use a technique called the Critical Decision Method (CDM) which relies on the user recalling memorable incidents while doing a certain task. CDM is used to provide data to Activity theory. Miao [17] presents a conceptual framework for the design of virtual problem-based learning environments in the light of activity theory. We build upon this framework.

4 MACVILLE Framework Based on Activity Theory

The proposed framework uses AT to analyze the components of activities which may take place in a collaborative multi-agent VLE. As described next and depicted in Fig. 3, each component has a part in the real world and another part in the virtual world.

4.1 The Components of the Virtual Activity

Using Activity Theory we may analyze the virtual activities that may happen between different agents and the human learner as follows:

Subject – subjects in the virtual world are the multi-agents in different locations. When the learners sign into the system, they will be able to visit different locations; in each location there will be an activity to do. The agent in each location will collaborate with learners in fulfilling the required task.

Tools – include the 3D graphic character, 3D character animation and agent reasoning model. The 3D graphic of the agent should be interesting and believable so the human learner would be encouraged to collaborate with these agents (characters). 3D character animation is related to the reasoning model of the agent. The motion or animation is a reflection of what the agent reasons or decides to do. The reasoning model of the agent will follow the model of Belief-Desire-Intention (BDI) which helps the agent in realizing the environment around him and carrying out the required activity; the agent should also realize the role of the learning in performing the activity and the intersection with his role. As the agent is going to collaborate with the learning the reasoning model of the agent should be integrated with both the social and collaborative model which enables the agent to be friendly and cooperative with the human user.

Community - is the environment in which the virtual activity is carried out and includes virtual communication that happens between the learners and the agent while performing the activity. Virtual communication includes interaction between the human learner and virtual agent. If the learners face some difficulty in doing the activity the agent will guide the learner to understand how to carry out the virtual activity.

Rules - are those controlling the performance of the activity in the virtual world. The initial rule is to determine the target of each activity the learner will participate in each location in the environment. Learners and agents should specify their roles in the activity, their roles should integrate together to fulfill the target determined in the first rule. The agent should be able to check the learners' behavior and progress in achieving the activity. If the agent detects inactivity, the agent may begin role-playing to encourage and direct learners to continue collaboration in the activity.

Division of Effort- is the division of roles in the virtual world. Human-agent group will be responsible for doing the virtual activity.

Object- the object in the virtual world would be to achieve the shared tasks between the learners and the agent; combining the objects of the virtual and physical activities will lead to reaching the final outcome of real world activity.

4.2 The Components of the Physical Activity

In considering the activities that may happen between the companion learners or the colleague learners in the physical or online communication, the components of AT may be represented as follow:

Subjects- the subject in the physical world are the learner companions who physically collaborate with the learner in exploring the virtual ecosystem, collecting data, finding evidence and writing the final report. Other subjects are learner colleagues, and they are other human learners who explore the virtual ecosystem, make their own conclusion and write their final report.

Tools- voice communication is the tool which is used between each learning companion; text message is the communication tool among online learner colleagues. Database and Internet protocols are technical tools used to connect different collaborative learning groups.

Community- the environment in which physical collaboration takes place includes face-to-face communication between collaborative learning companions who share the same physical place and interact. Together the learning companions explore the environment, discuss, adopt roles such as leader, provide guidance to one another, draw conclusions and complete activities such as writing and submitting reports. Each learning group will be able to see others' reports and comment on their conclusion. Online communication between learning groups may include debating with a colleague learner to defend a conclusion and negotiating a conclusion.

Rules - the rules that control the communication between learners may be divided into two sections: the first section is the rules that control the physical communication between the two learning companions who share the same computer and exist in the same physical place. These rules include discussion that may take place between the two learners, guidance of navigation and conclusion, decision making, role playing, knowledge sharing between the companion learners, note taking and making conclusion represented in the final report. The second section is the rules that manage the online communication among groups which may be located in different places. These rules may include justifying to each group their own decision, defending their idea if the other groups criticize their report, questioning to inquire about the reasons why other groups adopt a different point of view about the causes to the problem, comparison of the pretexts of other groups and their own reasons on which the conclusions were taken, making conclusion about whether the result of their report is reasonable enough or whether other groups have reached a better conclusion and finally updating their first conclusion. Updating the first conclusion could be by adopting the other group's conclusion or by persisting on their first one.

Division of Effort- human learners could be divided into learner-learner physical group and learner-learner online group. Collaboration learning takes place by discussion and guidance in the first case and by debating and negotiation in the second case.

Object- while the objective in the virtual world is to collect data and perform activity in order to be able to write the final report, the objective in the physical world is to collaborate with the companion learner in making the conclusion, and debate with colleague learners about their own conclusion.

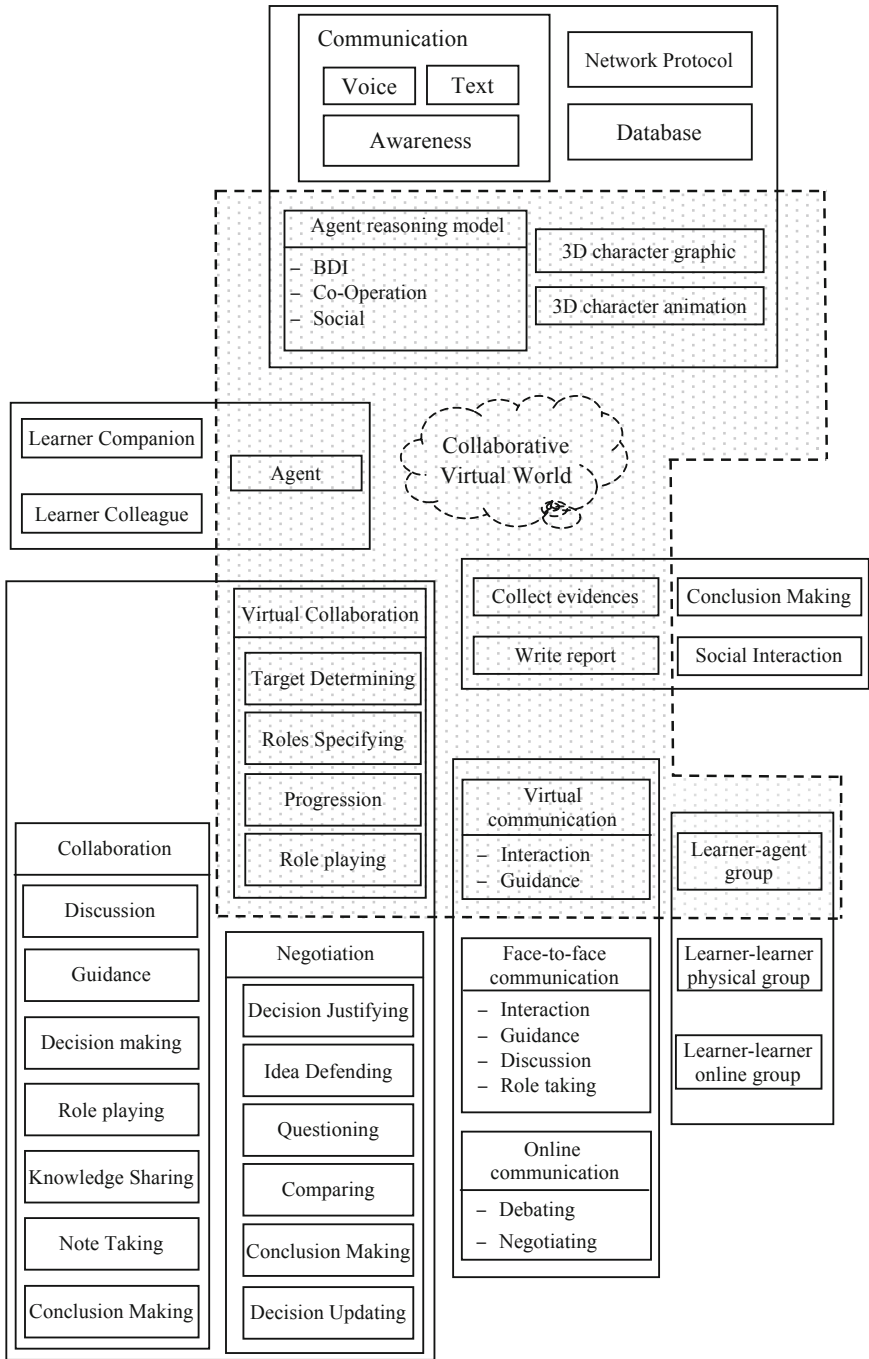


Fig. 3. The proposed framework for virtual collaborative learning based on Engestrom's expanded Activity Theory model

5 Agent Architecture in MACVILLE

A number of agent-based architectures for collaborative learning have been proposed [18-21]. Liang, Ruo and Bai [15] have also proposed the use of activity theory in their abstract architecture. However, these approaches tend to have simple multi-agent architectures comprised of agents with different roles, such as teacher, instructor or students. Incorporating the concepts, philosophy and assumptions on which AT is based, we propose that a multi-agent collaborative learning environment should contain components to handle the cognitive processes including memory and its related processes as well as the social and collaborative processes as shown in Fig. 4:

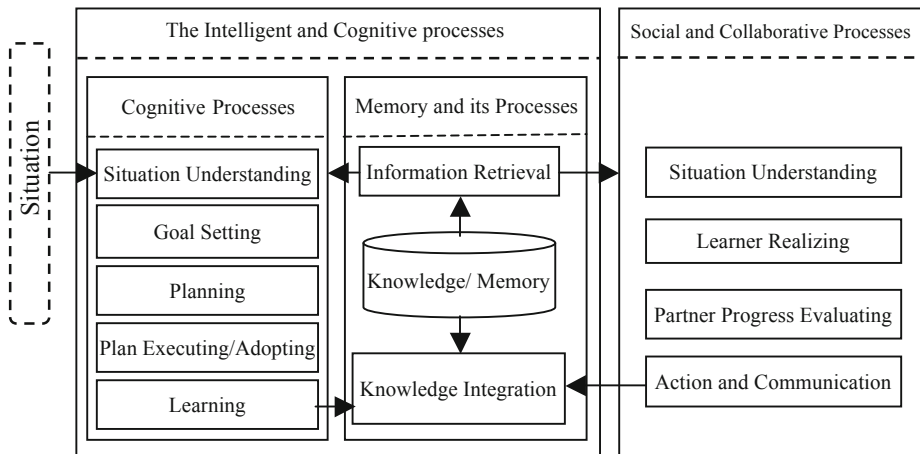


Fig. 4. Agent Architecture

The Intelligent and Cognitive processes. This component includes two processes: Memory and Cognitive. The memory or the knowledge base is where the agents store information, knowledge and experience. There are two processes related to the memory: knowledge Integration to add new experience to the stored knowledge, and Information retrieval to get the appropriate piece of information to the current situation. Retrieving information from the memory could be done during any process.

Cognitive Processes include the reasoning model that the agent has to perform in the situations s/he faces. The Cognitive Processes begin with Understanding the situation and determining if this situation has appropriate knowledge in the memory or it is a new situation and the agent will have to perform an inference on the situation. As the proposed framework of collaboration between the human learner and an agent includes a role to be played by the agent, the agent should have the ability to plan what activity to do and what is his share in this activity and what is human learner's role. After specifying the role the agent is going to play, s/he needs to plan how this share in the activity should be done. By knowing his share in the activity, the agent should have solutions for possible problems s/he may encounter. At the end of the

reasoning processes, the agent may add a piece of knowledge that is not in his memory for later use; this process is similar to learning.

Social and Collaborative Process. One of the most important elements of learning which is absent from a traditional text based learning system is social interaction with other learners. The agent in a collaborative learning environment should have the ability to socially interact with the human learner and encourage collaboration together. The social processes begin with understanding whether the social situation is competitive or collaborative. The agent may have a mechanism to identify the learning and the social properties of the learner. The Partner Progress Evaluation process is a continuous process that the agent should do during execution of the collaborative activity with the human learner, the agent should make sure that the learner is participating with him/her. The agent is going to evaluate the progress of the task relying on another two processes: planning process to determine the share of the agent and the learner and a process to acquire the properties of learner which may lead the virtual agent to adopt different evaluation criteria. At the end of the social and collaborative process the agent will need to take a social action such as encouraging the learner to do more effort, or congratulate the learner for his hard working.

5.1 Applying the Framework

In order to design and implement MACVILLE we follow the six-step methodology designed by Zurita and Nussbaum [14]. We apply these steps to an existing VLE we have created that currently only supports collaboration in the real world and only between companion learners and not among learning groups; interaction and activities in the virtual world are not collaborative. Our VLE is an ecosystem for an imaginary island called Omosa created to help students learn scientific knowledge and science inquiry skills. The goal is to determine why the fictitious animals, known as Yernt, are dying out. The island consists of four different locations the learner can visit. In each location there will be a virtual agent waiting for the visit of the group of companion learners to ask their help in doing a specific activity which at the end will give the learners data, material and evidence that may all them to hypothesize the cause of the problem on the island. There will be a manager agent to welcome the learning group and introduce the problem facing the island. The group members will collaborate to explore the island and visit the four different locations. The four locations are village, research lab, hunting ground and weather station. Inside each location group members will meet a different agent and they will have to collaborate with him/her in collecting data and evidence to understand the cause of the changes in the ecosystem of the island. In the village, group members will meet an agent who will collaborate with them in collecting data from the village about its past and current state. In the hunting ground, group members will encounter another agent with whom they can count the numbers of prey, predator and human hunters. The hunting-ground agent gives learners clues about past numbers. In the location of the research lab, the researcher agent will ask for the help of the learner in getting samples from soil cores to examine the earth for pollen grains and char deposits. In the weather station location,

group members will encounter a climatologist agent who will ask their help in measuring today's temperature and rain level and write it down in a certain table where they can see past measurements about temperature and rain level.

After exploring the virtual world and collecting data and evidence from the imaginary island, the manager agent will ask the group members to write a report that summarizes their conclusion about what is the cause of the changes in the ecosystem of the island (Omosa) and what is the reason the imaginary animal (Yernt) are dying out. After finishing writing the report and submitting it online the group of companion learners will be transferred to a window where they can see other groups' reports and they can send text messages or speak online to the other groups in order to discuss the conclusions they reached and defend their own conclusion. As this activity has no specific right answer, each group could adopt another group's conclusion if they find it is more persuasive than theirs, or they could defend and retain their conclusion.

In this context, we now apply the six steps [19] to Omosa as follows:

1. Characterize collaborators. The learners will be grade 9 students of both genders. The learners will be divided into groups of two on each computer, so they can collaborate while navigating the virtual world and in the same time keep the number of the group small to keep them focus on the learning activity.
2. Define the global educational objective. The high level aims are to learn the processes in scientific inquiry including data collection and hypothesis testing and to learn concepts from biology particularly concerning ecosystems.
3. Establish the desired social interaction skills. The social aim is to practice face-to-face communication among the two learning companions, online communication among different learning groups, and virtual communication among the learning companion and the virtual agent. Specific social skills will include discussion, negotiation, role-taking, etc. as identified in Fig. 3.
4. Identify the type of collaborative activity. The type of collaborative activity needed to teach the student the educational skills will involve drawing conclusions and (virtual) communication amongst learning companions, groups and virtual agents to teach the social skills.
5. Define activity tasks. Learners perform a variety of tasks, either learner-agent task, learner-learner physical task, learner-learner online task.
6. Define the roles and rules. Types of social relationships among companion learners can include guidance from one learner to the other and discussion between the two companion learners to find out one or more possible scientific conclusions to the problem presented by the virtual world. Types of social interaction between both companion learners and the virtual agent include guidance from the agent to the learner in case the learner stops interacting in the virtual environment, and social interdependency relation as both the companion learners and the virtual agent will be involved in doing activities in the virtual virtual ecosystem. The types of social interaction among the learning groups of colleagues may include cognitive conflict when groups have different conclusions; each learning group will try to negotiate with other learning colleagues to defend their own report results.

5.2 Collaboration in MACVILLE-Omosa

MACVILLE-Omosa has six agents (manager agent, hunting ground agent, climatologist agent, village agent, research lab agent and virtual animal agents) that are going

to interact with the learners. Additionally there are other agents that perform pre-determined tasks in the background. They do not directly communicate with the learners but their existence adds more believability to the environment (such as villagers who talk and interact with each other but not with learners). The multi-agent virtual collaborative activity diagram is presented in Fig. 5. Note that while we anticipate that each companion group will complete every activity in each of the four locations, it is possible that groups could divide up the activities. If multiple groups/processors were available the activities can be conducted in parallel and in any case all activities must be completed before the report can be written; hence, the use of synchronisation bars.

- **Manager agent:** The role of manager agent is to give hints to the learners about the problem that faces the island, and direct them to the different locations where they may find clues to make a scientific conclusion about the cause of the problem.
- **Hunting ground agent, climatologist agent, village agent, research lab agent:** They are the four agents that the learners are going to meet in different scenes on the island and each agent will collaborate with learners to do a different activity which will provide more information about what has changed the ecosystem of the island.
- **Animal Agents:** They are the flocks of prey and predators, in one of the tasks the learner will have to count the number of individual animals and find a possible relation between the numbers of prey and predators in the present and past time.

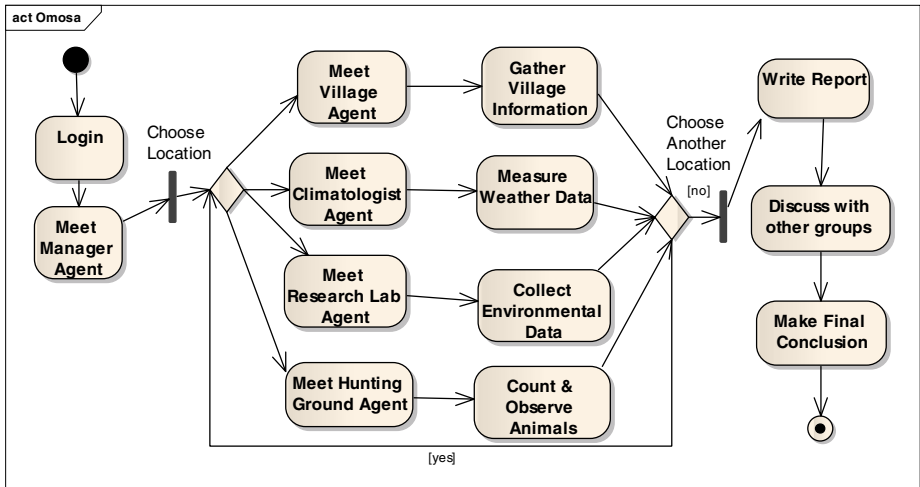


Fig. 5. Multi-agent activity diagram

6 Evaluating MACVILLE

The 3D part of the virtual world has already been designed and implemented [22]. MACVILLE extends the current system to provide inworld/virtual collaboration involving virtual agents. After completing the implementation, three types of evaluation will be conducted to measure:

1. *Skills in making scientific conclusions*: To evaluate the effect of the collaborative virtual environment on making scientific query, a pretest and posttest technique will be applied to test how much the scientific query skill has developed.
2. *Social skills involving interaction with the virtual agent*: As part of the Social and Collaborative Processes in the architecture of the agent, there is a process called Partner Progress Evaluation in which the agent will evaluate learner collaboration in the virtual task by using some parameters such as idle time of learner, incorrect responses from learner, and duration of time spent in each activity/location.
3. *Social skills involving interaction among human learners*: As a way to evaluate the development of the social skills among learners as a result of human-human interaction, a pre and post assessment of learners' attitude toward collaborative learning and companion learner will be made, also data related to collaboration of learner will be gathered and learning session will be video and audio recorded and evaluated using DFCS coding [23]. Further to measure 2 and 3, we have also created methods for automatic data acquisition of user models and system usability [24].

7 Conclusion and Future Work

This paper presented the design of a framework for a multi-agent collaborative virtual learning environment based on Activity Theory which is used to design the activities in both the virtual and the physical world. Also, the paper provides the design of an agent architecture that is appropriate for this special type of virtual environment characterized as containing cognitive and social/collaborative cores. The proposed framework builds on an earlier framework for collaborative virtual environment based on AT [17]. Future work will include implementing the reasoning model of the collaborative agent and group-group collaboration plus evaluation of the system.

References

1. Johnson, R.T., Johnson, D.W.: *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning*. Allyn and Bacon, Boston (1999)
2. Brandon, D.P., Hollingshead, A.B.: *Collaborative Learning and Computer-supported Groups*. *Communication Education* 48, 109–126 (1999)
3. Kuutti, K.: *Activity Theory as a Potential Framework for Human-Computer Interaction Research*. In: Nardi, B.A. (ed.) *Context and Consciousness: Activity Theory and Human-Computer Interaction*. MIT Press, Cambridge (1996)
4. Engeström, Y.: *Knotworking to Create Collaborative Intentionality Capital in Fluid Organizational Fields*. *Collaborative Capital: Creating Intangible Value* 11, 307–336 (2005)
5. Engeström, Y., Mietinen, R.: *Perspectives on Activity Theory*. Cambridge University Press, New York (1999)
6. Engeström, Y.: *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Orienta-Konsultit Oy, Helsinki (1987)
7. Kaptelinin, V., Kuutti, K., Bannon, L.: *Activity Theory: Basic Concepts and Applications*. In: Blumenthal, B., Gornostaev, J., Unger, C. (eds.) *EWHCI 1995*. LNCS, vol. 1015, pp. 189–201. Springer, Heidelberg (1995)

8. Bouras, C., Fotakis, D., Kapoulas, V., Koubek, A., Mayer, H., Rehatschek, H.: Virtual European School-VES. In: IEEE International Conference on Multimedia Computing and Systems (ICMCS 1999), pp. 1055–1057 (1999)
9. Chee, Y.S., Hooi, C.M.: C-VISions: Socialized Learning through Collaborative, Virtual, Interactive Simulations. In: Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community, pp. 687–696. International Society of the Learning Sciences, Boulder (2002)
10. Monahan, T., McArdle, G., Bertolotto, M.: Virtual Reality for Collaborative E-learning. *Computers & Education* 50, 1339–1353 (2008)
11. Zhi, L., Hai, J., Zhaolin, F.: Collaborative Learning in E-Learning based on Multi-Agent Systems. In: 10th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2006), pp. 1–5 (2006)
12. Gifford, B.R., Enyedy, N.D.: Activity Centered Design: Towards a Theoretical Framework for CSCL. In: Roschelle, C.H.J. (ed.) Proceedings of the 1999 Conference on Computer Support for Collaborative Learning, pp. 22–37. International Society of the Learning Sciences, Palo Alto (1999)
13. Lim, C.P., Hang, D.: An Activity Theory Approach to Research of ICT Integration in Singapore Schools. *Computers & Education* 41, 49–63 (2003)
14. Zurita, G., Nussbaum, M.: A Conceptual Framework Based on Activity Theory for Mobile CSCL. *British Journal of Educational Technology* 38, 211–235 (2007)
15. Liang, X., Wang, R., Bai, G.: A Multi-Agent System Based on Activity Theory for Collaborative Network Learning. In: First International Workshop on Education Technology and Computer Science (ETCS 2009), pp. 392–397 (2009)
16. Norris, B.E., Wong, B.L.W.: Activity Breakdowns in QuickTime Virtual Reality Environments. In: Proceedings of the First Australasian User Interface Conference (AUIC 2000), pp. 67–72. IEEE Computer Society Press, Canberra (2000)
17. Miao, Y.: An Activity Theoretical Approach to a Virtual Problem Based Learning Environment. In: The 2000 International Conference on Information in the 21 Century: Emerging Technologies and New Challenges, pp. 647–654 (2000)
18. Chuan, Z., Jianqing, X., Xiangsheng, Y.: An architecture for intelligent collaborative systems based on multi-agent. In: 12th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2008), pp. 367–372 (2008)
19. Liu, Y., Chee, Y.S.: Intelligent Pedagogical Agents with Multiparty Interaction Support. In: Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2004), pp. 134–140. IEEE Computer Society (2004)
20. Marin, B.F., Hunger, A., Werner, S., Meila, S., Schuetz, C.: A Generic Framework for an Interface Tutor Agent within a Virtual Collaborative Learning Environment. In: Proc. of IEEE International Conference on Advanced Learning Technologies, pp. 31–35 (2004)
21. Liu, X., Peng, G., Liu, X., Hou, Y.: Development of a Collaborative Virtual Maintenance Environment with Agent Technology. *Jrnl of Manufacturing Systems* 29, 173–181 (2010)
22. Jacobson, M.J., Richards, D., Kapur, M., Taylor, C., Hu, T., Wong, W.-Y., Newstead, A.: Collaborative Virtual Worlds and Productive Failure: Design Research with Multi-disciplinary Pedagogical. In: Technical and Graphics & Learning Research Teams. Proc. Comp. Supported Collaborative Learning (CSCL 2011), Hong Kong, pp. 1126–1129 (2011)
23. Poole, M.S., Holmes, M.E.: Decision Development in Computer-Assisted Group Decision Making. *Human Communication Research* 22, 90–127 (1995)
24. Hanna, N., Richards, D., Jacobson, M.J.: Automatic Acquisition of User Models of Interaction to Evaluate the Usability of Virtual Environments. In: Richards, D., Kang, B.H. (eds.) PKAW 2012. LNCS (LNAI), vol. 7457, pp. 43–57. Springer, Heidelberg (2012)

Emergence of Personal Knowledge Management Processes within Multi-agent Roles

Shahrinaz Ismail¹ and Mohd Sharifuddin Ahmad²

¹Universiti Kuala Lumpur, Malaysian Institute of Information Technology, Malaysia
shahrinaz@miit.unikl.edu.my

²Universiti Tenaga Nasional, College of Graduate Studies, Malaysia
sharif@uniten.edu.my

Abstract. In conceptualising a multi-agent reputation point system, we discovered that software agents' roles were similar to human's common processes of personal knowledge management (PKM), namely get/retrieve knowledge, understand/analyse knowledge, share knowledge and connect to other knowledge (GUSC). The proposed reputation point system entails that the 'connect' process is mediated, but the other three were found to be within and related to that process. This paper discusses the emergence of personal intelligence within the roles assigned to software agents in mediating the PKM of their human counterparts. Recommendation of future work includes an agent simulation to prove this emergence within the assigned roles.

Keywords: Multi-Agent System, Reputation Point, Personal Knowledge Management, Personal Knowledge Network.

1 Introduction

Recent researches in agent-mediated knowledge management (KM) have diversified into the more focused aspect of KM, which is personal knowledge management (PKM) [1, 2, 3, 4]. In investigating PKM, it is found that knowledge workers connect to other knowledge workers within and outside of the organisations in order to get, understand, and share knowledge to improve their work performance. Subsequently, this expands to another focused domain called the personal knowledge network, which is introduced in the multi-agent reputation point framework.

Researches in PKM suggest the existence of a series of processes called get/retrieve knowledge, understand/analyse knowledge, share knowledge, and connect to other knowledge sources and software agents are proposed to mediate these processes on behalf of their human counterparts. However, there is a difference between the human aspects of PKM processes and the human-agent PKM processes when the system view of organisational KM is included. The question in mind is: if it is possible for agents to have their own sequence of PKM processes, how should this be implemented?

This paper proposes a multi-agent reputation point system framework and proposes the process flow of multi-agent PKM roles based on GUSC model. The GUSC model is derived from the preliminary study on PKM processes among knowledge workers in the Malaysian context, which is proposed to be implemented in agent-mediated domain.

2 Related Works

It is important to understand the theoretical background of this study to further enlighten the future research in the domain of agent-mediated PKM where the need of multiple roles in agent-mediation is found significant. This section includes the two main areas of research that are closely related to this study: personal knowledge management (PKM) processes; and software agents and human-agent systems.

2.1 Personal Knowledge Management (PKM) Processes: Theory to Technical

McFarlane [5] suggested that a good definition for PKM by Gorman and Pauleen [5], who state that PKM is “an evolving set of understandings, skills and abilities that allows an individual to survive and prosper in complex and changing organisational and social environments”. It is argued that this definition is suitable due to the idea of being ‘personal’ in KM, suggesting that the main focus is on the individual, his or her understanding of knowledge, and the classification and systematisation of that knowledge for personal and professional growth. From this definition, some pattern of ‘systemisation’ is revealed from the social sciences aspect of the PKM theory. It brings the adoption of PKM concepts into the application of information and communication technology.

Considering that PKM encompasses an understanding of “the idea of knowledge applied to individual tasks and needs” [5] and concurrent with the emergence of Web 2.0 tools since the past decade, a number of studies investigate the technicality of the PKM processes. Apart from the previous works on PKM over Web 2.0 [6, 7, 10], recent works looked into the automation of PKM processes using agent technology with the concept of GUSC (Get-Understand-Share-Connect) Model [1, 3]. The evolution of PKM processes from theory to technical in the past decade progressed from total conceptual theory of human practice [5, 8, 9, 11] (shown in italics in Table 1) to application of the concept on existing behaviour identified over Web 2.0 tools usage [6, 7, 10], and eventually to the current study of agent-mediated PKM processes [1, 3]. Table 1 shows how the processes under the theoretical and application of PKM processes over Web 2.0 matches the current GUSC Model. Table 1 also shows the fundamentals of ‘get/retrieve’ and ‘understand/analyse’ knowledge in the PKM processes, which are both theoretically and technically significant. On the other hand, the processes of ‘share’ and ‘connect’ to other knowledge (people and/or artefact) on emergence is significant with the usage of technology in general, and Web 2.0 in specific. The agent-mediated PKM is suggested to take ‘share’ and ‘connect’ processes to ensure that the fundamental processes are implementable.

Table 1. PKM Processes from Theory to Technical

Authors (according to year)	Significant PKM Processes			
	Get / Retrieve	Understand / Analyse	Share	Connect
McFarlane (2011) [5]	<ul style="list-style-type: none"> • Acquired • Readily Retrieved • Stored 	<ul style="list-style-type: none"> • Internalised • Categorised • Classified 		<ul style="list-style-type: none"> • Readily Accessed
Jarche (2010) [6]	<ul style="list-style-type: none"> • Aggregate 	<ul style="list-style-type: none"> • Understand 		<ul style="list-style-type: none"> • Connect
Razmerita, Kirchner & Sudzina (2009) [7]	<ul style="list-style-type: none"> • Create 	<ul style="list-style-type: none"> • Codify • Organise 	<ul style="list-style-type: none"> • Share • Collaborate 	
Verma (2009) [8]	<ul style="list-style-type: none"> • Find 	<ul style="list-style-type: none"> • Learn • Explore 		<ul style="list-style-type: none"> • Connect
Grundspenkis (2007) [9]	<ul style="list-style-type: none"> • Gather • Search • Retrieve • Store 	<ul style="list-style-type: none"> • Classify 		
Pettenati, et al. (2007) [10]	<ul style="list-style-type: none"> • Create 	<ul style="list-style-type: none"> • Organise 	<ul style="list-style-type: none"> • Share 	
Avery, et al. (2001) [11]	<ul style="list-style-type: none"> • Retrieve 	<ul style="list-style-type: none"> • Evaluate • Organise • Analyse 	<ul style="list-style-type: none"> • Collaborate • Present 	<ul style="list-style-type: none"> • Secure

Table 2. Software Agents Characteristics and Capabilities

Authors	Definition of Software Agents	Application in GUSC
Coen (1991) [12]	Programs that <i>engage in dialogs and negotiate and coordinate the transfer</i> of information	Get, Share, Connect
Russel and Norvig (1995) [13]	Anything that can be viewed as perceiving its environment <i>through sensors and acting</i> upon that environment through effectors	Connect
Gilbert, et al (1995) [14]	Software entities that <i>carry out some set of operations</i> on behalf of a user or another program with some degree of independence or autonomy, and in so doing, <i>employ some knowledge or representation</i> of the user's goals or desires	Understand
Maes (1995) [15]	Autonomous agents are computational systems that inhabit some complex dynamic environment, <i>sense and act</i> autonomously in this environment, and by doing so <i>realise a set of goals or tasks</i> for which they are designed	Understand, Connect
Jennings, et al. (2000) [16]	An encapsulated computer system that is situated in some environment and that is capable of <i>flexible action</i> in that environment in order to meet its design objectives	Get, Understand, Share, Connect
Ali, Shaikh and Shaikh (2010) [17]	Computational systems that inhabit some complex dynamic environment; <i>sense and act</i> autonomously in this environment and by doing so <i>realise set of goals or task</i> for which they are designed	Understand, Connect

2.1 Software Agents and Human-Agent Systems

Software agents are considered to have the capabilities that PKM processes would need. Table 2 concisely summarises the capabilities of software agents that allow further exploration of agent-mediated PKM processes, especially in defining the role of GUSC for the system based on the definitions given by authors in the past two decades. This summary does not consider the BDI (Belief-Desire-Intention) notions of software agents, since those are the intangible aspects of the agent-mediated system not discussed here.

The role of agents is applicable in the perspective of social intelligence, which operates within a social network where “agents search the space within the control of the agent-mediated system” [3]. In this environment, socialisation between software agents and their human counterparts is consistent with the concept of SECI model interactions by Nonaka and Takeuchi [18]. The forms of knowledge (i.e. explicit and tacit knowledge) are interchanged within the interactions that occur between humans, between human and agent, and between agents as shown in Table 3.

Table 3. Social interactions within agent environment

SECI Interactions	Social Interactions within Agent Environment
Externalisation Tacit → Explicit	Human → Agent The task of finding the knowledge expert is mediated by an agent, when the knowledge seekers SHARES by passing the messages and documents to the agent in explicit form.
Combination Explicit → Explicit	Agent → Agent Agent GETS the messages and documents from other agents, in explicit form.
Internalisation Explicit → Tacit	Agent → Human The knowledge seeker UNDERSTANDS the messages and documents found by the agents.
Socialisation Tacit → Tacit	Human → Human The knowledge seeker and the knowledge expert (the agents' human counterparts) CONNECT to each other.

In supporting the human-agent interactions shown in Table 3, a meta-level approach to building agent systems is suggested [3], which requires a comprehensive analysis of humans' and agents' functions. A novel technique is suggested for building multi-agent systems based on the concept of nodes, where a human entity is conceived (e.g. a knowledge worker), working cooperatively with a software agent in a virtual workspace called a node [3]. “A node consists of a knowledge worker and one or more agents, also known as role agents, to perform some roles of the knowledge worker”, and the knowledge worker has a set of functions, some of which could be delegated to the agents [3]. This work is supported by the analysis of human-human interactions, which reveals the two types of function of a knowledge worker: common functions (e.g. open document, create/edit document, upload/download document, delegate role, request, request response); and unique functions based on the knowledge held (e.g. analyse problem, propose solutions, response-to-request).

3 GUSC Roles in Reputation Point System Framework

A multi-agent reputation point system is proposed in the current study, which is the basis of this research. The methodology for this research is confined to the analysis on the process flow design of the reputation point system based on related works. In other words, the literature and design analyses are triangulated to conclude the findings of an emerging PKM processes within the roles assigned to agents.

The multi-agent reputation point system is based on one of the PKM processes, which is 'connect to other knowledge', elaborated from a preliminary case study on a researcher's PKM, where people connection and networking is vital in locating knowledge experts and that connecting to others is an important process in research [1]. "The basic concept of formalising software agents to mediate the work processes usually handled by the human counterpart is by having him/her to delegate the tasks to software agents" [1]. In order to understand the 'connect' process, this system is separated into three main processes, based on the preceded interview survey analysis: (i) identify knowledge source (as shown in Figure 1), (ii) search the network and merit reputation point (as shown in Figure 2), (iii) initiate connection to knowledge source (as shown in Figure 3).

The 'get knowledge' process suggested in the PKM model has a similar process flow as Figure 1, until the initial stage of Figure 2. Compared to Figure 1, the agent with the role of 'getting the knowledge' (or 'GET agent') needs to check the log for a similar task done previously by the system. If the 'knowledge' is similar to previously found, the agent reports the result back to the base, with a query whether the human knowledge seeker would like to get the latest updates for less than a month old (or any other duration justified to be suitable as default); the agent continues to look for updates for existing results that are older than this default duration.

Figure 1 includes a subprocess called 'analyse,' which consists of a series of discrete processes that the agent performs like compare or match keywords or criteria of knowledge expert before deciding or choosing on a knowledge source. For new task that does not match any records in the log, the agent searches the network to find the 'explicit knowledge', and further search the Internet if the 'knowledge' does not exist in the organisational network.

In Figure 2, where documented 'knowledge' is found and 'analysed', the GET agent's task is completed once the analysis returns with some results worth reporting back to the human knowledge seeker. From the point where GET agent achieved its goal, the CONNECT agent continues (if the task to connect is assigned by the human counterpart) with the merit of reputation point (shown in Figure 2 in dotted box).

A reputation point means a merit point assigned to a profiled knowledge expert based on reviews by and/or references from others. In other words, whatever happens in this area in Figure 2, splitting of work involving multiple agents takes place: the GET agent analyses the explicit knowledge and reports back to the base, and the CONNECT agent analyses the reputation point. The remaining process after this juncture is for the CONNECT process of PKM.

In 'understanding knowledge', the reputation point system includes the 'analyse' segment in its process flow, in both Figures 1 and 2. 'Analyse' compares or matches

keywords or criteria of ‘knowledge’, which can be explicit knowledge or knowledge expert. This depends on the location of the ‘analyse’ stage required in the process flow. It can be a role of an independent UNDERSTAND agent, or a role that belongs to the GET and CONNECT agents.

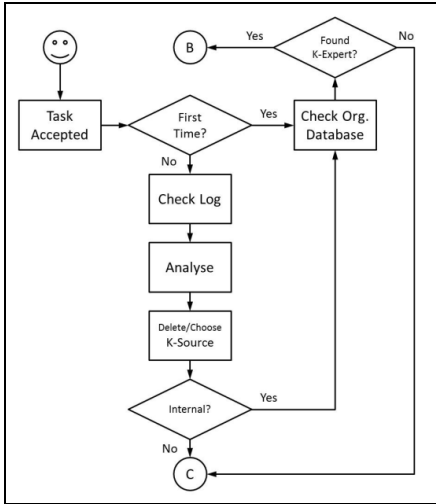


Fig. 1. Identify knowledge source

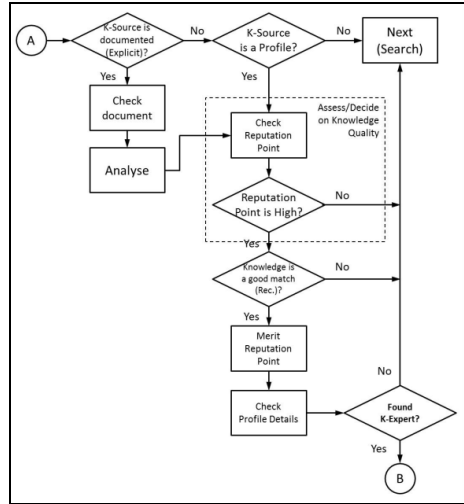


Fig. 2. Search the Internet and Merit Reputation Point

The concept of ‘sharing knowledge’ is very subjective in real life, since it is totally relied upon the behaviour of the human knowledge experts’ willingness to share with the community. In Figure 3, there is another agent with a ‘sharing’ role (which we called as SHARE agent) that ‘advertises’ their human counterpart’s profile. The SHARE agent performs the role of sharing updates in the profile and extends this updates to the rest of the nodes. For example, knowledge seekers or experts who have tried to connect to this profiled agent may be interested to get updates from this SHARE agent, since the (nodal) connection is logged. When there is an update from the knowledge expert, the SHARE agent retraces the connection trials previously made by the CONNECT agents, by checking the log and sends the updates to the base of those CONNECT agents.

4 Discussions

From the findings, it is shown that the reputation point system embeds the GUSC framework within it by assigning multiple agents that carry different roles similar to the GUSC concept in PKM. As agents are expected to be ‘intelligent’ enough to perform the roles they are assigned to, it is believed that a form of ‘personal intelligence’ emerges in these roles. While a human knowledge worker manages

his/her personal knowledge, the multi-agent system could mediate the whole process, even though the initial idea (of designing the reputation point system) is to focus on only one of the four processes.

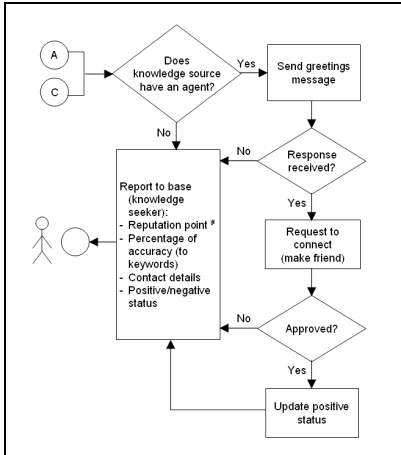


Fig. 3. Initiate connection to knowledge source

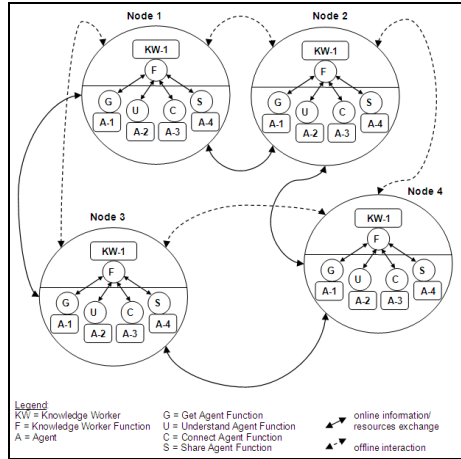


Fig. 4. Multiple nodes replicated from a single multi-agent environment for PKM processes

Figure 4 shows an ecosystem of agents that carry different roles in fulfilling the tasks of managing personal knowledge on behalf of their human counterparts. In mediating a human knowledge seeker, the roles include GET, UNDERSTAND, and CONNECT, whereas the SHARE role may be useful in mediating the task of a human knowledge expert. Both knowledge seeker and knowledge expert can be the same human who manages personal knowledge and is willing to share via his/her profile (i.e. SHARE role agent).

The online information/resources exchange include interactions and necessary communication across the nodes for fulfilling the CONNECT role and SHARE role agents' tasks. The offline interactions include any offline communication between the humans that are carried out after the role agents have fulfilled their tasks in connecting them.

5 Conclusion

By narrowing down the details of each segment of the reputation point system into agents' roles, the further work to be carried out is to develop the primitives based on these roles: GET, UNDERSTAND, SHARE, and CONNECT. With the well-designed primitives, a simulation of such processes will be made using the BDI (Belief-Desire-Intention) architecture because the intelligence of the GUSC agents may highly depends on the strong notions expected to be the characteristics of the software agents.

References

1. Ismail, S., Ahmad, M.S.: Personal Intelligence in Collective Goals: A Bottom-Up Approach from PKM to OKM. In: 7th International Conference on IT in Asia (CITA 2011), Kuching, Malaysia, pp. 265–270 (2011)
2. Ismail, S., Ahmad, M.S.: Modelling Social Intelligence to Achieve Personal Goals in Agent-mediated Personal Knowledge Management. In: International Workshop on Semantic Agents 2011 (AIW 2011), pp. 11–17. MIMOS/UNITEN, Malaysia (2011)
3. Ismail, S., Ahmad, M.S.: Emergence of Social Intelligence in Social Network: A Quantitative Analysis for Agent-mediated PKM Processes. In: ICIMu 2011 Conference. UNITEN, Malaysia (2011)
4. Ismail, S., Ahmad, M.S.: Modeling Social Intelligence to Achieve Personal Goals in Agent-mediated Personal Knowledge Management. In: ICRIIS 2011, Kuala Lumpur, Malaysia (2011)
5. McFarlane, D.A.: Personal Knowledge Management (PKM): Are We Really Ready? *JKMP* 12(3) (2011), <http://www.tlainc.com/article1270.htm>
6. Jarcho, H.: PKM in 2010. Harold Jarcho - Life in perpetual Beta (2010), <http://www.jarcho.com/2010/01/pkm-in-2010>
7. Razmerita, L., Kirchner, K., Sudzina, F.: Personal Knowledge Management: The Role of Web 2.0 Tools for Managing Knowledge at Individual and Organisational Levels. *Online Information Review* 33(6), 1021–1039 (2009)
8. Verma, S.: Personal Knowledge Management: A Tool to Expand Knowledge about Human Cognitive Capabilities. *IACSIT International Journal of Engineering and Technology* 1(5), 435–438 (2009)
9. Grundspenkis, J.: Agent based approach for organization and personal knowledge modeling: Knowledge management perspective. *Journal of Intelligent Manufacturing* 18(4), 451–457 (2007)
10. Pettenati, M.C., Cigognini, E., Mangione, J., Guerin, E.: Using Social Software for Personal Knowledge Management in Formal Online Learning. *Turkish Online Journal of Distance Education - TOJDE* 8(3), 52–65 (2007)
11. Avery, S., Brooks, R., Brown, J., Dorsey, P., O'Connor, M.: Personal knowledge management: framework for integration and partnerships. In: Small Computer Users in Education Conference, South Carolina, United States, pp. 29–39 (2001)
12. Coen, M.H.: SodaBot: A Software Agent Construction System. MIT AI Laboratory, Cambridge (1991)
13. Russell, S., Norvig, P.: *Artificial Intelligence: A Modern Approach*. Prentice-Hall, Englewood Cliffs (1995)
14. Gilbert, D., Aparicio, M., Atkinson, B., Brady, S. et al.: *IBM Intelligent Agent Strategy, White Paper* (1995)
15. Maes, P.: Artificial Life Meets Entertainment: Life like Autonomous Agents. *Communications of the ACM* 38, 108–114 (1995)
16. Jennings, N.R., Faratin, P., Lomuscio, A.R., Parsons, S., Sierra, C., Wooldridge, M.: Automated Negotiation: Prospects, Methods and Challenges. *International Journal of Group Decision and Negotiation*, 1–30 (2000)
17. Ali, G., Shaikh, N.A., Shaikh, A.W.: A Research Survey of Software Agents and Implementation Issues in Vulnerability Assessment and Social Profiling Models. *Australian Journal of Basic and Applied Sciences* 4, 442–449 (2010)
18. Nonaka, I., Takeuchi, H.: *The knowledge creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press, New York (1995)

Towards an Ontology-Based Approach to Knowledge Management of Graduate Attributes in Higher Education

Amara Atif, Peter Busch, and Deborah Richards

Department of Computing, Faculty of Science, Macquarie University, NSW 2109, Australia
{amara.atif, peter.busch, deborah.richards}@mq.edu.au

Abstract. Knowledge around graduate attributes (GAs) is an area in need of knowledge management strategies. GAs are the qualities and skills that a university agrees its students should develop during their time with the institution. The importance of GAs and ensuring they are embedded and assessed is widely accepted across higher education. This research paper uses Grounded Theory and Network Maps to gain insights into the issues of similarities and differences in the discourse around our sample universities. To cover these two perspectives, we had two researchers involved in data analysis, one with the goal of distilling key ideas and identifying similarities and the other with the goal of untangling and drawing out differences. There is no unified taxonomy of managing GAs. The motivation to create such ontology is to push the standardization process that will enable the connection among academic systems and improve educational workflows, communication, and collaboration between universities.

Keywords: Graduate Attributes, Grounded Theory, Knowledge Management, Ontologies, Higher Education.

1 Introduction

Just as the world's financial and economic situation is changing very quickly, the requirements for higher education are growing very quickly [1]. Therefore, actively managing the comparatively large amount of knowledge in a university is a complex and subtle process that involves priorities, needs, tools, and administrative support components [2]. Branin [3] states that Knowledge Management (KM) has been applied to the education industry since the post-World War II and sputnik era of 1950 to 1975. According to Petrides & Guiney [4] the use of KM in the Higher Education (HE) sector enables the encouragement of practical know-how, and effectiveness of educational institution management. KM in higher education also offers the benefits of a practical assessment framework that depends on the effectiveness of information management [4].

The knowledge needed by an organisation is closely related to the mission and activities of the organisation. In the case of universities, knowledge plays an obvious and central role; the core mission can be summed up as knowledge creation (research) and dissemination (teaching and outreach). Management of research related knowledge is

primarily handled at the discipline, departmental, research group or even individual researcher level. Management of data and knowledge related to administrative activities is supported by information systems (IS) common to many large organisations, such as financial accounting and human resource management systems. The concept of Knowledge Management Systems (KMS) in Higher Education (HE) is not as common or as widely used when compared to the corporate world. KMS “refer to a class of information systems applied to managing organizational knowledge. That is, they are IT-based systems developed to support and enhance the organizational processes of knowledge creation, storage/retrieval, transfer, and application” [5, p.114].

A less considered area for technology-support, but one in desperate need of KM strategies, is the handling of Graduate Attributes (GAs). GAs are the qualities, skills and understandings that a university agrees its students should develop during their time with the institution such as communication skills, critical thinking, team work, creativity, ethics, social responsibility etc. Treleaven & Voola [6] highlight the various, interchangeable terms related to graduate attributes: key skills; key competencies; transferable skills; generic skills; employability skills [7]; soft skills [8] [9]; graduate capabilities [10]; generic graduate attributes [11] [10]; professional skills; personal transferable skills [12] and generic competencies [13]. However, such phrases as generic skills, graduate attributes, graduate qualities, etc. can be thought of as synonymous/hyponymous [10]. This research paper uses the term graduate attributes as this term is most commonly in current use in Australia, the country in which this research was conducted.

The importance of GAs and ensuring they are embedded and assessed within a program of study is widely accepted across the sector. Furthermore, GAs are being used as a key strategy for defining and distinguishing one institution from another. There is currently confusion over how GAs should be defined, what these attributes look like within each discipline, how they should be taught, assessed and evaluated, and how their adoption should ultimately shape teaching practices in higher education. One popular definition nonetheless, is that adopted by the Australian Technology Network “the qualities, skills and understandings a university community expects its students to develop during their time at the institution and, consequently, shape the contribution they are able to make to their profession and as a citizen” [14]. Our initial investigations lead us to believe that graduate attributes are a form of soft, typically discipline specific knowledge that place importance on communication and critical thinking and problem solving skills. Furthermore, with the changing nature of Australian society we note there has been added emphasis on greater indigenous and international cultural awareness along with an expectation of wider community engagement on the part of university graduates.

Nonetheless, research has suggested there is a lack of shared understanding on what graduate skills are, and when and how to integrate and develop such skills in the curriculum [11] [15]. A key weakness in the literature is the vagueness in the conception of generic skills and the proliferation of terms in the literature [15]. For example, the notions of sustainability and ethical practice are difficult to conceptualize and capture as “skills” or “attributes”. While critical thinking and teamwork are clearly

enabling skills, sustainability and ethical practice are perhaps better conceptualized as graduate dispositions [16].

With the plethora of terms used to describe GAs across the sector, the Department of Education, Science and Training [17] and employer groups like Business Council of Australia (BCA) cited in [18] are calling for a more consensual and systematic approach to developing graduate attributes. To ensure high standards, the Australian Government has established the Australian Universities Quality Agency (AUQA), now called TEQSA (Tertiary Education Quality and Standards Agency), as an independent not-for-profit national organisation that promotes, audits and reports on quality assurance in Australian HE. There is growing pressure from TEQSA for Higher Education institutions to demonstrate that graduate attributes not only exist but they are implemented in a way that delivers evidence-based higher levels of attainment of the unit learning outcomes.

In this paper we present some of the data analysis we have performed in the GA domain using grounded theory. We also outline an ontology based approach to handle the complexity which exists in the domain. The domain analysis and approach offered in this paper seeks to contribute towards our research questions:

- What is the underlying ontology of Graduate Attributes present in Australian institutions of higher education?
- How can we create a common core of Graduate Attributes for use across higher education institutions?

In the next section, we develop further our understanding of the GA domain through grounded theory. After that, we outline the research we are undertaking to bring some clarity to field. Conclusions and future work appears in the end.

2 Related Work on Ontologies in Higher Education

There exist many definitions of ontology; some of them are given in [19] [20]. However, most researchers conclude that an ontology can be helpful when there is a need to define a common vocabulary of terms for some application domain. Ontologies can be developed at different levels and for different purposes. For example, there are top-level (or high-level/upper/foundation), domain, task and application ontologies.

Studying the ontologies in HE, Milam [20] points out that ontologies for HE are created for: representing the institution, representing academic discipline, documenting the data, creating meta data about learning and management systems, describing the nature of higher education enterprise, creating online resources for training materials and teaching.

Ontologies offer possibilities for organising and visualising educational knowledge and for this knowledge to be shared and reused by different educational applications [21]. Although ontologies exist in the domain of HE [22] [23], these ontologies are mostly oriented to the problems of teaching process formalization, sharing courses, etc. Until now, the complexity in the GA domain has not been the point of consideration for ontology researchers.

3 The Research Methodology

We have used Grounded Theory (GT) [24] in this paper for the primary data analysis of the GA domain within the context of the HE sector in Australia. Grounded Theory is a data analysis approach involving searching for concepts by looking for the codes [25], which can be thought of as themes. Codes are generated by analysis of the data, and a process of constant comparative analysis is used, which compares these codes until a foundation is discovered leading to theory generation. This paper does not provide an in-depth analysis or discussion of the different perspectives and approaches to GT, but presents the approach used for the purposes of this research. The method of GT enabled us to work systematically through a basic data corpus, generating codes to refer to both low level concepts and to more abstract categories. A picture paints a thousand words, so for this reason the Network Maps are able to present the concepts through Grounded Theory.

One key strength of Grounded Theory is that it can be used to uncover themes arising in the literature (incl. documents, repositories, websites, etc.) on a given topic or Hermeneutic Unit that would otherwise not be so visible. The underlying themes we may label “codes”. The “groundedness” of these codes, that is to say the occurrence of the codes in the literature tell us how important a particular theme is in the literature. We then construct a network map whereby codes are subjectively joined to one another. The number of times any given code may be joined with another is referred to as the “density” of the code. The combination of the code groundedness with its density permits us as researchers to gain a more complete understanding of a topic, which in turn informs further interpretivist or positivist research approaches.

In 2012 the Australian HE sector consists of 37 public universities and 2 private universities that are autonomous and self-accrediting; 4 other self-accrediting higher education institutions; and about 150 other institutions accredited by state and territory governments (such as theological colleges and providers specializing in professional and artistic courses of study). While our interest is in the GA domain for the whole of Australia, we restricted this ‘pilot’ analysis to 8 Australian Universities, trying to include at least one representative institution from each of the major groups (Go8 – University of Melbourne (UNIMELB), University of Queensland (UQ); IRU – Griffith University (GU), Murdoch University (MURDOCH); ATN - Curtin University (CURTIN); New Generation – Edith Cowan University (ECU), Regional – University of Southern Queensland (USQ) and one non-aligned – Macquarie University (MQ)). The universities and their identifier codes are shown in table 1.

Using the workbench Atlas.ti™ we have created a Hermeneutic Unit (HU) containing a set of primary documents as sources of GAs from the sample of eight Australian Universities. The primary documents were the list of graduate attributes and their descriptions available publically on the eight university websites to create eight textual primary documents to be examined in the Hermeneutic Unit of Graduate Attributes.

From these documents codes are created. The codes then are used to draw the Network Maps or Diagrams showing subjective relationships between different codes

Table 1. List of sample Australian Universities studied using GT in the GA domain

Source #	Name of University	University Code
1	Macquarie University, NSW	MQ
2	Griffith University, QLD	GU
3	Curtin University, WA	CURTIN
4	University of Melbourne, VIC	UNIMELB
5	Murdoch University, WA	MURDOCH
6	University of Queensland, QLD	UQ
7	University of Southern Queensland, QLD	USQ
8	Edith Cowan University, WA	ECU

or nodes. Afterwards, the network maps are created providing a graphical view of the graduate attributes with their differences. Map creation involves simply using the codes established, as nodes in a network map. The task the researcher then faces is to construct the links or relationships between the nodes. Whilst there is only one relation able to be expressed between any two nodes, and this relation is fixed throughout the hermeneutic unit, any number of relations may be created, representing many kinds of relationships between the other nodes.

Some relations used are: is-associated-with; is-part-of; is-cause-of; contradicts; is-a; is-cause-of and is-property-of. Carefully looking at the network maps shows two integer values on each code. The former relates to the “groundedness” or occurrence of this term in the source documents. The second numerical value on each code relates to the density of this code (showing how many other codes this code is related to) in the network. Both of these values permit the researcher to gain some understanding of the importance of one code over another.

Our research questions encompass two ways of looking at the GA data. Firstly, our question involves the establishment and evolution of multiple set[s] of assessable graduate attributes which could be at the institutional/discipline/department/group/individual level. Secondly, we are concerned with benchmarking and assurance of a minimum set of equivalent graduate capabilities. To cover these two perspectives, we had two researchers involved in data analysis using GT, one with the goal of distilling key ideas and identifying similarities in keeping with the more standard use of GT to draw out themes; the other with the goal of untangling and drawing out differences. We have used the network maps to explore the GA domain extensively from both of these angles as presented below.

4 Findings

Network maps provide interesting ways to explore the data. However, due to limitations in how much content can be legibly presented in one graph, it is not possible to explore all the data concurrently. Thus, we display only selected specific codes and their connections. For example in table 2, researcher one has selected the codes with the highest groundedness (above 10) or occurrence or ‘embedding’ in the literature;

while figure 1 explores the codes that relate to only two selected universities, namely Melbourne and Macquarie; one university that is a Go8 and another that sits just outside that group but aspires to be a Go8. We present a small selection of our maps.

4.1 Distilling and Simplifying the GA Domain: Researcher One

Based on the work of the researcher seeking to extract common concepts, 48 codes were identified. Viewing the network map in Figure 1 shows two values on each code. The first value relates to the GROUNDEDNESS or occurrence of this term in the graduate attributes source documents. The second value on each code relates to the DENSITY (with how many other codes this code is related to) of this code in the network. Both of these values permit some understanding of the importance of different codes over one another. Table 2 shows the codes with groundedness of equal to or greater than 10 in descending order of groundedness while figure 1 provides a comparison of two universities (Macquarie and Melbourne). We see in table 2, the relative emphasis on each GA code for each of the 8 universities. For example, we see that Curtin, Griffith and UQ have the strongest focus on Discipline Knowledge, while for others this GA takes second place to Professional Skills at Macquarie, Problem Solving at USQ, Communication at Murdoch, while Edith Cowan does not mention discipline specific skills at all.

In Figure 1 we see some overlapping codes (cultural awareness and understanding, international perspectives, academic excellence, community engagement, social responsibility). By looking at the groundedness (of 3) we can see that while academic excellence is shared with Macquarie and the University of Melbourne, only one other university shares it.

4.2 Unpacking and Expanding the GA Domain: Researcher Two

Grounded Theory is an interpretive research method and does not claim to be replicable, and admits subjectivity is acceptable in research as the human experience or interpretation is central to the research process. Also, as previously mentioned, our two researchers had different goals. Researcher One had distilled 48 codes from the source documents. Researcher Two in seeking to tease out differences marked up 159 codes. This larger number of codes also results in codes with lower groundedness and density values. We dedicate more space to this perspective, as much of the literature on embedding GAs has focused on the shared and common perspective rather than on what differences may exist. To identify how codes were connected as part of the same GA, a code for each GA for each institution was created. For example MQ_1 (i.e. Macquarie GA number 1) is comprised of Discipline Specific Knowledge and Discipline Specific Skills. Similarly, to identify the relationships between GAs we broke each one down according to the number of terms/concepts included within them. For example, USQ's tenth GA (USQ_10) is "Knowledge and skills required in your discipline to develop sustainable practice in relation to communities, economies and the environment." This GA includes the concepts of knowledge, skills and sustainability.

Table 2. Top 10 codes for the Graduate Attribute domain for 8 Australian Universities

Graduate Attribute Code	CURTIN	ECU	GU	MQ	MURD-OCH	UNI-MELB	UQ	USQ	Total
Discipline knowledge	4	-	6	4	5	1	7	4	31
Communication	2	2	5	2	6	1	5	2	25
Critical thinking	3	1	3	3	3	2	6	3	24
Problem solving	2	1	2	3	5	2	4	5	24
Professional skills	3	-	4	5	5	1	2	4	24
Cultural awareness & understanding	2	2	4	2	3	4	1	3	21
Interpersonal skills	-	2	2	2	6	2	2	5	21
Ethical	1	-	1	1	6	2	4	3	18
Community engagement	-	-	3	3	1	6	1	1	15
Social responsibility	-	-	2	4	4	-	3	1	14
Using technology	3	1	1	3	2	1	1	2	14
Synthesising information	2	-	3	7	1	-	-	-	13
International perspectives	2	-	2	2	2	2	1	1	12
Research capability	1	-	1	2	1	1	5	1	12
Creativity	1	1	-	3	3	1	1	1	11
Independence	1	-	-	1	2	1	4	2	11
Literacy	-	-	-	2	2	1	2	3	10

As this paper concerns KM in HE, it is interesting to consider how other GAs with a knowledge-focused outcome are related. As depicted in Figure 2, we can also see that USQ_10 (bottom right) links to knowledge through its application to sustainability. We find that knowledge is a recurring theme across many GAs with a groundedness of 23, and density of 25.

Some interesting codes and relationships include:

- Discipline Specific Knowledge is the most popular attribute connecting to most of our sample universities.
- 50 % of our sample universities want their graduates to be creative.
- Next prominent capabilities in the universities GA list are applying the knowledge through ethics, social and discipline specific skills.
- Being engaged in their communities, cultural diversity, problem solving, research capability, Interdisciplinarity.

Table 3 provides a summary of the use of the term “Knowledge” and reveals, in general, the relative (un)importance or (de)emphasis on knowledge across GAs. Out of the eight sample universities, ECU is the only one that does not use the term “knowledge” in any of its GAs, though it uses related terms such as critical thinking, problem solving, cultural awareness and understanding and creativity. Focusing on codes related to “knowledge” is an example of how network complexity can be managed by the creation of code families. A code family is created that summarizes the codes that are similar; this process may be labelled selective coding. Three code families were created on the subjective understanding of the researcher: COGNITIVE, SOCIAL/INTERPERSONAL and SKILLS. The GAs from the sample universities may be seen as a mix of cognitive, skill and social attributes.

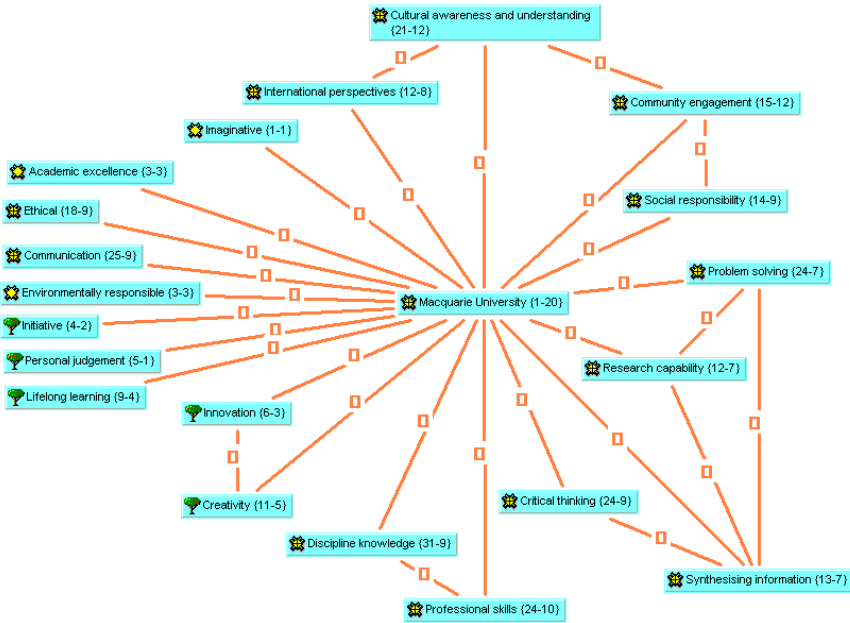


Fig. 1. A network map in Atlas.ti for Macquarie University and University of Melbourne

[] represents is-associated-with

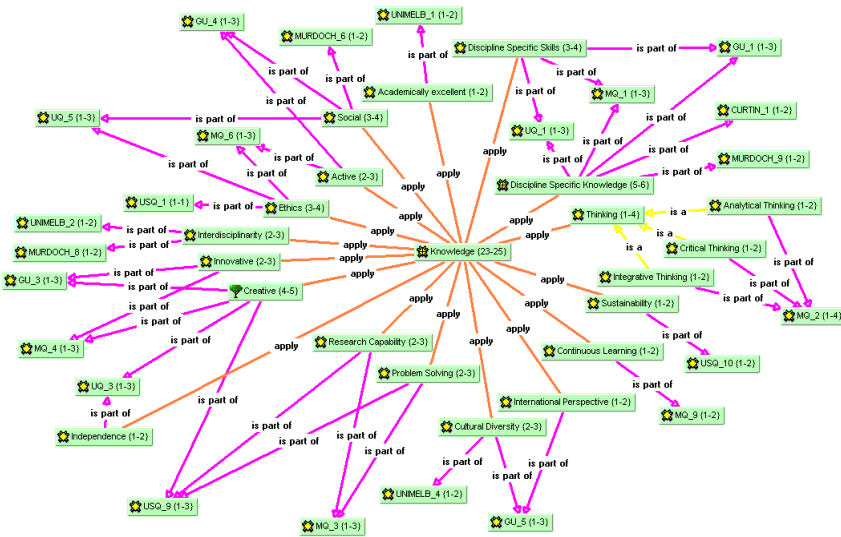


Fig. 2. A KNOWLEDGE network map with groundedness of 23 and density of 25. To increase the semantic value of the map and avoid the loss of information through summary of the data, coloured nodes and links permit the viewer to identify the types of relationships, e.g. orange represents GAs that apply KNOWLEDGE.

Table 3. Total number of GAs and those including KNOWLEDGE for included Universities

University Code	MQ	ECU	GU	CURTIN	UNI-MELB	MUR-DOCH	UQ	USQ
Total # of GAs	9	5	5	9	5	9	5	10
# GAs involving KNOWLEDGE	6	0	4	1	3	3	3	3
Percentage (%)	66.6	0	80	11	60	33.3	60	30

To minimize misinterpretation and reduce bias, the codes were assigned to the code families on the basis of the description of the GAs given in the source documents. Table 4 shows the number of each type for each institution.

Table 4. Number of university GA and each type of code for the eight universities

Code Families/UniCode	MQ	ECU	GU	CURTIN	UNI-MELB	MURDOCH	UQ	USQ
Cognitive	5	2	2	4	2	4	3	3
Social	3	3	3	4	4	4	2	6
Skills	3	4	3	5	3	4	4	7

Table 5 shows the SOCIAL/INTERPERSONAL and SKILLS GAs assigned to each code family in the respective university. Table 5 shows that there exist some GAs in each university, that appear in more than one code family. Some GAs, for example sustainable practice (USQ_10), ability to communicate (ECU_1), ability to work in teams (ECU_2), critical appraisal skills (ECU_3) and ability to generate ideas (ECU_4) are only understood as SKILLS. This demonstrates the difficulty with interpretation and classification of the GAs.

In Table 6, we can see the summary of a number of GAs in the sample universities, belonging to the COGNITIVE family. In figure 3 we have created a network map for the COGNITIVE family. The DISCIPLINE SPECIFIC KNOWLEDGE (groundedness of 4, density of 7) is noticeable from the network map. There are some other key codes such as thinking (3-5), knowledgeable (2-5), creative (5-6), critical thinking (3-5), life-long learning skills (2-3). The low groundedness of the codes academically excellent (1-1) and sustainable practice (1-1) highlight the problems of use of different terms and different use of the same terms and the need to capture semantics. For example MQ_7 = “Socially and Environmentally Active & Responsible”; GU_4 = “Socially Responsible and Engaged in Their Communities”, use alternative terms related to sustainability.

Particularly based on the GT analysis conducted by the second researcher, we draw the conclusion that while the data is not substantial in this domain, it is nevertheless very difficult and time consuming to model existing similarities and differences. This leads us to propose an ontology-based solution.

Table 5. Themes with id for SOCIAL/INTERPERSONAL & SKILLS code families

Social / Interpersonal	Uni_Code	Skills	Uni_Code
Effective Communication	MQ_5	Discipline Specific Skills	MQ_1
Engaged and Ethical Local & Global citizens	MQ_6	Effective Communication	MQ_5
Socially & Environmentally Active & Responsible	MQ_7	Capable of Professional & Personal Judgment & Initiative	MQ_8
Effective Communicators and Team Members	GU_2	Skilled in their Disciplines	GU_1
Socially Responsible & Engaged in Their Communities	GU_4	Effective Communicators and Team Members	GU_2
Competent in Culturally Diverse & Int'l Environments	GU_5	Innovative and Creative with Critical Judgement	GU_3
Communicate effectively	CURTIN_4	Apply discipline knowledge, principles and concepts	CURTIN_1
Recognize and apply international perspectives	CURTIN_7	Think critically, creatively and reflectively	CURTIN_2
Demonstrate cultural awareness & understanding	CURTIN_8	Communicate effectively	CURTIN_4
Applying professional skills	CURTIN_9	Use technologies appropriately	CURTIN_5
Academically excellent	UNIMELB_1	Applying professional skills	CURTIN_9
Leaders in communities	UNIMELB_3	Academically excellent	UNIMELB_1
Attuned to cultural diversity	UNIMELB_4	Knowledgeable across disciplines	UNIMELB_2
Active global citizens	UNIMELB_5	Leaders in communities	UNIMELB_3
Social interaction	MURDOCH_3	Communication	MURDOCH_1
Ethics	MURDOCH_5	Critical and creative thinking	MURDOCH_2
Social justice	MURDOCH_6	Ethics	MURDOCH_5
Global perspective	MURDOCH_7	In-depth knowledge of a field of study	MURDOCH_9
Effective Communication	UQ_2	In-depth knowledge & skills in field of study	UQ_1
Ethical and Social Understanding	UQ_5	Effective Communication	UQ_2
Ethical research and inquiry	USQ_1	Independence and Creativity	UQ_3
Written & oral communication	USQ_4	Critical Judgement	UQ_4
Interpersonal skills	USQ_5	Problem solving	USQ_2
Teamwork	USQ_6	Academic, professional and digital literacy	USQ_3
Cultural literacy	USQ_7	Written & oral communication	USQ_4
Sustainable practice	USQ_10	Interpersonal skills	USQ_5
Ability to work in teams	ECU_2	Cultural literacy	USQ_7
Cross-cultural and international outlook	ECU_5	Mgt, planning and organizational skills	USQ_8

Table 6. Total number of GAs in each university and belonging to the “COGNITIVE” family

University Code	MQ	ECU	GU	CURTIN	MELB	MURDOCH	UQ	USQ
Total # of GAs	9	5	5	9	5	9	5	10
Cognitive	5	2	2	4	2	4	3	3
Percentage (%)	55.55	40	40	44.44	40	44.44	60	30

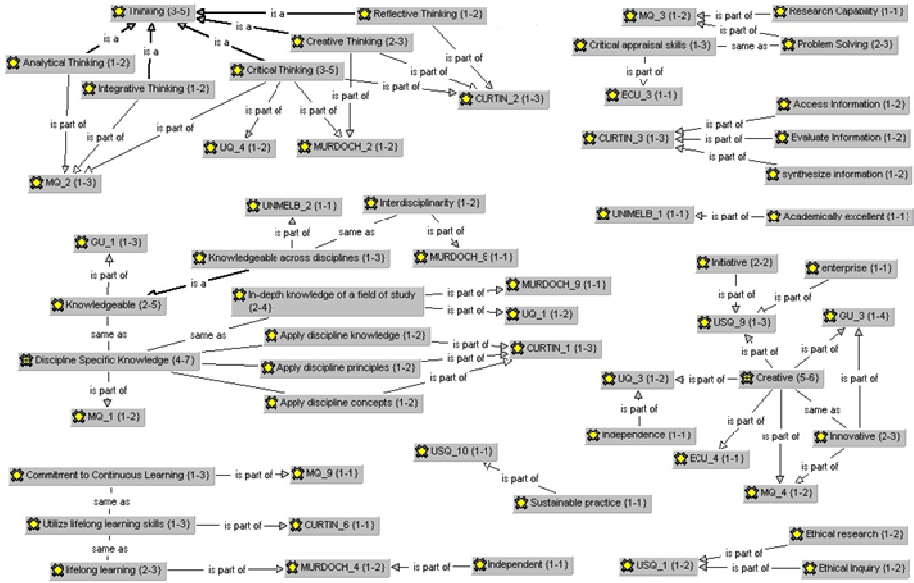


Fig. 3. A network map for the code family “COGNITIVE”

5 An Ontology-Based Approach for Managing GAs

Ontologies and the increased functionality they offer will bring new opportunities to the GA domain. Ontologies assure “a shared and common understanding of a domain that can be communicated between people and application systems” [23]. They attempt to put together a thorough representation of a domain by specifying all of its concepts, hierarchical links and the meaningful relationships.

Higher education scholars [11, 26] quite rightly argue that the lack of conceptual clarity has confused the GAs. Ontologies are one of the answers to knowledge structuring and modelling, by providing a formal conceptualization of a particular domain that is shared by a group of people in an organization [27]. The general consensus is that ontologies are able to improve communication, sharing and reuse [28]. In HE, many different specifications have been termed ontologies, the IS experts equate them as thesauri/glossaries/controlled vocabularies to taxonomies, and AI experts link them

as logical theories/frame languages etc. The extant research literature suggests that the foundation for developing standards of academic achievement can be found in learning taxonomies [29]. For example, the SOLO (Structure of Observed Learning Outcome) and Bloom's taxonomy described in [29] [30] [31] and the classification systems relating to GAs include the work conducted by Barrie [32].

There is no unified taxonomy of managing GAs. The motivation to create such an ontology is to push the standardization process that will enable the connection among academic systems and improve educational workflows, communication, and collaboration between universities. The approach to developing a standard of terminology and concepts relating to GAs is not an easy task. It requires techniques from information retrieval and linguistics. For example, corpus/concordance techniques will be applied to volumes of data captured from university websites. As there is currently no standard way in which this data is made available (for example, data may consist of any or all of the following: electronic text, PDF documents, unit outlines within a password restricted learning management system, or electronic data structured in table format), we will not invest time into developing an automated technique to gather this initial data. Sufficient data for the purposes of our prototype can be found online or obtained in documents provided by contacts that we have at many institutions to produce a corpus to which our automated methods are applied. This corpus can be added to and contain any text relevant to GAs.

For example, in the future we could add descriptions of the knowledge and skills assessed in the graduate skills assessment (GSA). The data to be included in the prototype is discussed further below. The data collected will be tagged with the source and will include the institution, discipline, department and faculty so that these can be used as selection items and used for identification when comparisons are presented.

The strength of our approach is that it includes an automatic method for identifying terms and key phrases directly from the data by their relative frequencies. It draws attention to not-yet-standardized terminology by the left- and right- collocates of individual words. It lends itself to our (project's) need to identify newly fledged terms/phrases in the graduate attribute (GA) discourse of other universities. Some of the key terminology is of course highlighted up-front in official documents, but other relevant terminology will simply appear (more often than one might expect) in explanations of institutional GA vocabulary. These alternative terms will be very useful in exploring the values and defining semantic relations among the whole set.

Once the corpus of GA text is together, there are tools (e.g. WordSmith) to automate the frequency counts and concordances (Key-word-in-context displays) to show up collocations. The concordances also draw attention to the linguistic relations between terms; and "manual" inspection of sets of instances allows us to interpret the semantic relations with the clusters of terms. This is a finite task, since the set of terms belonging to GA discourse in 8-10 universities is not open-ended.

In addition, we need a sophisticated methodology to help develop an ontology for managing GAs. Although ontology building methodologies are not mature enough, there are some methodologies available like the Enterprise Ontology (EO), TOVE methodology, METHONTOLOGY, On-To-Knowledge methodology, Activity-First

Method (AFM). In the next phase of this study, we could select good features of the above methodologies in our ontology building process. Usually, an ontology development methodology has its own support language/tool which has a function to generate the ontology. We could later select any one of the following popular ontology languages/tools such as, Ontolingua, RDF (Resource Description Framework), OWL (Web Ontology Language), OntoEdit, OilEd, WebODE, Protege-2000, Ontosaurus, and LinkFactory.

6 Conclusion

Clearly, there are some difficulties involved in creating ontologies. The ontology development process can be difficult and costly. However, by bringing the work together into a taxonomy/ontology, we will be validating and adding value to the work already conducted in the HE sector. At the discipline specific level, the approach would allow levels of compliance to standards specified by accreditation bodies and also comparison between programs and institutions within that discipline. For disciplines without accreditation bodies which have fewer opportunities for guidance and benchmarking, comparison across disciplines may assist in the development of threshold learning outcomes. The taxonomy of terms and concepts would allow a better understanding of the graduate outcomes achieved at various levels/degrees in the HE sector in keeping with the Australian Quality Framework (AQF) and support the future development of a myDegree and myUniversity site similar to the mySchools website which received strong public support. We envisage that the ontology will allow comparison with the goals and content in tests such as Graduate Skills Assessment (GSA), the Collegiate Learning Assessment (CLA) test used in the USA, Graduate Destination Survey (GDS), Course Experience Questionnaire (CEQ) and the Australasian Survey of Student Engagement (AUSSE).

References

1. Sedziuviene, N., Vveinhardt, J.: The Paradigm of Knowledge Management in Higher Educational Institutions. *Engineering Economics* 5(65), 79–90 (2009)
2. Adhikari, D.R.: Knowledge management in academic institutions. *The International Journal of Educational Management* 24(2), 94–104 (2010)
3. Branin, J.J.: Knowledge Management in Academic Libraries: Building the Knowledge Bank at the Ohio State University. *The Journal of Academic Librarianship* 39(4), 41–56 (2003)
4. Petrides, L.A., Guiney, S.Z.: Knowledge Management for School Leaders: An Ecological Framework for Thinking Schools. *Teachers College Record* 104(8), 1702–1717 (2002)
5. Alavi, M., Leinder, D.E.: Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly* 25(1), 107–136 (2001)
6. Treleaven, L., Voola, R.: Integrating the Development of Graduate Attributes Through Constructive Alignment. *Journal of Marketing Education* 30(2), 160–173 (2008)

7. Curtis, D., McKenzie, P.: Employability Skills for Australian Industry (electronic resource): Literature Review and Framework Development Australian Council for Educational Research (2001)
8. BIHECC. Graduate Employability Skills. Graduate Employability Skills Report (2007)
9. Freeman, M., et al.: Business as usual: A collaborative investigation of existing resources, strengths, gaps and challenges to be addressed for sustainability in teaching and learning in Australian university business faculties. ABDC Scoping Report (2008)
10. Bowden, J., et al.: Generic capabilities of ATN University graduates (2002)
11. Barrie, S.C.: A Research-based Approach to Generic Graduate Attributes Policy. Higher Education Research & Development 23(3), 261–275 (2004)
12. Drummond, I., Nixon, I., Wiltshire, J.: Personal Transferable Skills in Higher Education: The Problems of Implementing Good Practice. Quality Assurance in Education 6(1), 19–27 (1998)
13. Tuning-Report. Reference Points for the Design and Delivery of Degree Programmes in Business (2008)
14. ATN. Generic Capabilities of ATN: Australian Technology Network University Graduates (November 20, 2011), <http://www.clt.uts.edu.au/ATN.grad.cap.project.index.html>
15. Sin, S., Reid, A.: Developing Generic Skills in Accounting: Resourcing and Reflecting on Trans-Disciplinary Research and Insights. In: Annual Conference for the Association for Research in Education (2005)
16. Reid, A., Petocz, P.: University Lecturers'. Understanding of Sustainability Higher Education 51(1), 105–123 (2006)
17. DEST. Department of Education, Science and Training: Administrative and corporate publications. Annual Report (2005)
18. Thompson, D., et al.: Integrating graduate attributes with assessment criteria in business education: using an online assessment system (2008)
19. Cherednichenko, O., Kuklenko, D., Zlatk, S.: Towards Information Management System for Licensing in Higher Education: An Ontology-Based Approach. Information Technology Study Group (2010)
20. Milam, J.: Ontologies in Higher Education. In: Metcalfe, A.S. (ed.) Knowledge Management and Higher Education: A Critical Analysis. Information Science Publishing (2003)
21. Aroyo, L., Dicheva, D.: Workshop on Concepts and Ontologies in Web-based Educational Systems. In: International Conference on Computers in Education (2002)
22. Ermolayev, V., Spivakovsky, A., Zholtkevych, G.: UnIT-NET IEDI: Ukrainian National Infrastructure for Electronic Data Interchange. Reference Architecture Spec. (2003)
23. Wilson, R.: The Role of Ontologies in Teaching and Learning. JISC Technology and Standards Watch Report TSW0402 (2004)
24. Glaser, B., Strauss, A.: The Discovery of Grounded Theory: Strategies for Qualitative Research. Sociology Press (1967)
25. Allan, G.: A critique of using grounded theory as a research method. Electronic Journal of Business Research Methods 2(1), 1–10 (2003)
26. Chan, C.C., et al.: Applying the Structure of the Observed Learning Outcomes (SOLO) Taxonomy on Student's Learning Outcomes: An Empirical Study. Assessment & Evaluation in Higher Education 27(6), 511–527 (2002)
27. O'Leary, D.E.: Using AI in Knowledge Management: Knowledge bases and ontologies. IEEE Intelligent Systems 13(3), 34–39 (1998)
28. Uschold, M., King, M.: Towards a Methodology for Building Ontologies. IJCAI 1995 Workshop on Basic Ontological Issues in Knowledge Sharing (1995)

29. Green, W., Hammer, S., Star, C.: Facing Up to the Challenge: Why is it so Hard to Develop Graduate Attributes? *HE and Development* 28(1), 17–29 (2009)
30. GSA. Critical Thinking, Team Work, Ethical Practice & Sustainability. An initiative of the Australian Government Department of Education, Employment and Workplace Relations funded by Australian Learning and Teaching Council Ltd. (2010)
31. Phillips, V., Bond, C.: Undergraduates' experiences of critical thinking. *Higher Education Research & Development* 23(3), 277–294 (2004)
32. Barrie, S.C.: A conceptual framework for the teaching and learning of graduate attributes. *Studies in Higher Education* 32(4), 439–458 (2007)

Commonsense Knowledge Acquisition through Children's Stories

Roland Christian Chua Jr. and Ethel Ong

Center for Language Technologies, De La Salle University, Manila, Philippines
rc_chua07@yahoo.com, ethel.ong@delasalle.ph

Abstract. Humans interact with each other using their collection of commonsense knowledge about everyday concepts and their relationships. To establish a similar natural form of interaction with computers, they should be given the same collection of knowledge. Various research works have focused on building large-scale commonsense knowledge that computers can use. But capturing and representing commonsense knowledge into a machine-usable repository, whether manual or automated, are still far from completion. This research explores an approach to acquiring commonsense knowledge through the use of children's stories. Relation extraction templates are also utilized to store the learned knowledge into an ontology, which can then be used by automatic story generators and other applications with children as the target users.

Keywords: Commonsense Knowledge, Story Generation, Relation Extraction.

1 Introduction

People acquire commonsense knowledge as part of their daily living. This includes the physical, social, temporal, psychological and spatial experiences. To make machines be able to mimic human abilities and behavior, especially in applications requiring a more natural man-machine interaction [1], an adequately-sized body of commonsense knowledge must be made available to them.

Several knowledge repositories have been developed like WordNet [2], Cyc [3], ConceptNet [4], and VerbNet [5]. These contain entries ranging from syntactic to semantic in nature. ConceptNet [6] is a large-scale ontology of commonsense knowledge that has been collected through the Open Mind Common Sense (OMCS) project. OMCS [7] enabled the general public to contribute by posting questions that pertain to commonsense. WordNet [2] and Cyc [3] are similar and notable repositories that give semantic meaning to information, but differ in the type of machine understanding that they want to achieve. WordNet is envisioned for lexical categorization and word-similarity while Cyc is envisioned for formalized logical reasoning. ConceptNet, on the other hand, is envisioned for making practical reasoning over real world context.

Building commonsense knowledge repositories is very tedious and time-consuming. OMCS turned to the community instead of the experts, while systems like BagPack [8]

use the Games with a Purpose (GWAP) [9] approach that uses entertainment to encourage user participation and gives incentives to users as they play.

Much research has already been done on probing adults while children are being taken for granted as prime contributors to the acquisition task [10]. The Verbosity [11] system harnesses knowledge from children age ten to twelve years old who are explicitly expanding their mental capacities as well as their commonsense repository. However, few relations are covered and a question-answer format is employed.

This paper presents an approach to commonsense knowledge acquisition from children through the use of stories. Because storytelling is a common way for people to share information and exchange experiences, it is worth exploring how this can be used to elicit participation from children. Applications that require commonsense knowledge, such as the story generation systems in [12] and [13], can then utilize the ontology to produce stories to entertain young children.

2 Ontology Structure

An ontology is used for storing and representing the acquired knowledge as a collection of assertions. An assertion is comprised of two concepts that are connected through relationships. This allows commonsense to be defined in the form of binary concepts, the approach used in ConceptNet [14]. For example, an *apple-isA-fruit* asserts the concepts *apple* and *fruit* to be related through the *isA* relation.

Relations and assertions are terms that are used interchangeably in this paper; they refer to two concepts being related together by a relationship. As these comprise the contents of the ontology, the knowledge that is stored is highly dependent on the types of relations that are supported. Table 1 lists some of the relations currently included.

Table 1. Sample Set of Relations

Relation Name	Answers the Question -	Example Assertions
IsA	What is it?	<i>Apple</i> is a fruit
UsedFor	What is it used for?	<i>Pencil</i> is used for <i>writing</i>
LocatedAt	Where can it be found?	You can find a <i>pencil</i> inside a <i>classroom</i>
ShapeOf	What is its shape?	The shape of a <i>ball</i> is <i>round</i>
EffectOf	What is its effect?	The effect of <i>not eating</i> is <i>being hungry</i>

3 System Architecture

Knowledge acquisition progresses in three major phases, namely template-based story generation, user input, and data validation. Commonsense knowledge is acquired with the use of children's stories. Partial stories (i.e., stories containing blanks to represent missing concepts) are automatically generated by instantiating story templates stored

as text files. Users then fill-up the blanks with new concepts to complete the stories. These concepts are used to form new assertions that are added to the ontology and are used for the subsequent generation of new story instances.

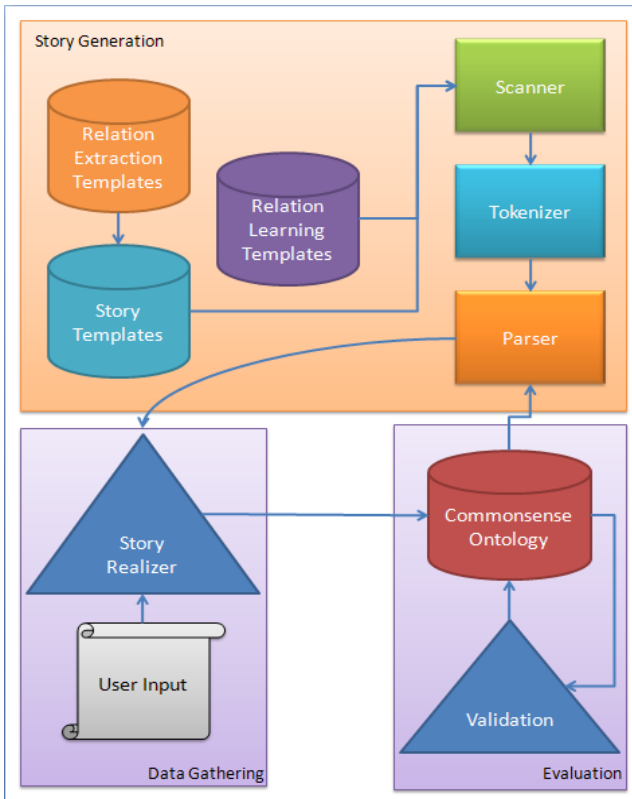


Fig. 1. System Architecture

3.1 Relation Extraction Templates

A set of relation extraction templates have been manually defined and forms part of the story templates, as depicted in Figure 2. This allows the system to identify the types of relations for the assertions provided by the users as they fill up blanks with concepts to complete the stories.

A relation extraction template is composed of three parameters, following the format of an assertion:

<Concept A><Relationship><Concept B>

Either <Concept A> or <Concept B> is left blank for the user to fill, while the other is filled with a concept taken from the current contents of the commonsense ontology. The blank concepts are denoted by a question mark (?).

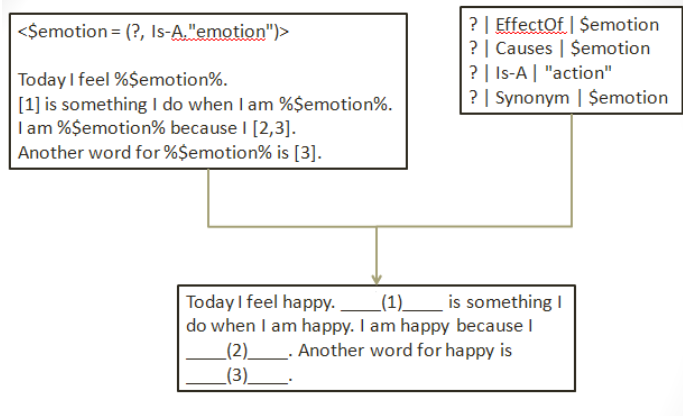


Fig. 2. Story Template with Relation Extraction Templates

3.2 Story Templates

Several story templates have been manually prepared with rules on how to use existing knowledge to generate various story instances, thus allowing the system to continuously acquire new assertions. These templates are stored as text files to facilitate future scalability by supporting the specification of more story variants.

Figure 3 shows a sample story template structure, and three sample surface forms of story instances that can be generated. Items in angled brackets represent queries to the ontology. The dollar (\$) symbol defines a variable and can be used to store query results for subsequent use in later part of the story. A pair of percentage (%) symbols is used to represent internal temporary memory. The pair of brackets with numbers ([1,2]) represents a blank in the generated story that must be filled by the user. Each number represents a rule in the Relation Extraction Templates that applies to the specific blank. These rules determine the new knowledge that the ontology will be learning. The ampersand (&) represents the conditional AND.

A query to the ontology contains three parameters – *concept1*, *relationship*, and *concept2*. Either *concept1* or *concept2* is replaced with the question mark (?) symbol to represent the concept to be retrieved. In the given example in Figure 3, the first query `<(?, Is-A, "location")>` retrieves a concept that is a “location” from the ontology and stores the result in the variable `$location`. The second query `<(?, Is-A, "object")>` retrieves an object and stores the result in the variable `$object`. The AND condition specifies that the resulting object must also be found in the given location stored in the variable `$location`.

The `#if` represents a conditional if-statement whose corresponding story sentence will be generated if the query `(?, ColorOf, $object)` is able to return a known color for the `$object` specified. However since the `#if` is followed by an exclamation point (!) the condition is treated as a negation; thus the corresponding story sentence will be generated if the ontology does not contain a relation specifying the color for the given `$object`. This is useful so that the system has a way to learn knowledge it does not yet know of.

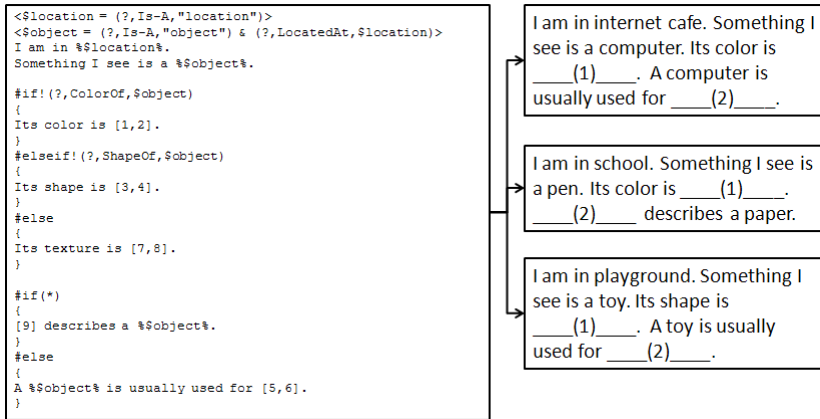


Fig. 3. Sample Story Template Structure and Some Corresponding Story Instances

The #if (*) with asterisk represents story text that can be generated randomly, to create variances by providing several story paths that the system can choose from.

The system also implements a memory-like storage, i.e., %\$object%, that temporarily stores query results, to prevent having to execute the same query to retrieve data needed again by the story generator.

Each story template is designed to learn about different categories of concepts, such as objects, places, functions, and feelings. The templates complement each other in a “story web” pattern (shown in Figure 4). This allows “Story Template 5”, for example, to integrate “Story Template 1” designed to learn assertions about objects, to generate longer stories and thereby facilitate the acquisition of further knowledge.

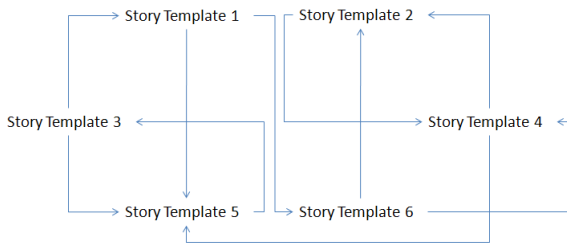


Fig. 4. Story Web

4 Commonsense Validation

The story ogenerator depends on the assertions in the ontology to instantiate new stories from the story templates. It is therefore necessary that assertions acquired from the users be validated correctly. As children are the ones that fill up blanks to add new assertions to the ontology, knowledgeable adults are then responsible for validating the provided inputs.

Validating commonsense is achieved with the use of a *confidence value* for each assertion, computed using the following formula:

$$\text{Confidence} = \sum \alpha \div (\beta * 4)$$

Where α is the total points earned from validation

β is number of times the assertion is validated

The formula has been inspired from a previous work on Community Refinement of Ontology (CoReO) [15]. COREO handled refining the initial contents of a given ontology for a certain domain with the use of a game where the assertions served as game objects that users interact with. The problem, however, was the validations made on the assertions were limited to a strict “yes” or “no” only. This did not allow the users to partially agree or partially disagree to the statement being shown to them.

In our approach, the validation provided a number of possible user responses, namely *agree*, *partially agree*, *neutral*, *partially disagree*, and *disagree*, with a corresponding scoring scheme of 4, 3, 2, 1 and 0, respectively.

5 Conclusion and Further Work

This paper presented a preliminary work on using generated stories to acquire commonsense knowledge from children. A set of story templates has been manually prepared and is used by the story generator to produce story instances using partial knowledge available from the ontology. As the user inputs concepts to complete a story, corresponding relation extraction templates associated to the source story template are used to acquire assertions that are then added to the knowledge base. A simple validation scheme has also been implemented to provide a confidence value of the assertions in the ontology before these can be used.

The commonsense ontology has been populated with a minimal set of assertions needed by the story generator to instantiate stories that will start the knowledge acquisition process. However, this minimal set is hard to determine and leads to situations where the ontology cannot provide the results to satisfy a query in order to generate the rest of the story. This problem is expected to be addressed by increasing the number of story templates to further saturate the story web shown earlier in Figure 4.

The approach implemented has provided positive initial results. As new assertions are learned, the stories are tweaked which allows the learning process to continue. However, there will come a point in time when the learning rate will decrease. This is the ideal period to add new story templates. The new story templates should address the shortcomings of the initial set of templates by providing new ways of acquiring more assertions.

As with any knowledge acquisition task, continued support from the community of users is needed in order to increase the amount of assertions in the system. An incentive scheme can be implemented to encourage the users to actually use the system on a regular basis. Embedding the system into an interactive learning environment can also provide an opportunity for language and literacy development through the combined storytelling and commonsense ontology approach.

References

1. Liu, H., Singh, P., Lieberman, H., Barry, B.: Beating Commonsense into Interactive Applications. *AI Magazine* 25, 63–76 (2004)
2. Miller, G.A.: WordNet: A Lexical Database for English. *Communications of the ACM* 38(11), 39–41 (1995)
3. Lenat, D.B.: CYC - A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38(11), 33–38 (1995)
4. Havasi, C., Speer, R., Alonso, J.: ConceptNet 3 - A Flexible, Multilingual Semantic Network for Common Sense Knowledge. In: *Proceedings of Recent Advances in Natural Language Processing* (2007)
5. Kipper, K.S.: VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. Dissertation. University of Pennsylvania, Philadelphia (2005)
6. Liu, H., Singh, P.: Commonsense Reasoning in and Over Natural Language. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) *KES 2004, Part III. LNCS (LNAI)*, vol. 3215, pp. 293–306. Springer, Heidelberg (2004)
7. Singh, P., Barry, B., Liu, H.: Teaching Machines about Everyday Life. *BT Technology Journal* 22(4), 227–240 (2004)
8. Herdagdelen, A., Baroni, M.: BagPack - A General Framework to Represent Semantic Relations. In: *Proceedings of the EACL 2009 Geometrical Models for Natural Language Semantics Workshop*, pp. 33–40. ACL, East Stroudsburg (2009)
9. Von Ahn, L., Dabbish, L.: Designing Games with a Purpose. *Communications of the ACM* 51(8), 58–67 (2008)
10. Bosch, A., Nauts, P., Eckha, N.: A Kids' Open Mind Common Sense. In: *Proceedings of the 2010 AAAI Fall Symposium Series on Common Sense Knowledge*, pp. 114–119. AAAI Press, Virginia (2010)
11. Von Ahn, L., Kedia, M., Blum, M.: Verbosity: A Game for Collecting Common-Sense Facts. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 75–78. ACM, New York (2006)
12. Ong, E.: A Commonsense Knowledge Base for Generating Children's Stories. In: *Proceedings of the 2010 AAAI Fall Symposium Series on Common Sense Knowledge*, pp. 82–87. AAAI, Virginia (2010)
13. Ang, K., Yu, S., Ong, E.: Theme-Based Cause-Effect Planning for Multiple-Scene Story Generation. In: *Proceedings of the 2nd International Conference on Computational Creativity*, Mexico City, Mexico, pp. 48–53 (April 2011)
14. Liu, H., Singh, P.: ConceptNet – A Practical Commonsense Reasoning Tool-kit. *BT Technology Journal* 22(4), 211–226 (2004)
15. Chua, A.M., Chua, R.C., Dychingching, A.V., Ang, T., Espiritu, J.L., Lim, N.R., Cheng, D.: Crowdsourcing through Social Gaming for Community-Driven Ontology Engineering, Results and Observations. In: *Proceedings of the Fifth International Workshop on Ontology Matching*, Shanghai, China, vol. 689, pp. 244–245. CEUR-WS.org (2010)

Externalizing Senses of Worth in Medical Service Based on Ontological Engineering

Taisuke Ogawa¹, Mitsuru Ikeda¹, Muneou Suzuki², and Kenji Araki²

¹ School of Knowledge Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai Nomi, Ishikawa, Japan
{t-ogawa, ikeda}@jaist.ac.jp

² Medical Information Technology, University of Miyazaki Hospital, Japan

Abstract. In the medical industry, sharing a sense of worth among medical staff members can be one method of improving the quality of service. This is because high-quality medical service comes from not only careful consideration with respect to service design but also creativity in service implementation. Since identifying a clear sense of worth can be difficult, it can be hard to externalize them in a communicative way. This article will present a method to externalize sense of worth through service modeling.

Keywords: Service Modeling, knowledge Acquisition, Sense of Worth, Ontology.

1 Introduction

Sharing senses of worth of medical staff members is an important issue in striving to improve the quality of medical service. For instance, in Berryfs Mayo clinic study, the clinic is managed based on sharing their sense of worth, and the sharing stems from a source of high-level education and excellent team building [1]. In other words, the senses of worth of service experts parallel their belief about their work. Some research about the growth process of professional service staff [2] notes that formations of beliefs effect empirical learning in the workplace, and quality of service is the result of learning. Communication is essential to the formation of beliefs. This author is interested in how communications can be designed rationally, and how the computer system supports such communications.

In this article, the author sketches a method for sharing the sense of worth of medical staff members on the clinical pathway (CP) design activity. CPs are the documents on which hospital procedures and activity goals (e.g. examination and treatment) are written. A role of clinical pathways is to support different types of experts in working and sharing knowledge. The goal of this research is to develop a method for externalizing senses of worth of CP designers.

In the next section, an approach based on ontological engineering is explained. Based on this approach, in Section 3, CP design activity is analyzed with ontological engineering. In Section 4, a method is described for externalizing senses of worth of medical staff members.

2 How Ontology Helps Externalize the Senses of Worth of Medical Staff Members

In this research, the senses of worth behind CP are expressed by problems intended by designers (expert staff members in the hospital) in CP design activities. The research site is the Miyazaki University Hospital in Japan.

In this hospital, many doctors think it is ideal that CPs are designed in the Problem Oriented System. The POS (problem oriented system) has been proposed by Weed in the 1960s [3]. The system caters to rational medical service design and assessment. However, the “problem” is quite vague. The POS holds the ideal that medical service should be designed not only from a medical viewpoint but also from patients and families’ viewpoints. Therefore, the “problem” includes many aspects of patient life and issues related to sense of worth (e.g., patient’s everyday life, work, emotions, etc). In addition, medical services are complex and some areas are experimental. This presents difficulties in externalizing the intended problems.

If problems are externalized more deeply and are well organized, explicit knowledge about the sense of worth can be revealed. This is the basic concept supporting our need to externalize senses of worth of medical staff members and patients.

In developing this method, the author analyzed CP design activity along with ontological engineering. Ontologizing the CP design activity makes the analysis more precise, and the ontology works as a meta-model in a support system guiding the externalization and providing a vocabulary of problems.

3 Analysis of Difficulties in Externalizing Problems on CP Design

3.1 Analysis Based on an Ontologized CP Design

The analysis was executed by ontologizing CP design activity. Development of the ontology was done with the ontology development environment “Hozo” [4]. The analysis adhered to the following steps.

- Interviews with medical doctors about employing CP design activities in practice.
- Ontologizing methods of CP design by the knowledge engineer (or the author).
- Re-interview using the ontology about what difficulties exist and when they appear.

In the second step, the author builds hypotheses about the difficulties, which are confirmed in third step. Confirmations are made using the ontology, which is interpreted by the author. Reasons for using ontology in this process are to improve the precision of discussion with medical staff members.

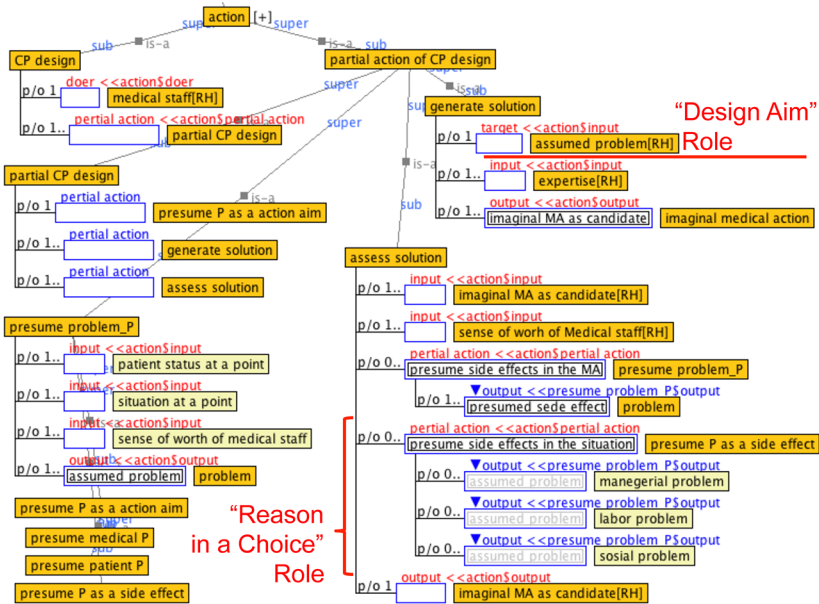


Fig. 1. Concept Definition of CP Design Activity

The interview process revealed the following difficulties:

- Medical problems are discussed adequately, but patient problems are under discussed.
- Managerial and labor problems are sometimes revealed, but there is no clear method for recording them as design intentions.
- The grain-size of the problems told and recorded in CP design is unclear.
- Some problems produce risks, but methods for expressing risk in CP design are unclear.

3.2 When Problems Are Revealed in Design Activity

Figure 1 outlines the concept definition of CP design activity. This model indicates that the problems assume two types of roles. One of these is “design aim” and the other is “reason in a choice.” The DESIGN AIM role is demonstrated by POS problems, and the REASON role is demonstrated by all types of problems. It is essential that the support system externalizing problems supposed by CP designers should be developed on the basis of this understanding, that the problems play two types of roles. The support method for this will be discussed in Section 4.

3.3 How to Decide the Grain-size of the Problems

Deciding the grain-size of problems presents some difficulties. In this section, the author considers one difficulty, stemming from the fact that a problem has several causes. For example, we think about a problem like “bleed.” We can find several underlying causes for this problem, such as “operative wound” or “abnormality after operations.” Should we divide the overall problem into categories for each potential cause? This is the question.

This question relates to the larger issue of how to decide the identity of a problem in an elementary sense. However, the absolute standard for deciding the identity of a problem is not considered in this research. The author operates from the approach that the identification of problems can be decided by the need to explain why a medical action exists. In the previous example, the problem is expressed in one or two instances, and solutions are decided by the level of detail that each CP designer wants to provide in explaining the intentions of medical activities.

3.4 How to Deal with Risk Problems

There are several types of problems, and some of these exist as risks. However the “risk” concept carries several meanings, and can cause confusion in CP design or modeling. In this section, the author considers the risk concept, and shows how to deal with problems as risks in modeling.

The following arguments were obtained through the interview process.

- CP is a plan, and then all problems exist as risk. (A)
- CPs are designed based on an intended problem. When the practitioners use the CPs, problems that are not intended impose risk. (B)
- CPs have a time-line. If we stand on a point of time, the problems that we expect to face in the future become risks. (C)
- If we stand on a point of time, there are problems that absolutely occur and problems that may potentially occur. The latter ones are risks. (D)

The modeling policy below addresses forming an understanding of these risks. If we have (A) understanding, then all problems are potential risks. It has no meaning. If we have (B) understanding, then the CP has no problems. Then the modeling policy does not include (A) and (B). With consideration to (C), a basic agreement of CP design reduces the complexity of the risk concept. The agreement is that a CP defines the ideal process of medical treatment. Based on this agreement, we need not deal with (C). Only (D) should be taken into account. The modeling policy is summarized as follows:

- Problems possess the “risk” attribute. The value is “Yes or No.”
- The “risk” attribute of a problem is fixed when the problem occurs.
- Realization of the risk problem is not described in CP. It is described as the condition on misuse of the CP.

4 Support System for Externalizing Sense of Worth of CP Designers

4.1 Methodology of Sense of Worth Externalization

As already discussed in Section 2, the author assumes the problems, which are assumed in CP design as the sense of worth of CP designers. However, the supposed problems are vague, and only some are discussed in CP design. Therefore, methodology of sense of worth externalization is required.

As the author mentioned in Section 3.2, the externalization is done in two phases. First, this methodology focuses on the “design aim” role, which is demonstrated by problems, and the system makes the users (or CP designers) model relationships between problems and medical actions (Section 4.2). Next, the system compares problems in several CPs and describes common design aims and different medical actions. Afterward, the system makes the users answer questions related to advantages and disadvantages (Section 4.3). Answers are formulated using the problem ontology. In this interview process, the system supports obtaining problems, which play the “reason in a choice” role in CP design.

4.2 Visualization of Relationships among Problems and Medical Actions

Figure 2 shows a user interface. This interface visualizes relationships among medical activities and problems, which play the “design aim” role.

The horizontal axis is the time sequence. Medical activities are at the top of this interface. Problems that play the “design aim” role appear at the left side. The circles on the intersection of the horizontal and vertical lines refer to the relationship between problems and medical actions. For each medical action, an order of priority for the problems is set by the users. It shows the CP designer’s thoughts about the seriousness among problems at the moment.

4.3 Reflection Support Based on the Visualization of Difference among Supposed Problems

The system semi-automatically generates a view (called “Focus-view”). This view shows differences in medical activities. When system users (or CP designers) select problems in two CPs displayed on each overview, the system generates this view. Figure 3 is an example. Using this view, the system interviews the users. Two questions are broadly handled broadly in this interview. These questions are “Are there any other supposed problems?” and “Do the medical actions lead to any problems as side effects?.” The users add problems to the model and set their order of priority. In this way, the users reflect their thoughts, which can be vague or hidden behind CPs, and externalize them as problems.

The example in Figure 3 shows the difference between two CPs, which are used in two of internal medicine departments in the Miyazaki University Hospital. The CPs are for the operation “endoscopic sub mucosal dissection for

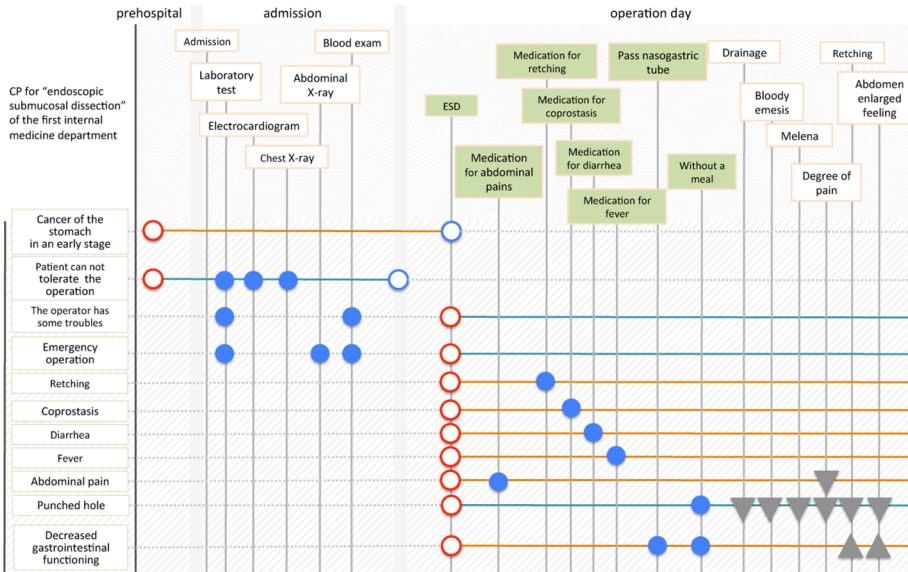


Fig. 2. Overview

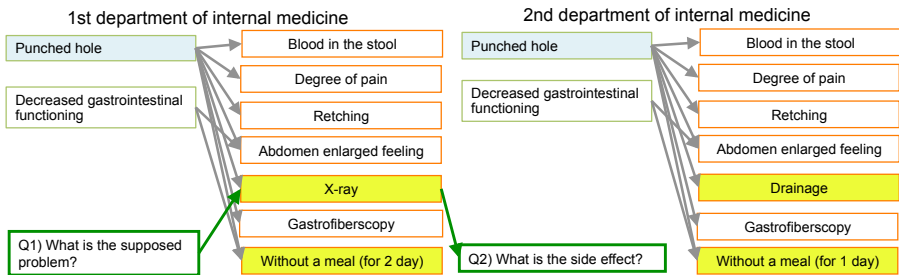


Fig. 3. Focus-view

early stage cancer of the stomach,” and the focused problem is “punched hole” (which happens during the operation, thinning the stomach wall). Doctors in each department answered the interview.

After this, highlights from the interview and some considerations about it are shown. Regarding the medical action “observation on the drainage through the nasogastric tube,” the doctor in the first department made the comment, “In our department, the nasogastric tube brings discomfort to patients. Then, we test the blood in the stool instead of using the tube and observing its drainage.” On the other hand, the doctor in second department made the comment, “Using the drain tube makes it easier to detect bleeding early, and it enhances safety of patients.” From these comments we can see the difference in their thoughts

about comfort and safety staying in the hospital. In addition, regarding the medical action “X-ray after the operation,” the doctor in the second department made the comment, “It is not indispensable. We cannot find a punched hole through an X-ray.” On the other hand, the doctor in the first department made the comment, “This X-ray is mainly to detect aspiration pneumonitis, and through this X-ray we can observe the punched hole at the same time.” After this interview, the author asked the doctor in the second department about the necessity to suppose the problem “aspiration pneumonitis.” The reply was that “If we have a minor case of aspiration pneumonitis, we need not describe it on CP. Of course, we check it in practice.” Detailed research about what takes place in each department is ongoing. Nevertheless, we can see the differences in thinking about the typicality of the CP. This provides some hints for ways to improve medical service.

The investigation of knowledge and the effectiveness of the system are ongoing. When the last trial was done in each department, staff members hesitated to communicate their thoughts about the CP of the other department. Part of the reason for this, they said, was “The medical staff cannot talk their thought without evidence.” Of course, this comment is reasonable considering the implementation of services to each individual patient. However, in the design phase and considering the necessary sense of worth (that is subjective), this attitude acts as a barrier. Breaking through this barrier is an issue in this investigation.

5 Conclusion

In this article, the author analyzed clinical pathway design activity, which is problem oriented. In the analysis, CP design activity was ontologized and difficulties surrounding the externalization of problems arose. Regarding the challenge of “how to decide the grain-size of problems,” we approached this dilemma by assuming that the grain-size is decided by the need to explain the intent of related medical activities. Regarding the difficulty of “how to describe the risk attribute of problems,” the author demonstrates the policy that risk attributes do not change in the modeling. Moreover, a method that externalizes the intended problem as the sense of worth of medical staff members is sketched. This method was explored in trial through interviews in the Miyazaki University Hospital, providing knowledge hinting toward ways to improve medical services.

References

1. Berry, L., Seltman, K.: *Management Lessons from Mayo Clinic: Inside One of the World Most Admired Service Organizations*. McGraw-Hill (2008)
2. Leonard, D., Swap, W.: *Deep Smarts*. Harvard Business School Press (2005)
3. Weed, L.L., Zimny, N.J.: The problem-oriented system, problem-knowledge coupling, and clinical decision making. *Physical Therapy* 69(7), 565–568 (1989)
4. Kozaki, K., et al.: Hozo: an ontology development environment - treatment of “role concept” and dependency management. In: *Proc. of Posters and Demos of the 4th International Semantic Web Conference, ISWC 2005* (2005)

Crowd-Sourced Knowledge Bases

Yang Sok Kim, Byeong Ho Kang, Seung Hwan Ryu, Paul Compton,
Soyeon Caren Han, and Tim Menzies

School of Computer Science and Engineering,
The University of New South Wales, Sydney, 2001, New South Wales, Australia
{yskim, compton}@cse.unsw.edu.au

Abstract. Crowdsourcing is a low cost way of obtaining human judgements on a large number of items, but the knowledge in these judgements is not reusable and further items to be processed require further human judgement. Ideally one could also obtain the reasons people have for these judgements, so the ability to make the same judgements could be incorporated into a crowd-sourced knowledge base. This paper reports on experiments with 27 students building knowledge bases to classify the same set of 1000 documents. We have assessed the performance of the students building the knowledge bases using the same students to assess the performance of each other's knowledge bases on a set of test documents. We have explored simple techniques for combining the knowledge from the students. These results suggest that although people vary in document classification, simple merging may produce reasonable consensus knowledge bases.

Keywords: crowd-sourcing, re-useable knowledge, knowledge acquisition, document classification.

1 Introduction

Crowdsourcing or human computation uses human power to solve specific problems. Many tasks such as natural language processing[1], image annotation [2], character recognition [3], sentiment analysis [4], and ontology development [5] are conducted successfully using this technology. Current approaches for crowdsourcing only exploit the direct contributions from a crowd, and do not collect the knowledge provided in some re-useable way. Obtaining re-useable knowledge from crowds has similarities to earlier work on combining knowledge from multiple experts [6-9], but there are also significant differences. Expertise is rare and valuable so work on combining knowledge from experts is generally concerned with a small number of experts. Secondly, experts are chosen for expert system development, whether single or multiple experts, because of their expertise, whereas crowdsourcing generally depends on common sense “expertise” and some members of the crowd may be erratic or biased.

This paper does not aim to produce a solution for crowd sourcing but to examine some of the challenges involved. A document classification problem was chosen for this purpose, since people readily understand documents and readily classify them, although their classification may be more idiosyncratic than consensual. Previously

we have developed a document classification system using Ripple-Down Rules (RDR) [10]. In building the knowledge base, the user chooses words from a document as the condition part of a rule and specifies the class as the conclusion part. For example, if a document contains the word ‘iPad’ then it might be classified under ‘Tablet’. New rules and refinements or corrections are added when a document is not classified appropriately. From previous experience a laypeople (who might provide crowdsourcing common sense judgements), can readily learn the classification system and build sizeable document classification knowledge bases. A classification evaluation study with domain specific documents showed individuals agreed about 90% of the system classification suggestions after classifying about 1000 documents [11]. However, this does mean that this user’s classifications would be accepted by others; so the aim of these studies here was to look at differences and explore ways of merging knowledge bases.

In the study here, lay people (students) created their own knowledge bases to classify a set of documents, and the resulting knowledge bases were used to classify a set of evaluation documents. Classification results from each knowledge base for the evaluation documents were evaluated by other participants. Using these datasets we then investigated simple ways of aggregating the knowledge obtained.

2 Related Work

2.1 Crowdsourcing

Crowdsourcing aims to resolve tasks traditionally performed by individuals using a group of contributors through an open call [12-13] and is sometimes referred to as ‘human computation’ [14-16]. Early crowdsourcing was offline. For example, the Oxford English Dictionary (OED) made an open call to the contributors to index all words in the English language and in a 70 year project, received over 6 million submissions [17]. Recently, crowdsourcing has been conducted via the internet. Amazon’s Mechanical Turk (MTurk) (www.MTurk.com), an internet marketplace for crowdsourcing, provides a successful environment for crowdsourcing [18]. The Mechanical Turk has had significant impact and a Google scholar search returns about 58,000 results (9/Jan/2012). Even though crowdsourcing is increasingly widely used there are problems such as exploitation and abuse of labour and cost inefficiency [13]. The crowdsourcing process normally consists of three components: assignment of the problem, aggregation of contributions, and remuneration for contributions [19]. Aggregation of contributions to produce the desired outcome is a critical step [20]. The crowd may *explicitly* contribute to the specific problem by evaluating, reviewing, voting, and tagging; by sharing items, textual knowledge, and structured knowledge; by networking (i.e. LinkedIn, MySpace, and Facebook); by building artefacts such as software, textual knowledge bases, structured knowledge bases, systems, etc.; and executing tasks. Crowds also *implicitly* contribute to specific problems by acts such as submitting keywords, buying products, browsing web sites, etc [12]. In this project we have looked at crowds *explicitly* contributing to a document classification problem by creating their own knowledge bases. This is critically different from other crowd

sourcing as it has the potential to produce re-usable knowledge able to carry out the same document classification task. Crowd contributions can be aggregated by *integrative* or *selective* processes [21]. *Integrative* crowdsourcing pools complementary input from the crowd (e.g., the Android market, iStockphoto, YouTube, and Wikipedia) or the collective opinion of the crowd (e.g., Delicious, Digg, and the Google Image Labeler). Once individual contributions meet certain quality requirements, they are used to get the final outcome. Conversely *selective* crowdsourcing processes choose the ‘best’ contribution from among a set of options (e.g., 99designs, Atizo, InnoCentive, the Netflix Prize, and the Dell IdeaStorm platform) [20-21].

2.2 Knowledge Acquisition from Multiple Experts

Knowledge acquisition from multiple experts has long been proposed to avoid some of the pitfalls of relying on a single expert or improve knowledge acquisition performance [22]. There have been a range of suggestions for the aggregation of knowledge from multiple experts using techniques such as Bayes’ theorem [6], Bayesian Network Inference [23], and Semantic Networks [24]. The key difference between knowledge acquisition from multiple experts and knowledge acquisition from a crowd is that the domain experts by definition are highly qualified, and the opinion of any single expert should not be ignored lightly. Secondly, because experts are involved there are likely to be very few of them – it is difficult enough finding a single expert to be involved in a knowledge based systems project. In contrast with a crowd there is likely to be a range of skill, and also a range of diligence in how the task is undertaken – and assessing the quality of what is provided is a major issue in conventional crowdsourcing [1]. As will be seen below the range of expertise and perhaps diligence from our crowd is very wide. Secondly, crowds are likely to be large. We used a crowd of 27 and although this is tiny in terms of normal crowd sizes, it is far larger than any multiple expert studies we are aware of. The closest to crowd-sourced knowledge acquisition is the work of Richardson and Domingos [23], using first order probabilistic reasoning to combine the first-order knowledge of a crowd. However, experiments were carried out using simulation studies and similar to other multiple expert human studies, the human study they carried out contained only four volunteer experts.

3 Crowd Sourced Knowledge for Document Classification

Fig. 1 illustrates how we carried out experiments. The contributors C_1, C_2, \dots, C_N create N knowledge bases to classify a large number of training documents. Evaluation documents are then processed by the knowledge bases and evaluators E_1, E_2, \dots, E_N , (who may be other contributors) who agree or disagree with the classifications provided by the knowledge bases. We then combined the knowledge bases from all the contributors in various simple ways to produce an overall crowd-sourced knowledge base and again used the evaluation results to assess the quality of the merged knowledge.

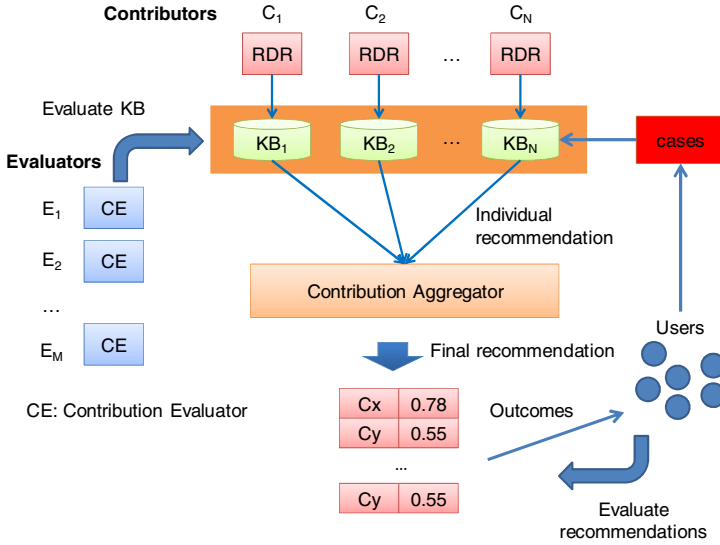


Fig. 1. The Knowledge-Based Crowdsourcing for Document Classification Problem

3.1 The RDR Document Classifier

The RDR Document Classifier is based on the research by Kim, Park et al. [10] and was developed to classify web pages collected by a web monitoring system [25]. It supports *hierarchical classification*, since people usually organize their classes, concepts, and the categories they choose into a hierarchical structure [26] and research results showed that hierarchical classification significantly improves classification performance [27-29]. The RDR Document Classifier maintains a hierarchical class structure using an n-ary tree, which we call a hierarchical class tree. Each classification rule is formed as shown in **Listing 1**.

Listing 1. Classification Rule in RDR Classifier

```

RULE x
IF
    LAST_FIRED_RULE : RULE i
    CONDITION : {w1, w2, ..., wN}
THEN
    CONCLUSION : CLASS A
    
```

The “RULE x ” represents the unique identifier of a rule. The “LAST_FIRED_RULE” part shows the context how “RULE x ” was acquired. That is, “RULE x ” was acquired to correct a wrong conclusion, “RULE i ”. Semantically this structure also means that “RULE x ” is an exception to the last fired rule, RULE i . A new rule is only acquired when the current knowledge base suggests a wrong recommendation(s). No recommendation is also considered a wrong recommendation. If a

recommendation is not given, the “LAST_FIRED_RULE” is at the ‘root’ of the knowledge base. The “CONDITION = $\{w_1, w_2, \dots, w_N\}$ ” consists of disjoint N words from a document (e.g., ‘Google’ AND ‘Android’). The ‘CONCLUSION’ part represents a class, CLASS A in a class tree, where the document that satisfies the rule will be located by the system. Each rule specifies only one node of the hierarchical class tree.

The inference process starts with four variables –document D , current rule R_c , temporary conclusion list C_t , and final conclusion list C_f . At the beginning R_c is the root rule and C_t and C_f are NULL. First the inference engine retrieves all child rules of the given current rule. If there are no child rules, the current rule is added into C_f . If there are child rules, the inference engine evaluates all child rules and adds the fired rules to C_t . If no child rule is fired, the inference engine adds the current rule to C_f . After processing the child rules, the inference engine examines whether or not C_t is empty. If C_t is empty, the inference engine returns C_f ; otherwise a rule R_j is retrieved from C_t and set it as the new current rule ($R_c = R_j$) and the inference process is performed with the new current rule. More details on the RDR Document Classifier approach are provided in [10].

3.2 Evaluation

The contributors C_1, C_2, \dots, C_N go through all the training documents and each builds a knowledge base to give what they believe are appropriate classifications for the training documents. There is no control over how carefully they consider each document. Particularly, for later documents they may simply glance at a document and decide to accept whatever conclusion their knowledge base assigns. The evaluation documents are classified by each contributor’s knowledge base and the classification results are then evaluated by other contributors (the evaluators). For each evaluation document and for each evaluator five conclusions from other contributors’ knowledge bases are randomly selected (excluding the evaluator’s knowledge base). The evaluator simply agrees or disagrees with each of the five classifications of the evaluation document. From the resulting data the percentage agreement with the conclusions from any contributor’s knowledge base can be obtained. Alternatively the evaluator can be evaluated by seeing how often their evaluation agrees with other evaluators.

3.3 Knowledge-Base Aggregation

We combined the knowledge bases in very simple ways:

- **Simple Voting (SV):** A document is processed by all the contributors’ knowledge bases. The best classification is considered to be the one most frequently assigned by the various knowledge bases. This scheme does not consider differences in expertise between contributors.
- **Weighted Voting (WV):** The same protocol is then followed as for simple voting except each classification provided by a particular knowledge base is weighted

by the evaluation score for that knowledge base. Again we take the highest ranked classification as the classification for the document.

- **Best Contributor (BC):** Here we simply take the classification for the document provided by the knowledge base with the highest evaluation score and which provides a classification for the document. If the best knowledge base provided a classification for all documents, then its classification would be used for all documents. In practice the best knowledge base will not classify all documents, so the best knowledge base for a particular document may not be the best overall.

4 Experimental Design

4.1 Document Classification

A total of 27 Masters students participated in the experiment to classify a total of 1,100 documents which were collected from five Australian online news websites: The Daily Telegraph, The Australian, News.com.au, The Age, and Sydney Morning Herald, and also the BBC news website. Each article has a title and main content, and the words from either could be used as rule conditions. The documents were divided into training set of 1,000 documents and an evaluation set of 100 documents.

Table 1. Classes given to the Contributors before Classification

State and Territory	Prime Minister and Cabinet Agencies / Australian Capital Territory/ New South Wales/ Northern Territory/ Queensland/ South Australia/ Tasmania/ Victoria/ Western Australia/ Norfolk Island
Australian Government	Commonwealth Parliament/ Agriculture, Fisheries and Forestry/ Broadband, Communication and the Digital Economy/ Climate Change and Energy Efficiency/ Defence/ Education, Employment and Workplace Relations/ Families, Housing, Community Services and Indigenous Affairs/ Finance and Deregulation/ Foreign Affairs and Trade/ Health and Ageing/ Human Services/ Immigration and Citizenship/ Infrastructure and Transport/ Innovation, Industry, Science and Research/ Resources, Energy and Tourism/ Sustainability, Environment, Water, Population and Communities/ Treasury/ Courts/ Attorney-General

Twenty seven Master of Computing course students were asked to use the RDR Document Classifier to build rules to appropriately classify the training set. The participants were asked to classify the documents as if they were a librarian who selects news articles for various government departments and/or agencies. Information about the Australian government presented in <http://australia.gov.au/directories/australian-government-directories/portfolios-departments-and-agencies> was used as a hierarchy of department and agencies. A class tree specifying this top-level hierarchy of

departments was provided at the beginning of the experiment. The tree consists of two top categories, “Australian Government” and “States and Territory”. A total of 19 sub-classes are under the “Australian Government” class and 10 sub-classes are under the “States and Territory” class (see **Table 1**). Students were expected to classify documents into this hierarchy, but could create further sub-classes. We suggested that participants might create sub classes up to four levels, and suggested they refer to the website above, but they were free to create any depth of classes and any classes they liked. Document titles were presented as a list and titles could be clicked to access the full document. Participants were expected check the content of a document before classifying it, but this was not controlled.

4.2 Evaluation of Individual Knowledge Bases

A total of 36 Masters students participated in the experiment. Each student’s knowledge base was evaluated by five other students. Since a knowledge base might provide multiple classifications for a document we randomly selected one of these classifications for the evaluation.

Evaluation of Contributors. After evaluation, we calculated an *error rate* for each student’s knowledge base which gives the proportion of all evaluated classifications not agreed to by the evaluators. It is defined as

$$Error = \frac{C_{disagree}}{C},$$

where C is the number of evaluations provided by all evaluators for the classifications of a student for the evaluation set and $C_{disagree}$ is the number of evaluations disagreed by the evaluators. Given there were 100 documents C is 500. If a knowledge base does not provide classification result for a document, we displayed “no classification” as conclusion.

Evaluation of Evaluators. The *accept rate* is the proportion of all classifications agreed to by an evaluator. It is defined as

$$Accept = \frac{E_{accept}}{E},$$

where E is the number of evaluations conducted by an evaluator, and E_{accept} is the number of evaluations where the evaluator agreed with the classification given. In this experiment, the evaluator should evaluate five classifications per evaluation document, so E is 500 (100 documents \times five evaluators). We would not expect the accept rate to be a linear indicator of evaluator performance, as both evaluators who agreed with none of conclusions given and evaluators who agreed with all the conclusions given are probably doing a poor job.

Aggregation of Knowledge Bases. The knowledge bases were then combined as above, but for simplicity, classes created by the contributors under the predefined sub-classes were subsumed into their parent class. For example, the ‘visa’ class under

'Immigration and Citizenship' class was subsumed into the latter class. In the following discussion we are only concerned with aggregation with respect to the top-level classes that all participants were expected to use. Error rates between top-level class and other classes were compared to assess the effect of aggregation.

5 Results

5.1 Individual Knowledge Base Characteristics

The results indicate considerable differences in the way knowledge was added. **Fig. 2** shows that number of sub-classes added ranged from 0 to 113. It is difficult to know whether this is because of differences in knowledge about the documents, a different view on the detail required to produce a useable hierarchy, or whether some students were unable or unwilling to devote more than minimal time to the task. On average contributors created 49 new classes for their classification task.

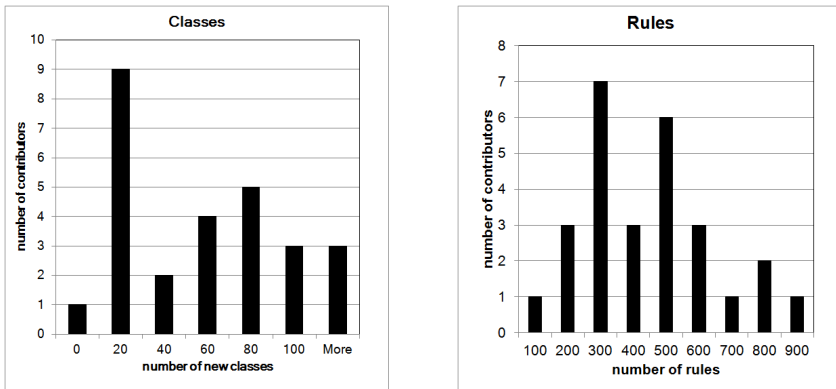


Fig. 2. Class and Rules

The number of rules constructed is another area of difference ranging from 89 to 884. On average the contributors created 396 rules with most contributors creating between 200 and 500. Those who create more classes do not necessarily create more rules. That is, there is no significant correlation between the number of classes and the number of rules ($r=0.13$).

Most contributors use from one to four condition words, but three contributors used more than five condition words per rule (**Fig. 3**). The contributors who created more rules tended to use more condition words, and the correlation between the number of rules and the number of condition words is relatively high ($r = 0.47$). However, there is no correlation between the number of new classes and the number of condition words ($r = 0.01$).

As the RDR Document Classifier allows multiple classifications, the number of classifications is greater than the number of documents. On average the knowledge bases provided 2,817 classifications for 1000 documents; on average 43.6 articles per class

and 2.8 classifications per article. The maximum classification count is 7,016 and the minimum is 1,004. The number of classifications seems to depend more on the number of classifications per document than the number of new classes added as the number of classifications is not correlated with the number of new classes ($r = 0.08$).

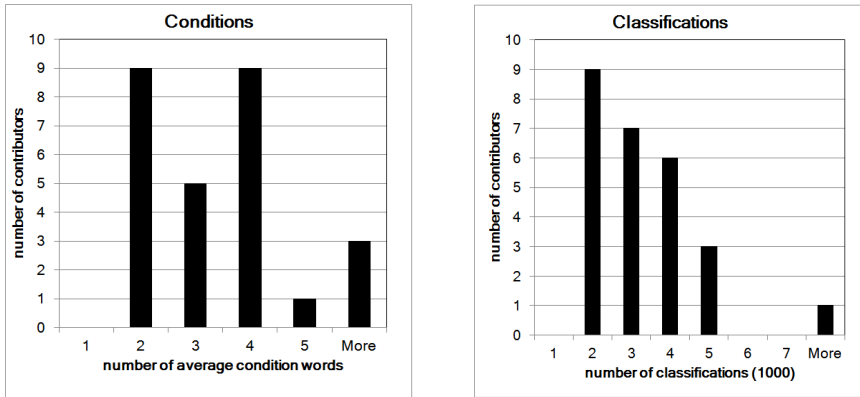


Fig. 3. Conditions and Conclusions

5.2 Evaluation of Individual Knowledge Bases

It is possible that the evaluation set may contain documents that are different from those in the training data so that the knowledge bases would be unlikely to classify such documents. To try to deal with this we looked at performance on all documents and excluding results were a knowledge base failed to classify a document. The results are summarized in Fig. 4. Most error rates for knowledge bases are distributed in the range of 30.0% to less than 70.0%, but the two highest errors were 84% and 93%. The error rates decrease, when ‘no class’ classifications were removed. The average error rate with all classification is 44.8%, but average error rate with the “no class” classification removed is 41.7%.

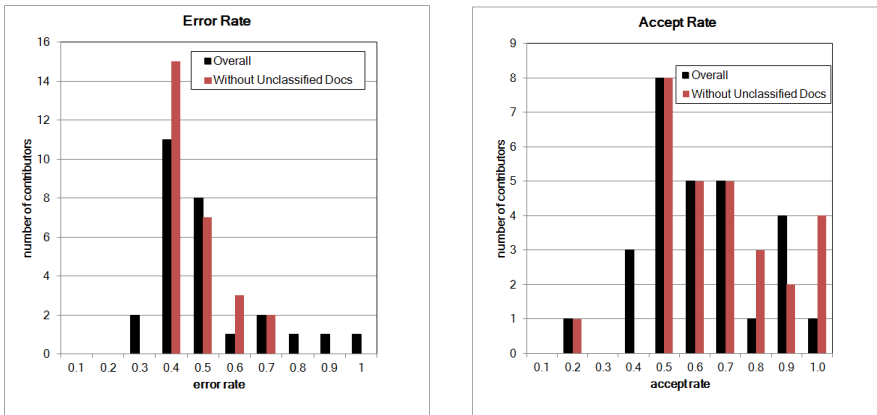


Fig. 4. Error Rate and Accept Rate Distribution

Fig. 4 also shows the accept rate which is a measure of how often the evaluators agree with the conclusions given. There are clearly more evaluators at very high agreement levels than knowledge bases at very high error levels. What is of interest is that the people who give very high levels of agreement also have the highest errors and overall there is a high correlation between error and agreement ($r=0.74$). The two people who had the highest error rates for their knowledge bases also had the highest agreement rates. Initially one might expect that if a person knew very little about the domain they would agree with about 50% with the evaluations; however, the data seems to suggest that if they know very little about the domain, they will agree with anything. On the other hand these may have been students who put little effort into the project in either building knowledge bases or assessing.

The overall error rate has no relation with the number of classes ($r=-0.15$). However, the error rate excluding ‘no class’ errors shows a slightly higher, but not significant, negative correlation with the number of classes ($r=-0.27$). The number of rules does not have a significant negative relation with the overall error rate ($r=-0.20$); however, the error rate excluding ‘no class’ errors shows a significant negative correlation with the number of rules ($r=-0.56$). This perhaps suggests that the contributors who create more rules probably approach the task thoughtfully and thus have smaller error rates.

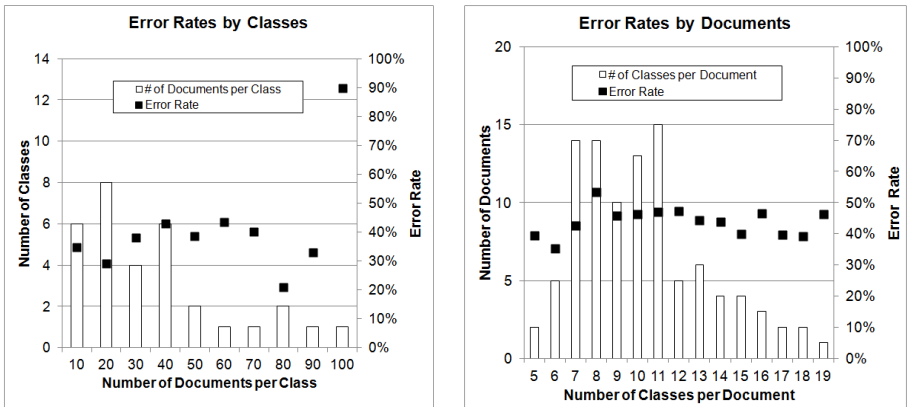


Fig. 5. Error Rate by Classes and Documents

It is interesting to speculate how knowledge base performance might relate to the characteristics of the concepts. Firstly, error rates may differ between classes perhaps because some classes may represent more specific concepts than others, also if a class has more documents, it may represent a more general concept. However as shown in **Fig. 5**, no clear picture emerges apart from very high error rates (90%) for the 100 documents with a no-class classification. This is simply because every document should be able to be classified somewhere.

5.3 Contribution Aggregation Results

We considered three methods of combining the classifications from the knowledge bases – simple voting (SV), weighted voting (WV) and best contributor (BC). These approaches are unlikely to be the optimal solution for aggregation, but they give surprisingly good results. Applying the various methods we found the single best conclusion according to that method for a specific document. As discussed this was necessarily a top-level classification, because lower level classifications were subsumed into their top-level class.

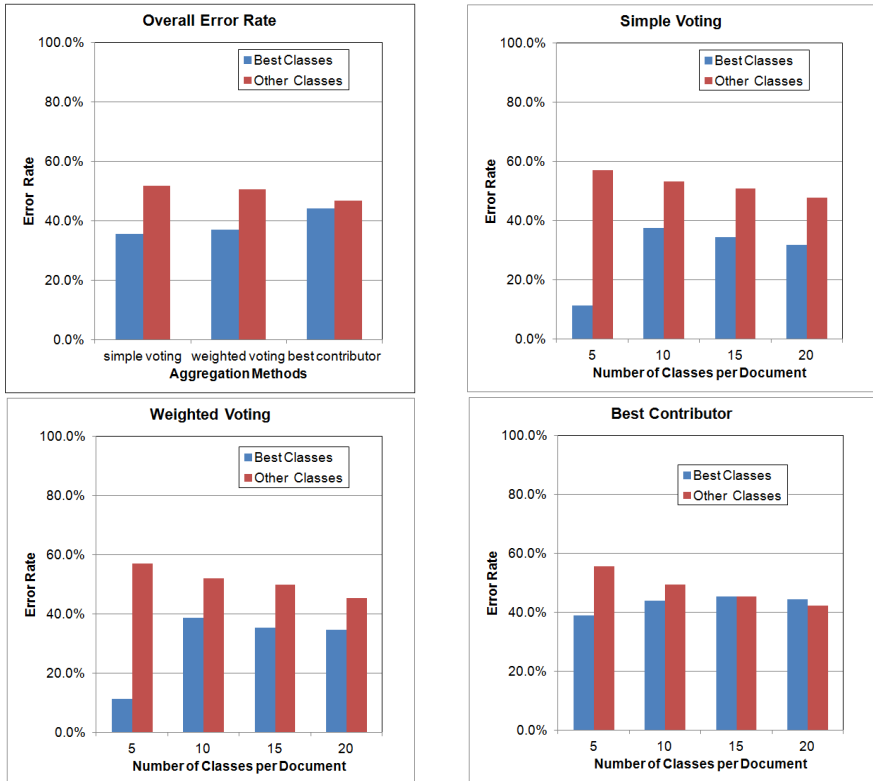


Fig. 6. Contribution Aggregation Results

In Fig. 6, we compare the error rate with the best class to the error rate with the other classes, where documents were grouped as having 5 or less classes, 5-10 classes and so on. If all documents are considered regardless of the number of classes the difference between the best classes and other classes is 16.2%, 13.7%, and 2.7% respectively for simple voting, weighted voting and best contributor. Simple voting and weighted voting show very low error rate for the documents that were classified into the small number of classes, but higher error rate when documents have larger numbers of classes.

6 Discussion

In this research, a number of lay contributors (students) have created knowledge bases to classify documents with the classifications being what they thought a government librarian might consider appropriate. The contributors evaluated each others' knowledge bases by agreeing or disagreeing with how they classified a set of test documents. The results were fairly poor in that the average error rate or disagreement rate was 45%. This did not suggest promising results for merging the knowledge bases. Surprisingly, however, when the merged classification selected for a document was the most frequent classification or the weighted most frequent classification, the agreement that such classifications were appropriate was close to 90% for documents with a small number of classes. This would clearly provide an appropriate classification framework from a librarian perspective.

Interestingly the agreement with the other classifications (i.e. excluding the best classification), was lower for documents with only a small number of classifications assigned than for documents with a larger number of classifications. We hypothesise that this is related to the type of errors being made. If a document has only a small number of classifications assigned, this is perhaps because the appropriate classification for the document is reasonably clear-cut and the selection of different classifications is not because of carefully considered different choices, but simply errors and perhaps carelessness. In contrast as the number of classes per document increases, the level of agreement increases. This might be because a greater number of classes indicates the document can be classified a number of ways or because it is unclear how best to classify it. Either of these might lead to increasing agreement when a document has more classes.

Perhaps surprisingly there was a lower level of agreement with taking the classification from the best contributor, that is the best knowledge base overall which provided a conclusion for that document. Just because a contributor got the highest agreement score for a particular document, does not mean that they chose the conclusion that most people preferred.

One limitation in the way we carried out the study was that lower level classifications were subsumed into upper level classifications. So although simple voting techniques might be appropriate for upper level classifications, they may not be appropriate more generally. In future work we will look at more sophisticated merging functions and will also allow the contributors to disagree with another contributor's conclusion by writing rules to correct that conclusion. This will effectively merge the assessment of errors from a knowledge base and the assessment of evaluators, as the rules a contributor adds, will not only specify why they chose a particular conclusion but why they disagreed with another contributors classification. At this stage, however, it is an open question of how to merge rules in this way.

Acknowledgement. This research has been funded by an Australian Research Council Discovery Grant and Asian Office of Aerospace Research and Development (AOARD).

References

- [1] Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263. Association for Computational Linguistics, Honolulu (2008)
- [2] Chen, K.-T.: Human Computation: Experience and Thoughts. In: *CHI 2011 Workshop on Crowdsourcing and Human Computation Systems, Studies and Platforms (2011)*
- [3] Ahn, L.V., Maurer, B., Mcmillen, C., Abraham, D., Blum, M.: reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321(12), 1465–1468 (2008)
- [4] Brew, A., Greene, D., Cunningham, P.: Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In: *Proceeding of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, pp. 145–150. IOS Press (2010)
- [5] Lin, H., Davis, J., Zhou, Y.: Ontological Services Using Crowdsourcing. In: *21st Australasian Conference on Information Systems (2010)*
- [6] O'Leary, D.E.: Knowledge Acquisition from Multiple Experts: An Empirical Study. *Management Science* 44(8), 1049–1058 (1998)
- [7] Medsker, L., Tan, M., Turban, E.: Knowledge acquisition from multiple experts: Problems and issues. *Expert Systems with Applications* 9(1), 35–40 (1995)
- [8] Turban, E.: Managing knowledge acquisition from multiple experts. In: *IEEE/ACM International Conference on Developing and Managing Expert System Programs, Washington, DC, USA*, pp. 129–138 (1991)
- [9] La Salle, A.J., Medsker, L.R.: Computerized conferencing for knowledge acquisition from multiple experts. *Expert Systems with Applications* 3(4), 517–522 (1991)
- [10] Kim, Y.S., Park, S.S., Deards, E., Kang, B.H.: Adaptive Web Document Classification with MCRDR. In: *International Conference on Information Technology: Coding and Computing (ITCC 2004)*, pp. 476–480 (2004)
- [11] Park, S.S., Kim, Y.S., Kang, B.H.: Personalized Web Document Classification using MCRDR. In: *The Pacific Knowledge Acquisition Workshop, Auckland, New Zealand (2004)*
- [12] Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the World-Wide Web. *Communications of the ACM* 54(4), 86–96 (2011)
- [13] Zhang, L., Zhang, H.: Research of Crowdsourcing Model based on Case Study. In: *8th International Conference on Service Systems and Service Management (ICSSSM)*, pp. 1–5. IEEE, Tianjin (2011)
- [14] Das, R., Vukovic, M.: Emerging theories and models of human computation systems: a brief survey. In: *Proceedings of the 2nd International Workshop on Ubiquitous Crowdsourcing*, pp. 1–4. ACM, Beijing (2011)
- [15] Heymann, P., Garcia-Molina, H.: Turkalytics: analytics for human computation. In: *Proceedings of the 20th International Conference on World Wide Web*, pp. 477–486. ACM, Hyderabad (2011)
- [16] Little, G., Chilton, L.B., Goldman, M., Miller, R.C.: TurkKit: human computation algorithms on mechanical turk. In: *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology*, pp. 57–66. ACM, New York (2010)
- [17] Winchester, S.: *The Surgeon of Crowthorne: A Tale of Murder, Madness and the Oxford English Dictionary*, Penguin (1999)

- [18] Buhrmester, M., Kwang, T., Gosling, S.D.: Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1), 3–5 (2011)
- [19] Davis, J.G.: From Crowdsourcing to Crowdservicing. *IEEE Internet Computing* 15(3), 92–94 (2011)
- [20] Geiger, D., Seedorf, S., Schulze, T., Nickerson, R.C., Schader, M.: Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. In: *Americas Conference on Information Systems, AMCIS 2011* (2011)
- [21] Schenk, E., Guittard, C.: Towards a Characterization of Crowdsourcing Practices. *Journal of Innovation Economics* 7(1), 93–107 (2011)
- [22] Mittal, S., Dym, C.L.: Knowledge Acquisition from Multiple Experts. *AI Magazine* 6(2), 32–36 (1985)
- [23] Richardson, M., Domingos, P.: Building large knowledge bases by mass collaboration. In: *Proceedings of the 2nd International Conference on Knowledge Capture*, pp. 129–137. ACM, Sanibel Island (2003)
- [24] Puuronen, S., Terziyan, V.Y.: Knowledge Acquisition from Multiple Experts Based on Semantics of Concepts. In: Fensel, D., Studer, R. (eds.) *EKAW 1999. LNCS (LNAI)*, vol. 1621, pp. 259–273. Springer, Heidelberg (1999)
- [25] Park, S.S., Kim, S.K., Kang, B.H.: Web Information Management System: Personalization and Generalization. In: *The IADIS International Conference WWW/Internet 2003*, Algarve, Portugal, pp. 523–530 (2003)
- [26] Bagno, E., Eylon, B.-S.: From problem solving to a knowledge structure: An example from the domain of electromagnetism. *American Journal of Physics* 65(8), 726–736 (1997)
- [27] Dumais, S., Chen, H.: Hierarchical classification of Web content. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 256–263. ACM, Athens (2000)
- [28] Wibowo, W., Williams, H.E.: Simple and accurate feature selection for hierarchical categorisation. In: *Proceedings of the 2002 ACM Symposium on Document Engineering*, pp. 111–118. ACM, McLean (2002)
- [29] Liu, T.-Y., Yang, Y., Wan, H., Zeng, H.-J., Chen, Z., Ma, W.-Y.: Support vector machines classification with a very large-scale taxonomy. *SIGKDD Explor. Newsl.* 7(1), 36–43 (2005)

Social Issue Gives You an Opportunity: Discovering the Personalised Relevance of Social Issues

Soyeon Caren Han and Hyunsuk Chung

School of Computing and Information System,
Tasmania 7005, Australia
{Soyeon.Han, hchung}@utas.edu.au

Abstract. Social networking services have received a lot of attention recently so that the discussion of certain issues is becoming more dynamic. Many websites provide a new service that displays the list of the trending social issues. It is very important to respond to those social issues since the impact on organisations or people may be considerable. In this paper, we present our research on developing the personalised relevance identification system that displays the relevance of social issues to a target domain. To accomplish this, we first collected social issue keywords from Google Trends, Twitter and Google News. After that, we setup an electronic document management system as a target domain that would include all knowledge and activities having to do with a target object. In order to identify the relevance of the social issues to a target, we applied the Term Frequency Inverse Document Frequency (TFIDF). Our experiments prove that we can identify the meaningful relevance of social issues to targets, such as individuals or organizations.

Keywords: Google Trends, Social Issues, Social Networking Sites, Twitter, Trending Topic.

1 Introduction

Social networking services (SNS) have received a great deal of attention recently [15]. These services enable the users to communicate with others in a new way and reflect the users' real-life interests [16]. SNSs do not only change the way that people communicate but also increase the speed of sharing information. There are two reasons for the latter. Firstly, unlike other online communication services, SNSs provide push-based information. For example, while the e-mail is like a letter that a person places in somebody else's mailbox, so that it can be opened when the user wants to, SNS can be likened to the user tapping another person's shoulder and forcefully placing a message in the latter's hand. Secondly, SNS messages are broadcasted to all the people linked to the sender while e-mails are sent only to the addresses specified by the sender. As the speed of the communication flow has been increased by SNS, a large amount of information exists on the Web; and because people are social beings and they are curious about what others are doing, there are those who want to see what information people are looking for.

Many large Internet-based companies think of this as an opportunity. They provide the new trend service, based on the data that they have collected. These trend services display the list of top trending social issue keywords. For example, Google and Twitter provide the new service of showing the list of trending topics in Google Trends and Twitter Trending Topics, respectively. While Google Trends displays the list of top 10 fastest-rising search terms based on hourly data from Google Search, Twitter Trending topic provides the list of top 10 most discussed topics based on tweets in Twitter [10]. The flow of Twitter is influenced by the “big mouths” like celebrities or special groups. Google Trends, however, is based on the search results so that it is affected by general users. According to Rech, among the existing trending services, Google Trends provides a highly reliable list of social issues. Google Trends is a good indicator of the evolution of world interests in certain topics of search [13]. Therefore, if you define the social issues as “the events that many people are interested in”, the keywords displayed by Google Trends should be considered as representing people’s interests.

It is very important to respond to these social issues since the impact on an organisation or an individual may be massive. If the relevance of a certain social issue is known to them as soon as possible, it will enable them to take the opportunities and avoid the threats that such a social issue may present for them. Unfortunately, there is currently no service that shows the relevance to people of those trending social issues.

In this paper, we present our research on developing the personalised relevance identification system that displays the relevance of social issues to target domains, such as individuals or organizations. First, in order to identify the social issues, we collected social issue keywords from Google Trends over a period of 195 days, approximately 5 months. Each keyword from Google Trends represents a certain social issue. However, it is hard to define the exact meaning of each social issue by using a certain term from Google Trends since there may be an ambiguity. To reduce this ambiguity, we decided to extract several related keywords. Since the top 10 social issue keywords from Google Trends contain real-time information, the related keywords should also be collected from the services, which include real-time information, such as micro-blog or Internet news. Therefore, we chose Twitter and Google News as the related keyword extractors. In order to identify the relevance of social issues to a target, we applied Term Frequency Inverse Document Frequency (TFIDF). TFIDF is a common technique for calculating relevance weight. We will show the effectiveness of our method by conducting several types of experiments.

The paper is structured as follows: Section 2 presents the related work, followed by the methodology of this proposed system in Section 3. In Section 4, we describe the evaluations conducted and discuss the results. Finally, we conclude this paper in Section 5.

2 Related Work

In various aspects, social networking services have been researched including their characteristics [2, 16] and the reason why people are enthusiastic about them. In this

regard, there are many works that analyse the behaviour of SNSs: Putnam described it as a social capital maintainer. Boyd and Ellison investigated the difference between SNSs and other communication services, such as email or messenger [2]. There are some researchers who have analysed different types of SNSs, such as Facebook [7], YouTube [12], and Twitter [10, 15]. Having the SNSs drawn enormous interest in a short time span, trending social issues are becoming more dynamic. Because of this, tracking trends recently becomes the important issue in every field [13]. Many websites did not miss this opportunity and, consequently now provide the service that displays trending topics [16]. The method of trends tracking can be classified as three main sections: search-based [10, 16], social networking-based [10, 15] and news-based tracking trends [11]. Even if they use the same tracking method, the result would be different.

Unfortunately, it is hard to identify the exact meaning of a trending topic by using only a keyword from trends tracking services. It is necessary to utilize query expansion. Query expansion is widely-used in the field of information retrieval. The process of query expansion generally includes four steps: resources selection, seed query construction, search results review, and query reformulation [4]. Most researchers perform query expansion based on either local or global analysis [1].

To develop the personalized system with a certain target object, such as individual users or organizations, they always need to provide the digitalized domain. Fortunately, most activities of both individuals and organizations are now saved in assortment of digital information [9]. Most users utilize the information management system that enables them to manage their knowledge in a well-structured and categorized. Moreover, those systems offer centralized storage, which covers almost all activities of a target object, such as email [3], blog [6] or knowledge management system (KMS) [8].

To identify the relevance of a query to a certain document, string comparison and matching methods are briefly reviewed. A method of string matching that enables the system to make decisions using the actual content flow. This method applied in many pattern-matching and Web search areas [5]. There are several kinds of methods that are widely-used, such as the edit-distance method [14], Jaro-Winkler distance [5], Jaccard distance [18] and TFIDF distance [17].

3 Personalised Relevance Identification

In this paper, we present our research on proposing the method that identifies the personalized relevance of trends to target objects, such as individuals or organizations. To provide the personalized relevance identification system, the methodology employed in this research can be divided into four parts, as follows: (1) trending social issue keyword collection; (2) related keywords extraction; (3) personalized/adapted domain identification; and (4) relevance identification.

3.1 Social Issue Collection

The first phase deals with the collection of those trending social issues that show what people are currently most interested in. Fortunately, many websites provide the services that display the trending social issues. For example, Google, Yahoo, and Twitter provide the trends service that shows the list of trending topics in Google Trends, Yahoo Buzz, and Twitter Trending Topic, respectively.

Google Trends has been chosen as the social issue keyword collector in this paper. Based on hourly data from Google Search, Google Trends provides the list of the top 10 fastest-rising search terms. The search-terms indicate what topics people are interested in and looking for. It is evident that Google Search is currently the most popular search engine. Because of this, Rech claims that Google Trends adequately provides the most sought after terms and phrases [13]. Thus, Google Trends has been chosen as the trending topic collector in this study so that more accurate results will be obtained.

3.2 Related Keyword Extraction

Even though the top 10 trending social issue keywords per hour were collected, ambiguity occurs when the exact meaning of a trend topic is obtained by using each trend from Google Trends. For example, let's assume that "Michael" is one of the fastest-rising search terms in Google Trends. Most people may think that the keyword "Michael" is a popular American recording artist, entertainer, and pop star. The keyword "Michael," however, may be related to the retired American professional basketball player. Therefore, it is necessary to expand a trending keyword by extracting some related keywords. As Google Trends displays the list of fastest-rising search terms, which are considered as real-time social issue keywords, the related keywords must be extracted from services that publish real-time publishing, such as micro-blog and Internet news [10]. If related keywords are extracted from general documents published at any time, semantically related keywords will be extracted, not keywords that are related to the trending social issue.

In this paper, Twitter and Google News were chosen as the micro-blog and Internet news service, respectively. To extract the appropriate related keywords from those services, articles related to a Google Trends keyword were first searched. As it is necessary to extract documents related to an hourly-trending social issue keyword, we extract only articles that people uploads in an hour. After the articles collection, we applied Term Frequency (TF) to figure out the most relevant nouns on a Google Trends issue keyword. TF weight will be defined by dividing the occurrence count of a certain term by the total number of words in the given document [17]. Then, term weights are sorted in descending order. The higher the term weight, the more the keyword is related. The best number of related keywords will be analysed in the evaluation session.

3.3 Personalised Domain

After finishing the trending topic collection, it is necessary to obtain the digitalized document management system that contains all activities and knowledge regarding target objects, such as individuals or organizations. The document management system should be well-structured and categorized. The typical examples of a digitalized document management system are email, blog, and Knowledge Management System (KMS). Most document management systems are categorized the sections by genre. The way to categorize the document is a personal decision so that it might be subjective. However, the relevance will be viewed by people who classified that way so that it is not an issue.

Since there is some possibility that KMS in a certain organization contains private information, we create the virtual personalized domain by collecting the several kinds of food blogs in this paper. Most individual's blog is concentrated on only few topics so it may not show the relevance into various trends. Hence, we constructed and used the combination of food blogs as a target domain for this project. The domain was categorized by the names of each continent and country. The combination of several kinds of food blogs is classified by continent and country folder that is defined by the International Cartographic Association (ICA). All food blogs are collected by the Google Search, with the form of such search terms as 'Nation_food_blog'. For example, to find the blogs for the 'japan' folder, we searched by using 'Japanese food blog'. We collected only the blogs that are shown in the first page of Google Search.

In the target domain, there are 4 continent categories (e.g. Asia), 14 area categories (e.g. East Asia) and 26 countries categories. We crawled 22933 documents.

3.4 Relevance Identification

The goal of this paper is to identify the relevance of the collected trending topics to a target object. As discussed before, trending social issues are collected by using Google Trends, Twitter, and Google news. The target domain for this project comprises the combination of various countries' food blogs. In order to calculate the relevance of social issues to the target domain, we applied the Term Frequency Inverse Document Frequency method that is widely used in machine learning area. It also usually used by search engines to rank a document's relevance to given a query.

The way to identify the relevance is described as follows. The set of trending keywords contains a Google Trends keyword and several related keywords extracted from Twitter and Google news. What we want to obtain is the relevance weight of each document to each set of trending keywords. First, the TF was applied. The system neglects the documents that do not contain trending keywords. After that, the system counts the number of terms in a document and totals them. However, if the trending social issue keywords contain common words, such as 'cook', it will emphasize these words. In order to filter out the common terms, Inverse document frequency (IDF) was applied. The IDF weight can be calculated by dividing the total number of documents by the number of documents that contain the trending

keywords. Therefore, the higher TFIDF weight is calculated by both a higher TF (in one document) and a lower DF of the term in the whole target domain [17].

After calculating the relevance weight, we decided how to visualize the relevance of trends to the target domain. Considering different characteristics of each target, the way that visualizes the relevance weight is separated by three different types as follows. The first type of relevance visualization is document-based relevance visualization. This is useful for a user who does not have a large number of documents or a complicated structure. The second type is category/folder-based relevance visualization. Most organizations have a complicated structure so that it is hard to identify all documents for them. Therefore, it might be essential for them to understand the highly-related category. The third combines both document-based and category/folder-based relevance visualization.

3.5 Summary

In this paper, the relevance value (RV) is defined as:

G_n is one of the top 10 search terms from Google Trends.

$$RV_n = \sum_{D=1}^k (TFIDF(TF(G_n, R_{m+i}), T_D))$$

n is a number from 1 to 10. Then, the system searched related documents by using Google Trends keyword (G_n). R_m and R_i represent the related documents from micro-blog, Internet news. To find the highest related keywords, TF was conducted. T_D is a digitalized domain that contains all information of a target object. D is the number of the documents, from 1 to the maximum number, k . To identify the relevance of the set of trending keywords to a target domain, we totalled all documents' TFIDF weight.

4 Evaluation and Discussion

Evaluations of the proposed system were carried out in order to examine the success of the method. With this in mind, we collected data for evaluating the proposed method. First, to extract trending social issues, we crawled Google Trends keywords for a period of 195 days, approximately over 5 months. As described in the introduction, we obtained 17559 unique topics. Secondly, in order to reduce the ambiguity of the social issues, we extracted several related keywords from Twitter and Google News hourly. The target domain is the combination of different countries' food blogs, which were collected from Google search. In the target domain, there are 4 continent categories (e.g. Asia), 14 area categories (e.g. East Asia) and 26 country categories. We crawled 22933 documents. We collected only the blogs, which are shown in the first page of Google Search. Each data set contains one Google Trends keyword, several related keywords, date, and relevance weight. We calculated not

only each target's relevance weight, but also the relevance weight of each document and each category.

The first part will describe the reason why we obtained several related keywords. To do this experiment, we extracted 10 related keywords for each Google Trends Keyword, and calculated their relevance weights. Fig. 1 gathers the relevance weights for the number of related keywords. First of all, if we did not obtain any related keyword, the relevance weights are almost 0, which is the blue line at the bottom of the Fig. 1. In this case, there may be some difficulty to define which social issue is highly related to a target object. However, if we extracted at least one related keyword, you can clearly see the big difference. This justifies why we need to extract the related keywords.

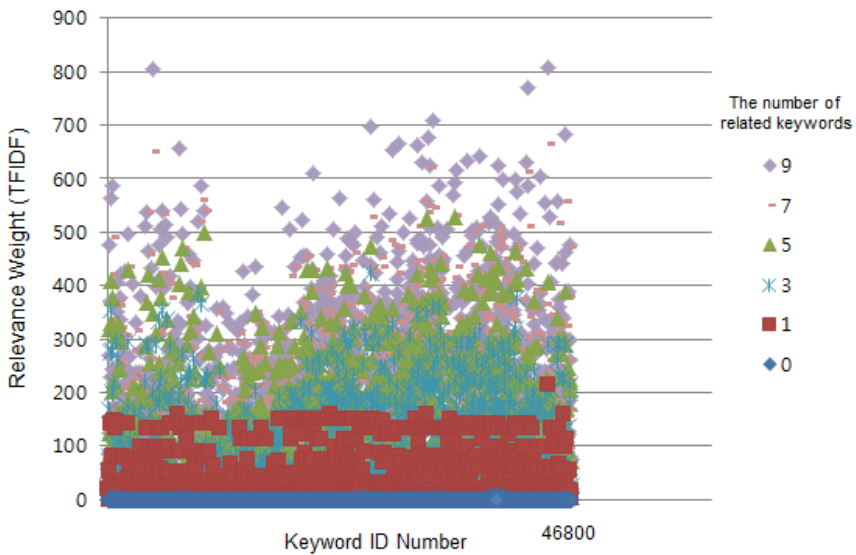


Fig. 1. Relevance weight (using TFIDF) for the number of related keywords

Even though Fig. 1 represents the importance of the related keyword extraction, it is not easy to see how the relevance weights are changed. In this section, we provide Fig. 2, which shows the standard deviation, median and average of relevance weights for the number of related keywords. As can be seen in the graph, the more we obtain the related keywords, the higher will be the standard deviation, median and average weights. According to this result, it will be easier to distinguish among documents if we extract more related keywords are extracted.

Next, we consider the appropriate number of related keywords. In Fig. 2, you can see the gap between each standard deviation is dwindling. This result might show the proper number of related keywords to identify the personalized relevance of social issues to a target object. There are two reasons why we would like to obtain the most appropriate number of related keywords. First, we have to consider the time needed.

We collected related articles from Twitter and Internet news hourly; Tweets are almost 90 and news articles are almost 10. It depends on the number of articles that people uploads in an hour. Extracting over 10 related keywords may not require a long time, but it does require a great amount of time to calculate the personalized relevance of a Google keyword and over 10 related keywords to a target. Secondly, the number of the related articles is limited. If we extract over 10 related keywords, some keywords might not be really related to that social issue. In other words, some keywords may just be very general words that have no relationship with a Google Trends social issue keyword. Therefore, for these two reasons, it is necessary to obtain the suitable number of the related keywords. With this in mind, we present the Fig. 3.

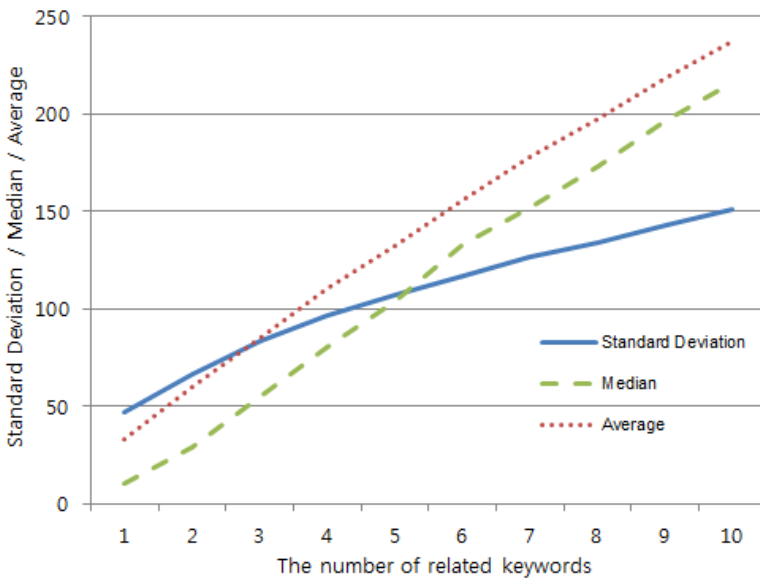


Fig. 2. Standard deviation, Median and Average for TFIDF

Fig. 3 indicates the difference between each standard deviation of the number of the related keywords. The number in x-axis represents the number of related keywords. For example, the '0->1' indicates that the difference in standard deviation between '1 Google keyword + 0 related keyword' and '1 Google keyword + 1 related keyword'.

As can be seen in the first section of the graph, 0 to 1, the difference is the highest in this graph. Then, the rate of those three sections, '1 to 2', '2 to 3', and '3 to 4', follows that of the '0 to 1' section. From the '4 to 5' section, the difference becomes similar or less. Therefore, it seems appropriate to extract 5 related keywords hourly. It is obvious that 5 related keywords are suitable, so we will conduct the user study of the relevance weight accuracy for the number of related keywords in the future.

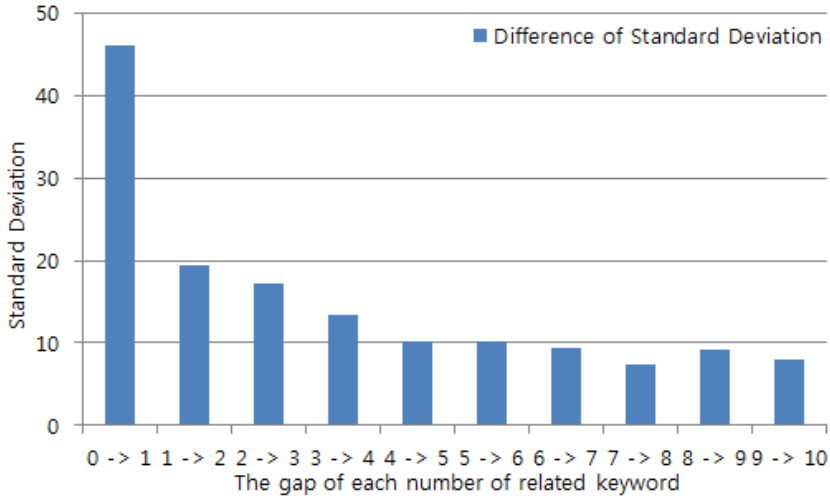


Fig. 3. The difference between each standard deviation of the number of the related keywords

Table 1. Top10 / Bottom10 of relevance weight (TFIDF) based on related keywords

	Keyword	Related Keywords					TFIDF
Top 10	air jordans	sniaggac	yahoosports	frenzy	cause	shoppers	679.4217
	x factor winner	nxvh	winnerthe	myspace	finale	winner	676.2475
	work it	time	things	day	one	relatablequote	637.5518
	friday the 13th	day	today	year	one	jason	636.1167
	truffles	world	food	minutes	chocolate	truffle	634.9386
	truffles	food	world	chocolate	posts	foie	614.8527
	phish	blog	york	post	year	city	594.3573
	friday the 13th	today	year	day	people	lt	588.8346
	restaurant week	time	lunch	food	site	diego	588.4931
	taylor lautner	people	blog	news	year	day	581.1989
Bottom 10	mega upload	megaupload	denovo	anonymouswiki	gomegaupload	gomegaupload	0.1557
	coachella	wid	rolln	snoopdogg	coachizzle	ticket	0.1532
	doj	fbi	riaa	anonymousirc	mpaa	warner	0.0995
	honey badger	beatthesaints	badgerrt	musburger	brent	gt	0.0984
	ifi	owens	iflfl	terrell	iflhttp	nfl	0.0934
	the weeknd echoes of silence	silencert	svdxoqzn	mrmiketubbz	np	chills	0.0802
	nfl playoff schedule	broncos	tebow	timtebow	wcommunities	help	0.0672
	ben roethlisberger	offseason	steelers	deadspin	broncos	tebow	0.0672
	doj	megaupload	fbi	umg	websitesattack	technica	0.0594
	emily maynard	bachelorette	enews	womack	keystone	brad	0.0053

Let’s move on to next step. As mentioned before, the reason for extracting the related keywords is to reduce the ambiguity of a Google Trends social issue keyword and improve the ability to identify the relevance of a social issue to a target.

In this section, we examined whether the related keywords that we extracted are useful to display the exact meaning of social issue and to identify the relevance of a social issue to a target, such as individual or an organization. To do so, we present a qualitative comparison among three types of related keywords; 5 related keywords by

using TF, related searches from Google Trends, and related keywords from Wordnet. To show the various keywords, we choose 20 Google Trends social issue keywords, which are the top10 and bottom10 in relevance weight rankings.

Table 1 covers the related keywords by using TF and their relevance weight. The related keywords seem quite understandable and help users to figure out the exact meaning of each related keywords. The related searches from Google Trends website are also quite understandable and enable the user to obtain the idea of that social issue.

Table 2. Top10 / Bottom10 of relevance weight (TFIDF) based on related searches, WordNet

	Keyword	Related Searches provided by Google					TFIDF	WordNet			TFIDF
Top 10	air jordans	air jordan 11 retro concords	air jordan	jordan shoes	new air jordans	air jordan 11 concord	0	N/A			0
	x factor winner	x factor winner 2011	x factor finale	melanie amaro x factor	x factor 2011	xfactor	0	N/A			0
	work it	bosom buddies	last man standing	new tv shows 2012	revenge	bachelor	1.1925	N/A			0
	friday the 13th	friday the 13	fredag den 13	freddy krueger	friday 13th	friday the 13th quotes	0	N/A			0
	truffles	truffle	fedora	perugia	stem cell	60 minutes	21.4043	fungus	vegetable	candy	26.1856
	truffles	fedora	perugia	stem cell	60 minutes	empathy	2.5708	fungus	vegetable	candy	26.1856
	phish	francesca woodman	sugarland	uk basketball	nikon	girl with the dragon tattoo	2.8419	N/A			0
	friday the 13th	friday 13th	friday the 13	viernes 13	13th friday	fredag den 13	0	N/A			0
	restaurant week	mutual funds	nyc.gov	nyc restaurant week	flowers of war	mik quotes	0.0000	N/A			0
taylor lautner	bolo	hallo pizza	hayley williams	la sirenetta	lily collins	0.4350	N/A			0	
Bottom 10	mega upload	megaupload	anonymous	department of justice	fbi	icefilms	0.0594	N/A			0
	coachella	coachella 2012	coachella 2012 lineup	coachella line up	coachella lineup	tomorrowland	0.1404	N/A			0.1404
	doj	department of justice	universal music	justice.gov	universal	anonymous	0	executive department			0
	honey badger	alabama crimson tide	alabama football schedule	alabama national championships	honeybadger	jarrett lee	0	musteline mammal			0
	iftl	ifc	indoor football league	itl	terrell owens	gary carter	0.3828	N/A			0
	the weeknd echoes of silence	the weeknd	clams casino	michael jackson	alan hansen	xfactor	0	N/A			0
	nfl playoff schedule	h and r block	raven	bcs	dr mercola	peyton manning	0.5864	N/A			0
	ben roethlisberger	ben roethlisberger	eli manning	peyton manning	andy dalton	big ben	0	N/A			0
	doj	universal music	department of justice	mpaa	universal	justice.gov	0	executive department			0
	emily maynard	brad womack	girl scout cookies	the bachelor	hysteria	joe magrane	0.5128	N/A			0

However, Wordnet is able to extract only very few related keywords. This is because most Google Trends keywords are related to a celebrity's name or an event. Wordnet provides semantically related words/terms so that the Wordnet system cannot find many related keywords. As you can see in the Table 2, the relevance weights of those two groups, the related searches from Google and the related terms from Wordnet, are almost 0. However, with the related keywords by TF, they display the relevance very clearly. Compared to the other two groups, it achieves much better and more recognizable results.

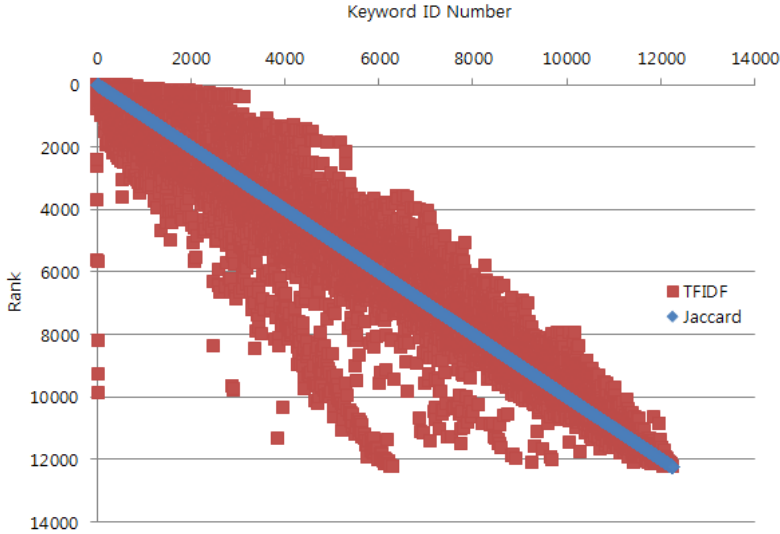


Fig. 4. Trend of TFIDF and Jaccard

In this paper, we used TFIDF as a primary approach to calculate the relevance weight of social issues to a target. TFIDF is a widespread technique to calculate relevance weights, and it is utilised for ranking relevance weights in most search engines. However, TFIDF has never been used in calculating personalised relevance of social issue before. Hence, we began an experiment to justify the efficiency of TFIDF by comparing it with another relevance calculation approach, Jaccard.

In order to show a difference of trend between TFIDF and Jaccard weight, we ranked each keyword on the basis of its applied relevance value. After that, we ranked in ascending order these trending social issue keywords that are applied TFIDF method, and compared them with the rank of same keywords that are applied in Jaccard. As a result, we can observe some similar trends in both of two methods, TFIDF and Jaccard. From this experiment, it is proved that we can obtain the similar relevance value regardless of relevance weight approach. For future work, it might be good to propose new relevance weighting approach that will suitable to this project.

5 Conclusion

As mentioned before, we present our research on developing a system to identify the personalized relevance of trends to target objects, such as individuals or organizations. The evaluation results proved that we have achieved these three aims: (1) trending social issue collection, (2) target domain identification, and (3) demonstration of the relevance of the social issue to a target domain. First, we collected social issues from Google Trends, Twitter and Internet news. The target domain for this paper is the combination of different countries' food blogs. We constructed the virtual target

domain, which is well-structured and categorized, so that the system can identify the relevance weight of each document and category. Finally, we applied the TFIDF method to obtain the personalized relevance of social issues to a target, such as an individual or an organization.

We conducted several types of experiments. Firstly, we proved that it is necessary to extract the related keywords and showed the appropriate number of related keyword in this paper. We analysed and compared the extracted related keywords by using TF with other related keywords groups. The advantage of our related keywords is proved. We also analysed the comparison between TFIDF and Jaccard to prove the efficiency of TFIDF. As mentioned in evaluation part, for the future work, we will conduct further analysis and evaluation, including user study.

Acknowledgements. This paper was supported by Asian Office of Aerospace Research and Development (AOARD), Japan. This paper was supported by Korea Association of Industry, Academy and Research Institute with the project called “Business support for academic-industrial common technology development”. We are grateful to Byeong Ho Kang and Hee-Geun Yoon for helpful discussions.

References

1. Aly, A.A.: Using a Query Expansion Technique To Improve Document Retrieval. *International Journal "Information Technologies and Knowledge"* 2, 343–348 (2008)
2. Boyd, D.M., Ellison, N.B.: Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication* 13(1), 210–230 (2007)
3. Brutlag, J.D., Meek, C.: Challenges of the Email Domain for Text Classification. Microsoft Research, One Microsoft Way, Redmond, WA, USA (2000)
4. Chum, O., Mikulik, A., Perdoch, M., Matas, J.: Total recall II: Query expansion revisited. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 889–896 (2011)
5. Cohen, W.W., Ravikumar, P.D., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: *IIWeb*, pp. 73–78 (2003)
6. Fitzpatrick, K.: *The Pleasure of the Blog: The Early Novel, the Serial, and the Narrative Archive*. Pomona Faculty Publications and Research (2007)
7. Joinson, A.N.: Looking at, looking up or keeping up with people?: motives and use of facebook. In: *CHI 2008 Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, pp. 1027–1036. ACM, New York (2008)
8. Juang, Y.S., Lin, S.S., Kao, H.P.: A knowledge management system for series-parallel availability optimization and design. *Expert Systems with Applications*, 181–193 (2008)
9. Kolbitsch, J., Maurer, H.: The Transformation of the Web: How Emerging Communities Shape the Information we Consume. *Journal of Universal Computer Science* 12(2), 187–213 (2006)
10. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: *WWW 2010 Proceedings of the 19th International Conference on World Wide Web*, pp. 591–600. ACM, New York (2010)
11. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: *IUI 2010 Proceedings of the 15th International Conference on Intelligent User Interfaces*, pp. 31–40. ACM, New York (2010)

12. Paolillo, J.C.: Structure and Network in the YouTube Core. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), p. 156 (2008)
13. Rech, J.: Discovering trends in software engineering with google trend. In: ACM SIGSOFT Software Engineering Notes, vol. 32(2), ACM, New York (2007)
14. Ristad, E.S., Yianilos, P.N.: Learning string-edit distance. *IEEE Trans. Pattern Annual Mach. Intell.*, 522–532 (1998)
15. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: WWW 2010 Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM, New York (2010)
16. Tirado, J.M., Higuero, D., Isaila, F., Carretero, J.: Analyzing the impact of events in an online music community. In: SNS 2011 Proceedings of the 4th Work-shop on Social Network Systems, vol. (6). ACM, New York (2011)
17. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting TF-IDF term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)* 26(3) (2008)

Identifying Important Factors for Future Contribution of Wikipedia Editors

Yutaka Yoshida and Hayato Ohwada

Department of Industrial Administration, Faculty of Science and Technology,
Tokyo University of Science,
2641 Yamazaki Noda-shi Chiba-ken 278-0022 Japan
y.yoshida01115@gmail.com, ohwada@ia.noda.tus.ac.jp

Abstract. This paper presents research relevant to predicting future editing by Wikipedia editors. We demonstrate the importance of each characteristic and attempt to clarify the characteristics that affect prediction. Clarifying this can help the Wikimedia Foundation (WMF) understand the editor's actions. This research adopted the increase in prediction errors as the means of evaluating the importance of a characteristic and thus computed the importance of each characteristic. We used random forest (RF) regression for calculating the importance. Characteristic evaluation in our experiment revealed that the past number of edits and the editing period increased predictive accuracy. Furthermore, information regarding earlier edit actions clearly contains factors that determine future edit actions.

Keywords: Wikipedia Participation Challenge, feature importance, random forest, prediction.

1 Introduction

This research problem was raised in a data-mining contest conducted by ICDM in 2011. Wikipedia is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation (WMF). Since 2001, Wikipedia has become the largest and most popular general reference knowledge source on the Internet.

However, Wikipedia growth has recently slowed significantly [1]. In particular, WMF has reported that between 2005 and 2007, *newbies* started having real trouble successfully joining the Wikimedia community. Before 2005, in the English Wikipedia, nearly 40% of new editors were still active a year after their first edit. After 2007, only 12 to 15% of new editors were still active in the same period. Post-2007, many people were still trying to become Wikipedia editors. What had changed, though, was that they were increasingly failing to integrate into the Wikipedia community, and failing was increasingly quickly. The Wikimedia community had become too hard to penetrate. Therefore, it is important to understand quantitatively what factors determine an editor's future (i.e., why they continue editing, change the pace of editing, or stop editing), in order to ensure that the Wikipedia community can continue to grow in size and diversity.

The Wikipedia Participation Challenge [2], sponsored by WMF and hosted by Kaggle, requested contestants to build a predictive model that would accurately predict the number of edits a Wikipedia editor would make in the next five months based on his edit history. Such a predictive model might help WMF determine how people can be encouraged to become, and remain, active contributors to Wikipedia. The winning solution to the challenge was announced at ICDM.

We now explain the problem setting of the contest. The problem was to construct a model that would predict how many edits would be made during the five months from 2010-09-01 to 2011-02-01 by using the past editing record, covering the period from 2001-01-01 to 2010-08-31. A prediction was made for each editor. Each predictive model's accuracy was measured by its Root Mean Squared Logarithmic Error (RMSLE). Ninety-six models were submitted in the contest. Wikipedia accepted some of the high-precision solutions. However, the best method was not determined. The importance of the used feature space was not calculated, even though that knowledge is important for understanding what factors determine an editor's future. Another project similar to Wikipedia will probably arise at some point and will face the same problem. Whether this leads to that understanding is important for future problem design. The purpose of this research is to clarify the importance of the characteristics found to affect prediction.

Section 2 describes related works. Section 3 identifies the selected characteristics. Section 4 discusses how to calculate characteristic importance. Section 5 describes our experiment and presents its results. Section 6 presents our conclusions.

2 Related Work

The global slowdown of Wikipedia's growth rate was studied in [3]. This source compared the growth problem in the present condition of Wikipedia with a society that has matured. There are also increasingly complex patterns of conflict and dominance, which may be the consequence of the increasingly limited opportunities to make novel contributions.

Ref. [4] used additional data uniquely extracted from Wikipedia in addition to the data provided by WMF and made highly accurate predictions. The central idea of this prediction method is to use many characteristics in the learning algorithm. The strategy was to construct a large number of characteristics to feed into the learning algorithm to ensure that maximum information would be available to the learning process. In addition, 206 of the characteristics involve elements that characterize an editor's work, such as the percent of articles edited by the user that were unique, the percent of edits that were new or original, the percent of edits that had comments, whether or not the user was blocked before the end of the observation period, and so on. The importance of each characteristic was computed by the random forest method, and the result suggested that the time scale and the number of edits were the most important for

predicting the future number of edits. However, this result is contrary to our expectation that characteristics involving an editor's own characteristics would have low importance. The result does not suggest that the amount of information can serve as a characteristic for prediction. Its predictive accuracy surpassed 43.2% of the WMF baseline predictive model. Ref. [5] made predictions using only the number of edits and the time. Its predictive accuracy surpassed the 41.7% of the WMF baseline predictive model. The number of edits and the time are thus highly important for predictive accuracy. In [6], predictions were made using several characteristics. Some characteristics are not included in the above-mentioned approaches. The predictive accuracy was 0.881. This improved the 40.5% standard of WMF. New knowledge was acquired, although the accuracy is low as compared with related research since the related research differs in the characteristics used.

In addition to Wikipedia, online user's behavior has been explored and exploited in social tagging [7,8], blogging [9], and on Twitter [10].

3 Feature Space

To construct a regression model, an explaining variable is chosen with reference to accumulated knowledge based on proof drawn from past research. Our research seeks to discover characteristics involved in prediction. Therefore, we chose a characteristic vector that may express an editor's characteristics in accordance with certain standards. Time information is first explained based on a standard that only partly exists.

Fig.1 indicates that the number of editors doubled in 2010. This suggests the importance of recent data over old data. In other words, data prior to 2010 contained only half the information of editors.

Next, Fig.2 expresses the relation between edit participation time and the number of edits. The edit participation date refers to the date of the first edit after January 1, 2001. A small number of edits were done if the edit participation date is close to the current date. The correlation of the edit participation date and the number of edits is low except for this. As mentioned above, the importance of past information is low, although information about the latest edit is important. The time element was also considered. The standards for characteristic selection are given below.

- (1) **Characteristic specifying the number of edits.** The number of previous edits is important for predicting the number of future edits.
- (2) **Editor characteristics.** The editor's own characteristics may influence the editor's future actions. It is necessary to investigate this relation.
- (3) **Characteristic describing the article.** The characteristic specifying the article that the editor edited is also used. It is necessary to investigate its hidden causal relationship with the editor's actions.
- (4) **Characteristic that considers time information.** Based on Figs.1 and 2, the characteristic considering time information is important and is required in order to verify the result.

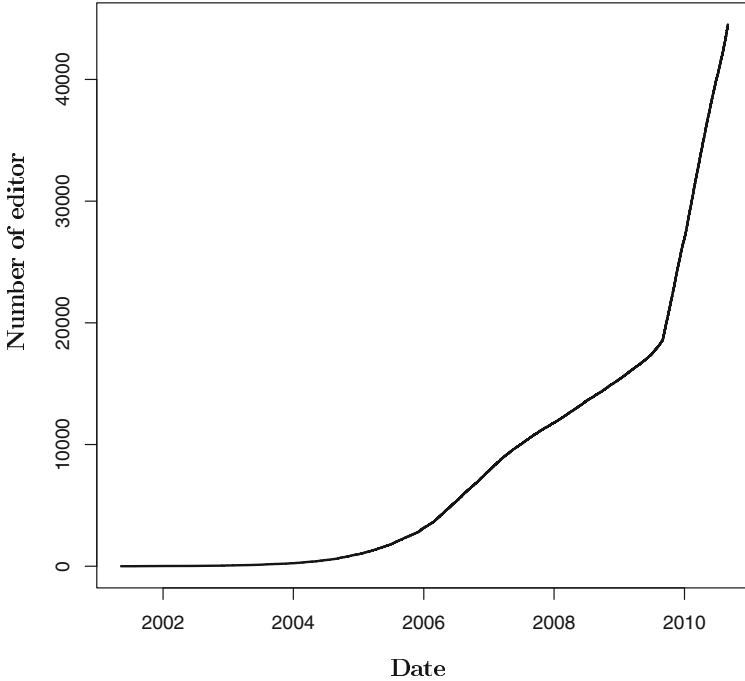


Fig. 1. Relation between time and the number of editors

The feature space chosen from the above standard is presented in Table 1. Section 3.1 discusses the time-division method used to increase the prediction accuracy. The characteristic chosen from each standard is explained in Section 3.2 to Section 3.5.

3.1 Time-Division Method

We searched for a time-division method whose prediction accuracy would be higher. Because time-series data are accumulated in each fixed period, the method of time division influences the prediction accuracy. In a model with high prediction accuracy, this is a very important characteristic. We compared the prediction accuracy of the period acquired from a division-into-equal-parts rate method and from the following formulas.

Step1. Let T_p be the end time of the period before the period that should be predicted. Let T_0 be the starting time.

Step2. $T_{(p-1)} = T_p - d$, where d is a constant representing the length of a period that should be predicted.

Step3. $T_{(p-n)} = T_{(p-1)} - 2^n$, where $n = -4, -3, \dots, 12$.

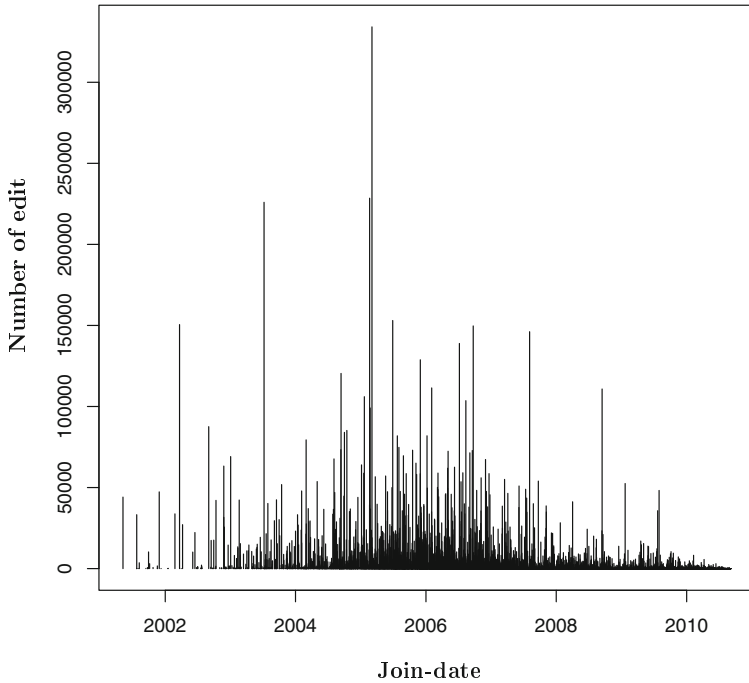


Fig. 2. Relation between edit participation date and the number of edits

3.2 Characteristic Indicating the Number of Edits

Total Number of Edits(1-2). Characteristic 1 is the sum of the number of edits obtained from the number of edits time-series vector (1), and characteristic 2 is the sum of the number of articles (2). An editor with a large number of previous edits is likely to produce a large number of future edits. Moreover, an editor who is editing many articles has a deeper relation with Wikipedia than those who edit a single article. We converted these values into natural logarithms because a very wide range was seen in the data.

Change in the Number of Edits(18-21). The change in the number of edits refers to the difference in the number of edits totaled for each period. The number of edits, whether increasing or decreasing, and the future number of edits should be closely related, so this characteristic was used. The rate of increase of the number of edits is the value which divide the number of periods in which the number of edits are increasing by the number of periods (18). The rate of decrease of the number of edits is the value which divide the number of periods in which the number of edits are decreasing by the number of periods (19). Moreover, the maximum increase in edits (20) and the maximum decrease in edits (21) are also characteristics.

Table 1. Outline of Feature Space

No.	Feature@
(1-2)	Natural logarithm of total number of revisions and articles.
(3-8)	Rate of edits to name spaces 0-5.
(9-10)	Average number of edited characters and rate of additional characters.
(11)	Sum of influences.
(12-13)	Number of edits reverted by oneself and others.
(14-15)	Number of articles in procedural and evaluative categories.
(16-17)	Number of redirect articles and articles with related content.
(18-19)	Increase rate and decrease rate of number of edits.
(20-21)	Maximum increase number and decrease number of number of edits.
(22)	Average number of article title characters.
(23-25)	Average, maximum and minimum number of comment characters.
(26-28)	Number of edit periods, active edit periods and length of time.
(29-32)	Sum of characters, influences, increases and comments edited during the last period.
(33-39)	Weighted sum of characters, influences, increases, comments, revisions and articles.
(40-55)	Natural logarithm of number of edits divided by time periods (y_t).

Number of Edits for Each Period(40-55). The number of edits is the total of the number of edits for each period. We converted these into natural logarithms because a very wide range was seen in the data. Let y_1 be the period from the start of a Wikipedia project. Let y_{17} be the most recent period. Since there were very few samples of y_1 , the characteristic after y_2 was used.

3.3 Editor Characteristics

Rate of Edits to Name Spaces(3-8). The edit rate R_i of a given editor's name space i is denoted by the following formulas.

$$R_i = \frac{C_i}{\sum_{i=0}^5 C_i} \quad (1)$$

$$C_i = \sum_{x=1}^n |\Delta c_i(x)| \quad (2)$$

Here, $\Delta c_i(x)$ is the number of characters in the x -th edit of name space i . The edit rate of name spaces 1 (3) through 5 (8) is used. As an editor's characteristic, those who edit an article in the user name space or the Wikipedia name space in addition to the standard name space can say that it is important. For example, those who edit name spaces 4 and 5, which are Wikipedia name spaces, are considered to manage other editors' work. The probability that they belong to the Wikipedia community is high.

The Number of Edit Characters(9,10). The average of the number of edit characters (9) is calculated without taking the absolute value. The character addition rate (10) is the rate at which the number of edit characters takes a positive value. Here, the addition of a character and the rate of deletion were

expressed in two characteristics. For example, editors can be divided into those with many character additions and those with few character additions. The average of the number of edit characters for a person with the same rates for added and erased characters is set to 0. For other cases, a higher rate of the number of added characters results in a positive value, and a higher rate of the number of erased characters results in a negative value.

Degree of Influence(11). The characteristic denoted by the following formulas is defined as the degree of influence on an article.

$$i = \begin{cases} \frac{\Delta c}{c} & (\Delta c \geq 0) \\ \frac{c - |\Delta c|}{c - \Delta c} & (\Delta c < 0) \end{cases} \quad (3)$$

Here, Δc is the number of edit characters, and c is the number of characters in the article after editing. This takes the maximum value 1 when $\Delta c \geq 0$ and $|\Delta c| = c$, i.e. when a new article is created. It takes the minimum value -1 when $\Delta c < 0$ and $|\Delta c| = c$ i.e. when an article is deleted. The degree of influence on an editor's Wikipedia article is measured by totaling this value for all of the periods. This characteristic is used in order to investigate the relation between the degree of influence and the number of edits.

Reversions(12,13). Ref. [3] indicated that the strife within the exclusive Wikipedia community may lead to a reduction in editing. The number of times revisions were made in person (12) and the number of times revisions were made for others (13) were obtained from the revision data. For example, an editor with a large number of revisions made in person is considered to perform such edit activities while highly motivated to correct the article content. An editor with a large number of revisions for others suggests the probability that the edits will be controlled is high.

Comments(23-25). Information about the number of characters in comments for each edit is used. The number of characters in comments could express the volition of the editor's activity. We use the average number of comment characters (23), the maximum number (24), and the minimum number (25). However, a comment given automatically is not classified at this time since its semantic importance is not high.

Periods(26-28). Information about the period and time required to perform edits is used. The number of periods (26) expresses the time from the first edit to the last edit. Time is measured as a difference in UNIX time (28). There is a period when an edit is not performed even once. A period which does not include such a period was defined as the active period (27). These characteristics were used in order to investigate the relation between future actions and the periods.

3.4 The Article(14-17,22)

Category, redirection, related article, and title information from an edited article is used. Information was classified into procedural and evaluated categories based on the meaning. The procedural category (14) refers to *Deletion*, *Mediation*, and *Arbitration*. Any article in which there is a discussion on deletion, mediation, or arbitration belongs in this category; good articles and featured articles belong in the evaluated category (15). Any article evaluated by the Wikipedia community belongs in this category. The article contents could be classified into these two types and totaled separately since the semantic difference is great. For example, a person who edits an article in an evaluated category may be considered to have a deep relation with the Wikipedia community.

For redirection (16) and related articles (17), the number of edits to the article was totaled. The number of title characters was also taken as a simple average (22). Although there is no semantic importance in these, the characteristic was introduced in order to explore any potential hidden relationships.

3.5 Consideration of Time Information

The number of edit characters, the degree of influence, the increase in the number of edits, the number of comment characters, the number of edits, and the number of articles can be totaled for each period. The number of edit characters (29), the degree of influence (30), the increase in edits (31), and the number of comment characters (32) of the final period are used as characteristics. The weighted sum of these (33-37), the number of edits (38), and the number of articles (39) are also used as characteristics. The weighted sum is the sum of the amount of information expressed by the following formula.

$$S_w = \sum_{t=1}^{p-1} \frac{100}{p-t} I_t \quad (4)$$

Here, I_t is the above-mentioned amount of information during period t . The weight takes a maximum value of 100 in the period before the prediction period. The result is the sum of what added the time meaning to the amount of information.

4 Calculating Importance

We now describe the method for calculating the importance of a characteristic. When evaluating the importance of a characteristic over a predictive model, it is useful to determine to what extent prediction errors increase if the characteristic is not used. This research defines the increase in this prediction error as the characteristic importance. If the prediction error increases when this characteristic is not used, the characteristic has high importance; if the inverse were true, the characteristic is not important.

The random forest (RF) regression method is used to calculate the importance. RF regression, as proposed in [11], is a form of ensemble learning employing many decision trees as a weak study machine and determines a final predicted value by averaging each decision tree. First, the sample set used for learning each decision tree is generated with a bootstrap method. The m characteristics of the M -characteristic set are then chosen at random, and a decision tree is built. The prediction error totaled from many such decision trees can be made into an importance measure.

We use two importance measures, MSE and purity, for calculating characteristic importance. The MSE importance of characteristic x_j is the difference in the mean squared error (MSE) between using and not using the characteristic. The Z score performs scaling using the standard deviation of this value. In [12], the result of not performing scaling suggests that the statistical value is high. This research demonstrates the MSE importance without performing scaling. If the MSE importance of characteristic x_j is set to $VI(x_j)$, the following formulas apply.

$$x_{i,\pi_j} = (x_{i,1}, \dots, x_{i,j-1}, x_{\pi_j(i),j}, x_{i,j+1}, \dots, x_{i,p}) \tag{5}$$

$$VI^{(k)}(x_j) = \frac{\sum_{i \in \beta} (y_i - f_k(x_i))^2}{|\beta|} - \frac{\sum_{i \in \beta} (y_i - f_k(x_{i,\pi_j}))^2}{|\beta|} \tag{6}$$

$$VI(x_j) = \frac{\sum_{k \in T} VI^{(k)}(x_j)}{|T|} \tag{7}$$

Here, x_{i,π_j} is a vector in which the value $x_{i,j}$ of variable j of sample i is replaced with $x_{\pi_j(i),j}$, which is the value of a randomly selected variable $\pi_j(i)$. T is a decision tree set. The purity importance is the average amount by which the decision tree error improved due to characteristic x_i . If the purity importance of characteristic x_j is $P(x_j)$, the following formula applies.

$$P(x_j) = \frac{1}{|T|} \sum_{k \in T} \sum_{d \in D_k} (S_d - (S_{d_R}(x_j) + S_{d_L}(x_j))) \tag{8}$$

Here, D_k is a node set of a decision tree, S_d is the error square sum before the decision tree branches, $S_{d_R}(x_j)$ is the error square sum on the right-hand side of the node branched according to characteristic x_j , and $S_{d_L}(x_j)$ is the error square sum on the left-hand side of the node branched according to characteristic x_j .

This research determines the relative MSE importance of each characteristic as the MSE importance divided by the maximum MSE value of the characteristic. This importance is also computed in [4]. If the relative MSE importance of characteristic x_j is $rVI(x_j)$, the following formula applies.

$$rVI(x_j) = \frac{VI(x_j)}{\max(VI(x))} \tag{9}$$

5 Experiment

R, an open-source statistical-analysis software package, was used for analysis and regressive prediction. We used Java for calculating the feature space and *random forest* [3], a component of R, for the RF method.

5.1 Data Set

The data set consisted of the full history of the editing activities of *semi-randomly sampled* active editors of the English Wikipedia (the first six namespaces only) active from January 1, 2001, to September 1, 2010. This data was used in the ICDM contest. During this period, 4,012,171 editors made a total of 272,213,427 edits in the English Wikipedia. Using the following sampling strategy, the data file is seen to consist of 44,514 sampled editors who contributed a total of 22,126,031 edits.

The WMF did not use a random sample because between 80% and 95% of those editors would have stopped editing. For such a sample, a constant prediction model would score very well (low RMSLE) but would not be helpful to the WMF. They therefore sampled editors who made at least one edit between September 2009 and September 2010, and extracted the full editing history for the sampled editors. For each edit, the available information includes its user ID, article ID, revision ID, namespace, and timestamp.

5.2 Result

We computed the random-forest importance as in the model of [4], in order to discuss the used characteristic. The MSE and purity of all of the characteristic quantities are seen in Fig 3. Furthermore, the top 20 most important characteristics based on relative MSE importance are listed in Table 2.

The strategy of [6] is the same as that of the model of [4], which is to construct many characteristics to confirm much of the information that would be available to the learning process and to find hidden relationships. In contrast, [5] achieves high prediction accuracy using data representing the number of edits, the number of articles, and the editing period. These characteristics also are included in [6], and the number of training samples made little difference. As mentioned above, a compensation algorithm reflecting an editor's latest prediction trend is considered to raise the prediction accuracy.

Next, we discuss the characteristics selected. In [4], many characteristics expressing an editor's attributes besides the number of edits are used. The results indicated that the time and the number of edits played the most important roles in predicting future editing; the attributes of the edits themselves, e.g. namespace, comments, and block, played a lesser role. Our study indicates the same result. Some characteristics, such as an article attribute indicating that a hidden relationship exists, had little importance. We had anticipated that reducing the number of edits due to reversion as seen in [3] would influence prediction.

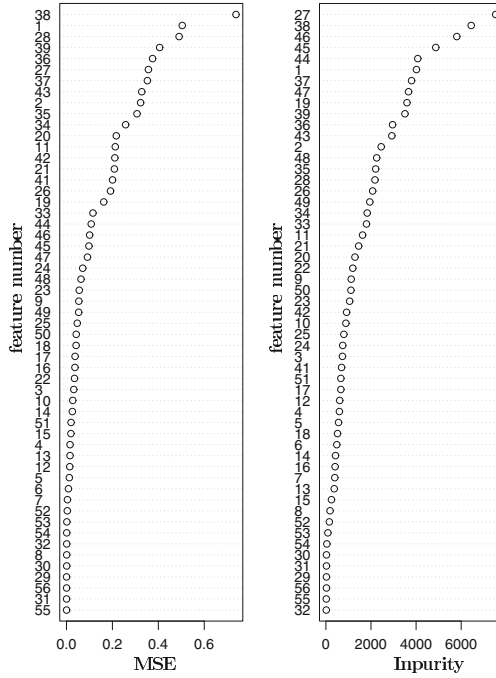


Fig. 3. MSE and purity importance

However, reversion also had little importance. In Fig. 3, the importance of reversion is less than half that of the other characteristics. The number of edits, number of articles, and editing time were highly important, as also indicated by related work. Furthermore, uniquely defined characteristics, such as weighted information and influence on the article, were highly important. In Table 2, the weighted sum of information on the number of edits was 1.464 times more important than the total number of edits. Furthermore, the characteristic ranked 20th in importance had only 14% of the importance of the top-ranked characteristic.

In Fig. 3, the top 20 characteristics contain only the number of edits, the edit period, the weighted sum of information, and the influence. The weighted sum of information was highly important. Based on this, a characteristic that considers the time information is useful in prediction. Characteristics concerning the article, such as its category, had little importance. These characteristics were hardly related to the number of edits. Characteristics indicating the number of edits were ranked $y_4, y_3, y_2,$ and y_5 sequentially from the top within the top 20. y_4 expresses the period from 767 days ago to 255 days ago. y_3 expresses the period from 1791 days ago to 767 days ago. y_2 expresses the period from 3839 days ago to 1791 days ago. y_5 expresses the period from 255 days ago to 127 days ago. These periods include rather old information, and recent information had little importance. This result differs from the original expectation. A factor related to future editing was revealed in these periods. To summarize the above,

Table 2. Relative MSE importance

No.	Feature	Importance
38	Weighted sum of information(the number of edits)	1.000
1	Total number of revisions	0.683
28	Time from the edit start to the latest edit(UNIX)	0.664
39	Weighted sum of information(the number of articles)	0.550
36	Weighted sum of information(the number of decreasing edits)	0.508
27	Number of active edit periods	0.482
37	Weighted sum of information(comment length)	0.483
43	The number of edits(y_4)	0.444
2	Total number of articles	0.437
35	Weighted sum of information(the number of increasing edits)	0.416
34	Weighted sum of information(influence)	0.349
20	Maximum increase number of edits	0.294
11	Sum of influences	0.288
42	The number of edits(y_3)	0.285
21	Maximum decrease number of edits	0.282
41	The number of edits(y_2)	0.271
26	Number of edit periods	0.260
19	Rate of decrease number of edits	0.220
33	Weighted sum of information(the number of characters)	0.156
44	The number of edits(y_5)	0.146

the characteristics that should be used to predict the number of future edits are the past number of edits, the number of editorials, and the editing time from the past to the present. Other characteristics (e.g., the comment head) that consider the time values also raise the prediction precision. We also found some factors clearly related to previous editing.

6 Conclusion

This research compared characteristics that could be used in a model for predicting a person's future edits based on past editing information and clarified the most important characteristics.

We used the number of edits, the editor's own data, data concerning the article, and time information as the feature space. The increase in the prediction error when a characteristic was not used was taken to represent the importance of that characteristic, and the importance of each characteristic was computed.

Characteristics that consider the past number of edits, the number of articles, the information acquired during the edit period from the past to the present, and time were found to raise the prediction accuracy as a result of characteristic evaluation in our experiment. These characteristics are thus more important than other characteristics. It also became quite clear that previous edits contain factors determining future edits.

References

1. Liere, D., Fung, H. (eds.): Trends Study, http://strategy.wikimedia.org/wiki/March_2011_Update (update March 11, 2011, last accessed at December 29, 2011)
2. Wikipedia's Participation Challenge, <http://www.kaggle.com/c/wikichallenge> (last accessed at January 13, 2012)
3. Suh, B., Convertino, G., Chi, E.H., Pirolli, P.: The singularity is not near: Slowing growth of Wikipedia. In: Proceedings of the 2009 International Symposium on Wikis (WikiSym), Orlando, FL, USA (2009)
4. Herring, K.T.: Wikipedia Participation Challenge Solution, http://meta.wikimedia.org/wiki/Research:Wiki_Participation_Challenge_Ernest_Shackleton (last accessed at December 29, 2011)
5. Zhang, D.: Wikipedia Edit Number Prediction based on Temporal Dynamics Only, <http://arxiv.org/abs/1110.5051> (last accessed at December 29, 2011)
6. Yoshida, Y., Ohwada, H.: Wikipedia Edit Number Prediction from the Past Edit Record Based on Auto-Supervised Learning. In: Proceedings of the 2012 International Conference on Systems and Informatics (2012)
7. Halpin, H., Robu, V., Shepherd, H.: The complex dynamics of collaborative tagging. In: Proceedings of the 16th International Conference on World Wide Web (WWW), Banff, Alberta, Canada, pp. 211–220 (2007)
8. Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J., Tseng, B.L.: Detecting splogs via temporal dynamics using self-similarity analysis. *ACM Transactions on the Web (TWEB)* 2(1), 1–35 (2008)
9. Zhang, D., Mao, R., Li, W.: The recurrence dynamics of social tagging. In: Proceedings of the 18th International Conference on World Wide Web (WWW), Madrid, Spain, pp. 1205–1206 (2009)
10. Abel, F., Gao, Q., Houben, G.-J., Tao, K.: Analyzing temporal dynamics in twitter profiles for personalized rec- ommendations in the social web. In: Proceedings of the 3rd International Conference on Web Science (WebSci), Koblenz, Germany (2011)
11. Breiman, L.: Random Forest. *Machine Learning* 45(1), 5–32 (2001)
12. Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. *BMC Bioinformatics* 9, 307 (2008)
13. Breiman, L., Cutler, A., Liaw, A., Wiener, M.: Breiman and Cutler's Random Forests for Classification and Regression R package version 4.6-2 (2011)

Network Analysis of Three Twitter Functions: Favorite, Follow and Mention

Shoko Kato¹, Akihiro Koide¹, Takayasu Fushimi¹,
Kazumi Saito¹, and Hiroshi Motoda²

¹ School of Management and Information, University of Shizuoka,
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
{b09032,j11103,j11507,k-saito}@u-shizuoka-ken.ac.jp

² Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We analyzed three functions of Twitter (Favorite, Follow and Mention) from network structural point of view. These three functions are characterized by difference and similarity in various measures defined in directed graphs. Favorite function can be viewed by three different graph representations: a simple graph, a multigraph and a bipartite graph, Follow function by one graph representation: a simple graph, and Mention function by two graph representations: a simple graph and a multigraph. We created these graphs from three real world twitter data and found salient features characterizing these functions. Major findings are a very large connected component for Favorite and Follow functions, scale-free property in degree distribution and predominant mutual links in certain network motifs for all three functions, freaks in Gini coefficient and two clusters of popular users for Favorites function, and a structure difference in high degree nodes between Favorite and Mention functions characterizing that Favorite operation is much easier than Mention operation. These finding will be useful in building a preference model of Twitter users.

1 Introduction

Grasping and controlling preference, tendency, or trend of the consuming public is one of the important factors to achieve economic success. Accordingly, it is vital to collect relevant data, analyze them and model user preference. However, quantifying preference is very difficult to achieve and finding useful measures from the network structure is crucial. The final goal of this work is to find such measures, characterize their relations and build a reliable user preference model based on these measures from the available data. As the very first step, we focus on Twitter data and analyze the user behavior of three functions (Favorite, Follow and Mention) of Twitter¹ from the network structural point of view, i.e., by using various measures that have been known to be useful in the graph

¹ <http://twitter.com/>

theory and identifying characteristic features (difference and similarity) of these measures for these functions.

User behavior of these three functions are represented by different directed graphs. Favorite function can be viewed by three different graph representations: a simple graph, i.e., single edge from a FAVORER to a FAVOREE, a multigraph, i.e., multiple edges from a FAVORER to a FAVOREE, and a bipartite graph, i.e., single edge from a FAVORER to a FAVOREE treating a user with both a FAVORER and a FAVOREE as two separate nodes. Likewise, Follow function can be viewed by one graph representation: a simple graph, i.e., single edge from a FOLLOWER and a FOLLOWEE, and Mention function can be viewed by two different graphs: a simple graph, i.e. single edge from a MENTIONER (sender) to a MENTIONEE (receiver) and a multigraph, i.e. multiple edges from a MENTIONER to a MENTIONEE. We have created these networks from three different Twitter logs (called "Favorites network", "Followers network", and "Mentions network") and used several different measures, e.g. in-degree, out-degree, multiplicity, Gini coefficient, etc. Extensive experiments were performed and several salient features were found. Major findings are that 1) Favorites and Followers networks have a very large connected component but Mentions network is not, 2) all the three networks (both simple and multiple) have the scale-free property in degree distribution, 3) all three networks (simple) have predominant three-node motifs having mutual links, 4) Favorites network have freaks in Gini coefficient (one of the measures), 5) Favorites network have two clusters of popular users, and 6) Favorites and Mentions networks differ in structure for high degree nodes reflecting that Favorite operation is much easier than Mentions operation. In this paper, we propose to analyze multigraphs by using two new measures, i.e., correlation between degree and average multiplicity, and correlation between degree and Gini coefficient. In our experiments, we show that these measures contribute to clarify a structure difference between Favorites and Mentions networks.

Twitter, a microblogging service, has attracted a great deal of attention and various properties have already been obtained [3] [4], but to our knowledge, there have been no work to analyze the user behavior from network structural point of view. We believe that the work along this line will be useful in understanding the user behavior and helps building a preference model of Twitter users.

The paper is organized as follows. We briefly explain the various measures we adopted in our analysis in [2], three networks (Favorite, Follow, and Mention) in [3]. Then we report the experimental results in [4] and provide some discussions regarding our observations in [5]. We end this paper by summarizing the major finding and mentioning the future work in [6].

2 Analysis Methods

According to [1], we define the structure of a network as a graph. A graph $G = (V, E)$ consists of a set V of nodes (vertices) and a set E of links (edges) that connect pairs of nodes. Note that in our Favorites, Followers or Mentions network, a node corresponds to a Twitter user, and a link corresponds to

favoring, following, or mentioning between a pair of users. If two nodes are connected by a link, they are adjacent and we call them neighbors. In directed graphs, each directed link has an origin (source) and a destination (target). A link with origin $u \in V$ and destination $v \in V$ is represented by an ordered pair (u, v) . A directed graph $G = (V, E)$ is called a bipartite graph, if V is divided into two parts, V_x and V_y , where $V = V_x \cup V_y$, $V_x \cap V_y = \emptyset$, and $E \subset \{(u, v); u \in V_x, v \in V_y\}$. In directed graphs, we may allow the link set E to contain the same link several times, i.e., E can be a multiset. If a link occurs several times in E , the copies of that link are called parallel links. Graphs with parallel links are also called multigraphs. A graph is called simple, if each of its links is contained in E only once, i.e., if the graph does not have parallel links. In what follows, we describe our analysis methods for each type of graphs.

2.1 Methods for Simple Graph

A graph $G' = (V', E')$ is a subgraph of the graph $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. It is an induced subgraph if E' contains all links $e \in E$ that connect nodes in V' . A directed graph $G = (V, E)$ is strongly connected if there is a directed path from every node to every other node. A strongly connected component of a directed graph G is an induced subgraph that is strongly connected and maximal. A bidirected graph $\tilde{G} = (V, \tilde{E})$ is constructed from a directed graph $G = (V, E)$ by adding counterparts of the unidirected links, i.e., $\tilde{E} = E \cup \{(v, u); (u, v) \in E\}$. A weakly connected component of a directed graph G is an induced subgraph from V' obtained as a strongly connected component of the bidirected graph \tilde{G} . We analyze the structure of our networks in terms of the connectivity using these notions.

In a directed graph $G = (V, E)$, the out-degree of $v \in V$, denoted by $d^+(v)$, is the number of links in E that have origin v . The in-degree of $v \in V$, denoted by $d^-(v)$, is the number of links with destination v . The average degree d is calculated by

$$d = \frac{1}{|V|} \sum_{v \in V} d^-(v) = \frac{1}{|V|} \sum_{v \in V} d^+(v) = \frac{|E|}{|V|}. \quad (1)$$

Here $|\cdot|$ stands for the number of elements for a given set. The correlation between in- and out-degree, denoted by c , is calculated by

$$c = \frac{\sum_{v \in V} (d^-(v) - d)(d^+(v) - d)}{\sqrt{\sum_{v \in V} (d^-(v) - d)^2} \sqrt{\sum_{v \in V} (d^+(v) - d)^2}}. \quad (2)$$

On the other hand, the in-degree distribution $id(k)$ and the out-degree distribution $od(k)$ with respect to degree k are respectively defined by

$$id(k) = |\{v \in V; d^-(v) = k\}|, \quad od(k) = |\{v \in V; d^+(v) = k\}|. \quad (3)$$

We analyze the statistical properties of these degree distributions.

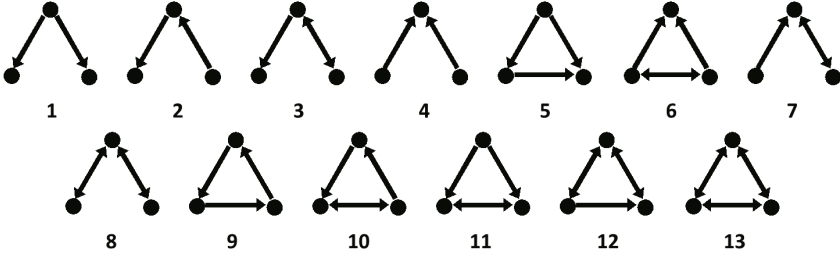


Fig. 1. Network motifs patterns

Network motifs are defined as patterns of interconnections occurring in graphs at numbers that are significantly higher than those in randomized graphs. In our analysis, we focus on three-node motifs patterns and Figure 1 shows all thirteen types of three-node connected subgraphs (motifs patterns). According to [5], we also use randomized graphs, each node of which has the same in-degree and out-degree as the corresponding node has in the real network [6]. A significance level of each motifs pattern i is evaluated by its z -score z_i , i.e.,

$$z_i = \frac{f_i - J^{-1} \sum_{j=1}^J g_{j,i}}{\sqrt{J^{-1} \sum_{j=1}^J (f_i - J^{-1} \sum_{j=1}^J g_{j,i})^2}}, \tag{4}$$

where J is the number of randomized graphs used for evaluation, and f_i and $g_{j,i}$ denote the numbers of occurrences of motifs pattern i in the real graph and the j -th randomized graph, respectively. By this motifs analysis, we attempt to uncover the basic building blocks of our networks.

2.2 Visualization of Bipartite Graph

A bipartite graph is a graph whose nodes can be divided into two disjoint sets V_x and V_y such that every links connects a vertex in V_x to one in V_y . We can construct a bipartite graph from a directed graph by setting $V_x = \{u; (u, v) \in E\}$ and $V_y = \{v; (u, v) \in E\}$, and regarding that any element in V_x is different from any element in V_y . Further, according to [2], we describe a bipartite graph visualization method for our analysis. For the sake of technical convenience, each set of the nodes, V_x and V_y , is identified by two different series of positive integers, i.e., $V_x = \{1, \dots, m, \dots, M\}$ and $V_y = \{1, \dots, n, \dots, N\}$. Here M and N are the numbers of the nodes in V_x and V_y , i.e., $|V_x| = M$ and $|V_y| = N$, respectively. Then, the $M \times N$ adjacency matrix $\mathbf{A} = \{a_{m,n}\}$ is defined by setting $a_{m,n} = 1$ if $(m, n) \in E$; $a_{m,n} = 0$ otherwise. The L -dimensional embedding position vectors are denoted by \mathbf{x}_m for the node $m \in V_x$ and \mathbf{y}_n for the node $n \in V_y$. Then we can construct $M \times L$ and $N \times L$ matrices consisting of these position vectors, i.e., $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_M)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^T$. Here \mathbf{X}^T stands for the transposition of \mathbf{X} . Hereafter, we assume that nodes in subset V_x are located on the inner

circle with radius $r_x = 1$, while nodes in V_y are located on the outer circle with radius $r_y = 2$. Note that $\|\mathbf{x}_m\| = 1$, $\|\mathbf{y}_n\| = 2$.

The centering (Young-Householder transformation) matrices are defined as $\mathbf{H}_M = \mathbf{I}_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^T$, $\mathbf{H}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ where \mathbf{I}_M and \mathbf{I}_N stands for $M \times M$ and $N \times N$ identity matrices, respectively, and $\mathbf{1}_M$ and $\mathbf{1}_N$ are M - and N -dimensional vectors whose elements are all one. By using the double-centered matrix $\mathbf{B} = \{b_{m,n}\}$ that is calculated from the adjacency matrix \mathbf{A} as $\mathbf{B} = \mathbf{H}_M\mathbf{A}\mathbf{H}_N$, we can consider the following objective function with respect to the position vectors $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_M)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)^T$.

$$S(\mathbf{X}, \mathbf{Y}) = \sum_{m=1}^M \sum_{n=1}^N b_{m,n} \frac{\mathbf{x}_m^T \mathbf{y}_n}{r_x r_y} + \frac{1}{2} \sum_{m=1}^M \lambda_m (r_x^2 - \mathbf{x}_m^T \mathbf{x}_m) + \frac{1}{2} \sum_{n=1}^N \mu_n (r_y^2 - \mathbf{y}_n^T \mathbf{y}_n), \quad (5)$$

where $\{\lambda_m \mid m = 1, \dots, M\}$ and $\{\mu_n \mid n = 1, \dots, N\}$ correspond to Lagrange multipliers for the spherical constraints, i.e., $\mathbf{x}_m^T \mathbf{x}_m = r_A^2$ and $\mathbf{y}_n^T \mathbf{y}_n = r_B^2$ for $1 \leq m \leq M$ and $1 \leq n \leq N$. By maximizing $S(\mathbf{X}, \mathbf{Y})$ defined in Equation (5), we can obtain our visualization results, \mathbf{X} and \mathbf{Y} for a given bipartite graph.

2.3 Methods for Multigraph

For multigraphs, we denote the number of links from node u to v , i.e., (u, v) , as $m_{u,v}$. Note that favoring or mentioning between a pair of users may occur several times during the observed period. We also denote an in-neighbor node set of node v by $A(v) = \{u; m_{u,v} \neq 0\}$, and an out-neighbor node set of node v by $B(v) = \{w; m_{v,w} \neq 0\}$. Then we can consider a node set $C(k) = \{v; |A(v)| = k\}$ for which the number of in-neighbor nodes is k , and a node set $D(k) = \{v; |B(v)| = k\}$ for which the number of out-neighbor nodes is k . Thus, by using these notations, with respect to the number of neighbors k , we can define the in-neighbor distribution $id(k)$ and the out-neighbor distribution $od(k)$ as follows:

$$in(k) = |C(k)|, \quad on(k) = |D(k)|. \quad (6)$$

Note that in case of simple directed graphs, the in- and out-neighbor distributions are simply called the in- and out-degree distributions, respectively.

Now, we define a set of nodes whose in-degree are not zero by $V^- = \{v \in V; deg^-(v) > 0\}$, and a set of nodes whose out-degree are not zero by $V^+ = \{v \in V; deg^+(v) > 0\}$.

Then, we can define the average in-multiplicity $m^-(v)$ for $v \in V^-$ and the average out-multiplicity $m^+(v)$ for $v \in V^+$ as follow:

$$m^-(v) = \frac{1}{|A(v)|} \sum_{u \in A(v)} m_{u,v}, \quad m^+(v) = \frac{1}{|B(v)|} \sum_{w \in B(v)} m_{v,w}. \quad (7)$$

For a multigraph, we can define the average in-multiplicity m^- and the average out-multiplicity m^+ as follow:

$$m^- = \frac{1}{|V^-|} \sum_{v \in V^-} m^-(v), \quad m^+ = \frac{1}{|V^+|} \sum_{v \in V^+} m^+(v). \tag{8}$$

On the other hand, with respect to number of neighbors $k(> 1)$, we can define the average link multiplicity $im(k)$ for a node set $C(k)$, and the average link multiplicity $om(k)$ for a node set $D(k)$ as follows:

$$im(k) = \frac{1}{|C(k)|} \sum_{v \in C(k)} m^-(v), \quad om(k) = \frac{1}{|D(k)|} \sum_{v \in D(k)} m^+(v). \tag{9}$$

Similarly, for each node $v \in V$, we can define the in-Gini coefficient $g^-(v)$ for $v \in V^-$ and the out-Gini coefficient $g^+(v)$ for $v \in V^+$ as follow:

$$g^-(v) = \frac{\sum_{(u,x) \in A(v) \times A(v)} |m_{u,v} - m_{x,v}|}{2(|A(v)| - 1) \sum_{u \in A(v)} m_{u,v}}, \quad g^+(v) = \frac{\sum_{(w,x) \in B(v) \times B(v)} |m_{v,w} - m_{v,x}|}{2(|B(v)| - 1) \sum_{w \in B(v)} m_{v,w}}. \tag{10}$$

For a multigraph, we can define the average in-multiplicity m^- and the average out-multiplicity m^+ as follow:

$$g^- = \frac{1}{|V^-|} \sum_{v \in V^-} g^-(v), \quad g^+ = \frac{1}{|V^+|} \sum_{v \in V^+} g^+(v). \tag{11}$$

With respect to number of neighbors $k(> 1)$, we can define the average Gini coefficient $ig(k)$ for a node set $C(k)$, and the average Gini coefficient $og(k)$ for a node set $D(k)$ as follows:

$$ig(k) = \frac{1}{|C(k)|} \sum_{v \in C(k)} g^-(v), \quad og(k) = \frac{1}{|D(k)|} \sum_{v \in D(k)} g^+(v). \tag{12}$$

Here note that the gini coefficient has been widely used for evaluating inequality in a market [7]. We use this index to evaluate inequality between favoring and mentioning.

3 Summary of Data

We briefly explain the data we used in our analysis. These data are retrieved from Favorite, Follow, and Mention of Twitter.

"Favorites" is a function which enables users to bookmark tweets, or to browse them anytime. We constructed a network with the users as nodes, and the Favored/Favoree relations as links. These data are retrieved from Favotter's "Today's best." [2] during the period from May 1st 2011 to February 12th 2012. Because of Favotter's specification, the retrieved tweets are bookmarked by more

² <http://favotter.net/>

than or equal to 5 users. This directed network has 189,717 nodes, 7,077,070 simple links, and 33,456,690 multiple links³.

”Follow” is the most basic function of Twitter. Users can get the new tweets posted by persons they are interested in by specifying whom to follow. We constructed a network with users who posted more than or equal to 200 tweets as nodes, and the follower/followee³ relations as links. These data are retrieved from Twitter search⁴ as of January 31st 2011. This directed network has 1,088,040 nodes and 157,371,628 simple links. Follow network does not have multiple links because users specify their respective followers only once.

”Mentions” are tweets which has the user’s names of the form ”@Screen_name” in the text. We constructed a network with users as nodes, and send/receive relations as links. These data are retrieved from Toriumi’s data⁸ for the period from March 7th 2011 to March 23rd 2011. This directed network has 4,565,085 nodes, 58,514,337 simple links and 193,913,339 multiple links.

Statistics of these networks are described for Tables 1 and 2. Here, WCC1 in Table 1 means the maximal weakly connected components, Em in table 2 means the number of multiple links. Others are defined in section 2.

Table 1 shows that Mentions network has a smaller WCC1 fraction than the other two networks. This is understandable in view of the communication aspect of Mentions because users do not send @-messages to people whom they do not well. Table 2 shows that Favorites network has smaller m^- , m^+ , g^- , and g^+ (see equations 8 and 11) than Mentions. This is understandable because only a few users are heavy favorers and the majorities have much less favorees whereas in Mentions the distribution of the number of mentions of each user is less distorted, which makes the average degree of Mentions network larger than that of Favorites network.

Table 1. Statistics of simple directed networks

	$ V $	$ E $	$ V _{WCC1}$ ($ V _{WCC1}/ V $)	d	c
Favorites	189,717	7,077,070	189,626 (99.9%)	37.3	0.2109
Follow	1,088,040	157,371,628	1,079,986 (99.3%)	144.6	0.7354
Mentions	4,565,085	58,514,337	1,839,189 (40.3%)	3.2	0.0387

Table 2. statistics of multi directed networks

	$ V $	$ Em $	d	m^-	m^+	g^-	g^+
Favorites	189,717	33,456,690	176.3505	2.1211	1.5024	0.2054	0.0851
Mentions	4,565,085	193,913,339	38.2894	3.6977	3.6574	0.3985	0.2138

³ The number of simple links means that we count the multiple links between a pair of nodes as a single link.

⁴ <http://yats-data.com/yats/>

4 Results

In this section, we report the results of analysis using various measures explained in [2](#)

4.1 Simple Directed Graph

As seen from Table [1](#), Favorites and Follow networks have each a large weakly connected component which includes almost all nodes but Mentions network is not so. Since Mentions network is too large to analyze for all nodes, we use WCC1 in the following analysis for Mentions network.

Degree Distribution. Figures [2](#), [3](#), [4](#), [5](#), [6](#), and [7](#) are the results of degree distribution of the three networks. Blue and red diamond marks indicate *id* and *od* (see equation [\(3\)](#)), respectively. The vertical axis indicates the number of nodes in logarithmic scale. From these pictures, we see that all the networks can be said to have a scale-free property for both in-degree or out-degree.

Network Motif. Figures [8](#) and [9](#) are the results of network motif analysis. The horizontal axis indicates the motif number explained in [4](#). In Figure [8](#) the vertical axis indicates the frequency of appearance in logarithmic scale, and in Figure [9](#) the vertical axis indicates *z*-score (see equation [\(4\)](#)) in logarithmic scale. Red and cyan bars mean positive score and negative score respectively. From these figures, we see that there are three predominant motifs: patterns 13, 12, and 8, which are all characterized by having mutual links. The results of Follow and Mentions networks are similar to these figures, so we omit showing these results.

4.2 Visualization of Bipartite Graph

Figure [10](#) is the result of visualization of bipartite graph of Favorites. In this analysis we used the data retrieved from only July 1st to 7th 2011 because so many links obscure the graph. Nodes on the outer circle are Favorers, and nodes on the inner circle are Favorees. Blue and Red nodes are users who are ranked Favored/Favoree's top 10. Only links with more than or equal to 10 multiplicity are shown by gray lines.

NHK_PR is the official account of NHK's PR section⁵, and sasakitoshinao is the account of freelance journalist. His tweets are on serious and important topics, for instance, current news or opinions about it. On the other hand, kaiten_keiku and Satomii_Opera are regular users of Twitter, and their tweets are often negative and/or "geeky".

From this figure, we see there are two clusters of popular users which are characterized by their content of tweets, one with serious and important tweets and the other with negative and/or geeky tweets.

⁵ Japan Broadcasting Corporation.

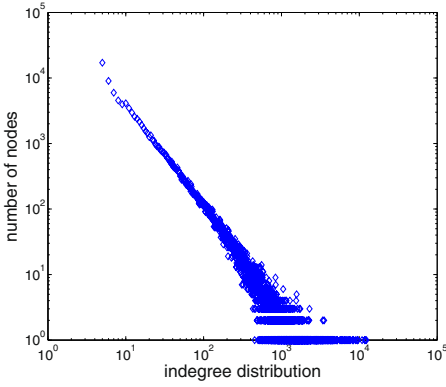


Fig. 2. Favorites network in-degree

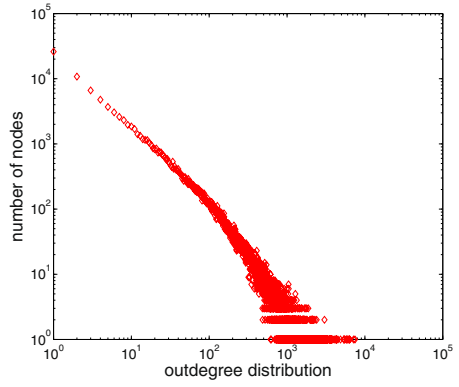


Fig. 3. Favorites network out-degree

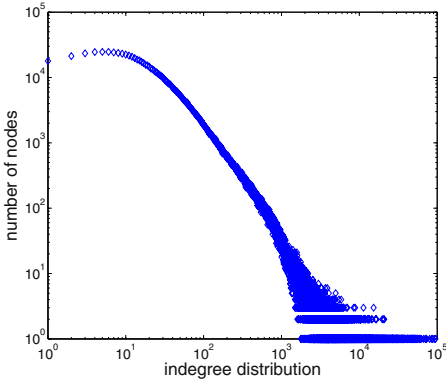


Fig. 4. Follow network in-degree

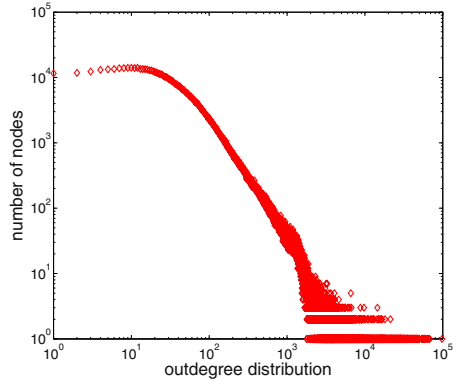


Fig. 5. Follow network out-degree

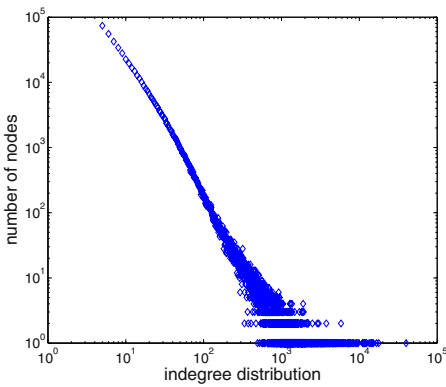


Fig. 6. Mentions network in-degree

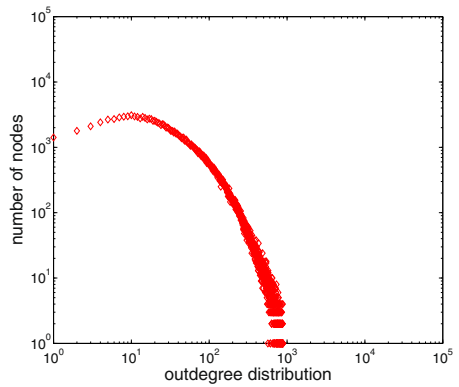


Fig. 7. Mentions network out-degree

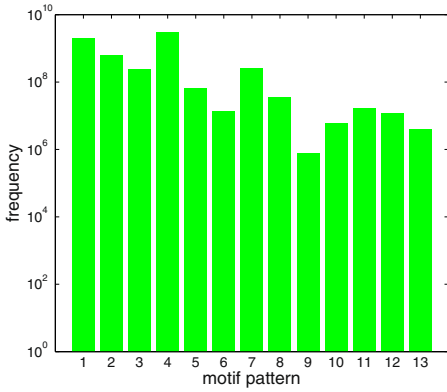


Fig. 8. Favorites network motif (frequency)

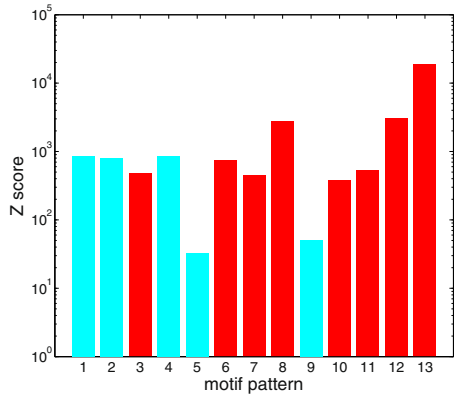


Fig. 9. Favorites network motif (z-score)

Only links with more than or equal to 10 multiplicity are shown

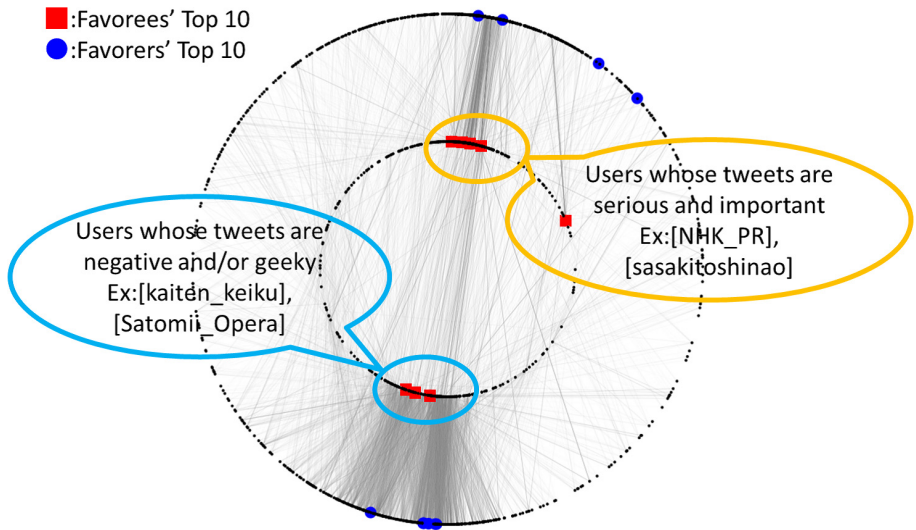


Fig. 10. Bipartite Graph Visualization

4.3 Multiple Directed Graph

In this subsection, we show the results of analysis using the measures explained in 2.3. In all the figures below (Figures 11 to 22), plots in blue squares are for in-degree, plots in red squares are for out-degree and plots in green circles are for randomized networks. Horizontal axes are all in logarithmic scale.

Degree Distribution. Figures 11, 12, 13 and 14 are the results of degree distribution (see equation (6)) for Favorites and Mentions networks. The vertical axes are frequency (the number of nodes) in logarithmic scale. From these figures, we see that both networks have a scale-free property, same as the simple directed networks 4.1. We notice that the distributions for the randomized Mentions network are shifted right to the real Mentions network, but this is not so for Favorites network.

Average Multiplicity. Figures 15, 16, 17 and 18 are the average multiplicity (see equation (7)) for the both networks. The vertical axes are in logarithmic scale. We notice the difference in correlation between the two networks. On the average, there are positive correlations between the average multiplicity and the degree for Favorites network (Figures 15 and 16), but the correlations change from positive to negative as the degree increases for Mentions network (Figures 17 and 18). Furthermore, the average multiplicity of randomized Favorites network behaves similarly to the real Favorites network, but that of randomized Mentions network is almost flat across all the range of degree.

Gini Coefficient. Figures 19, 20, 21 and 22 are the results of Gini coefficient (see equation (10)) for the both networks. The vertical axes are in linear scale. Correlations between the Gini coefficient and the degree and the relation between the real and the randomized networks are similar to those for the average multiplicity, i.e., positive correlations for Favorites network (Figures 19 and 20), positive to negative correlations for Mentions network (Figures 21 and 22) and more positive correlations for the randomized Favorites network than the randomized Mentions network.

5 Discussion

The results in subsections 4.1 and 4.3 revealed that all the three networks have the scale-free property, but we notice that the variance in the degree distributions for Mentions network is smaller in high out-degree nodes than others. We conjecture that this is due to the communication aspect of Mention function, i.e. users do not send many @-messages to people they do not know well and, thus, there are probably no big hub nodes in Mentions network. Further, this also explains that the fraction of the maximal weakly connected component (defined in subsection 3) is smaller than the other networks.

The results in subsection 4.1 revealed that there are a few numbers of predominant motifs that are characteristic of having mutual links. This accounts for the fact that, taking Favorites as example, mutual links are easily created between users who have similar tastes because Favorites network is driven by preference.

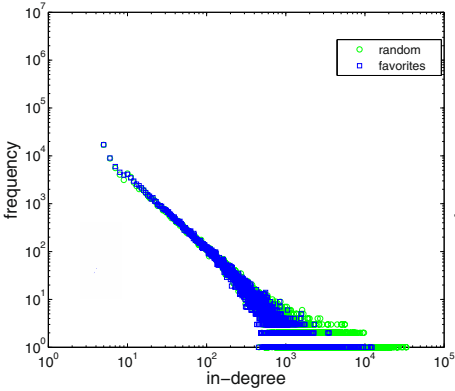


Fig. 11. Favorites in-degree

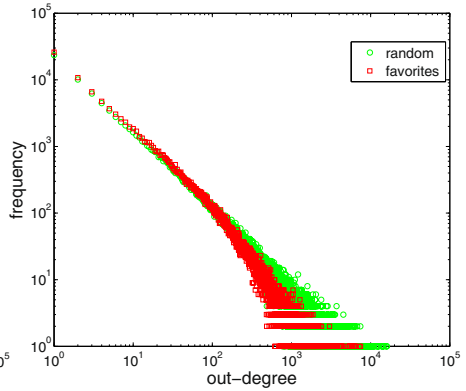


Fig. 12. Favorites out-degree

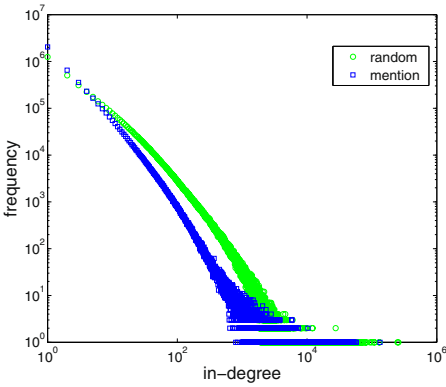


Fig. 13. Mentions in-degree

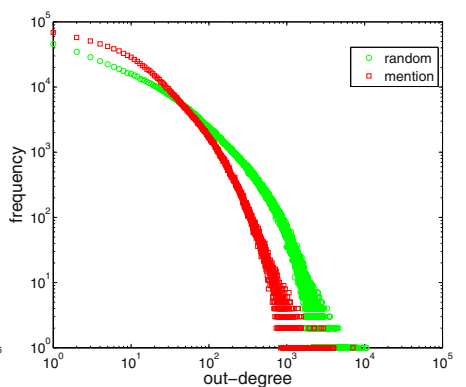


Fig. 14. Mentions out-degree

The results in subsection 4.2 that there are two clusters of popular users each corresponding to a particular type of tweets are quite natural and understandable. Whether these two are the unique tweets and there are no other such tweets remains to be explored.

The results in subsection 4.3 indicate that there are substantial difference in the distributions of multiplicity and Gini coefficient for high degree nodes between Favorites and Mentions networks. This is explainable considering the difference in nature of the two functions, Mentions network is driven by communications between users. Sending/receiving of @-message to/from many people become less practical, thus less frequent for high degree nodes. Favorites network is driven by preference. Expressing preference (bookmarking Favorees' tweets) is much easier than sending/receiving message, thus relatively more frequent for high degree nodes.

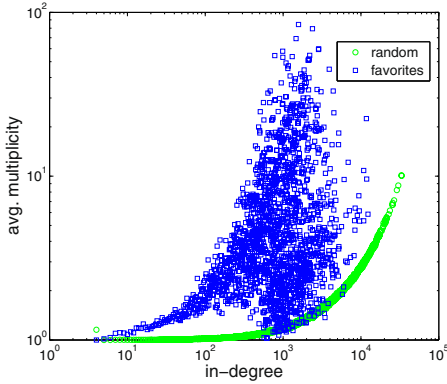


Fig. 15. Favorites in-degree

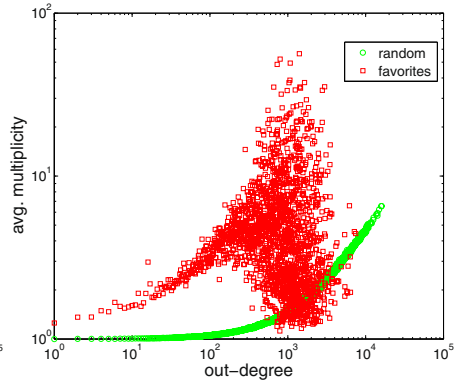


Fig. 16. Favorites out-degree

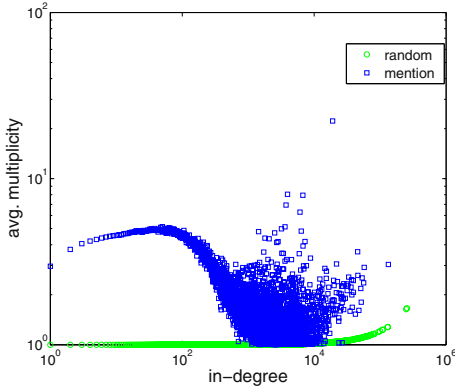


Fig. 17. Mentions in-degree

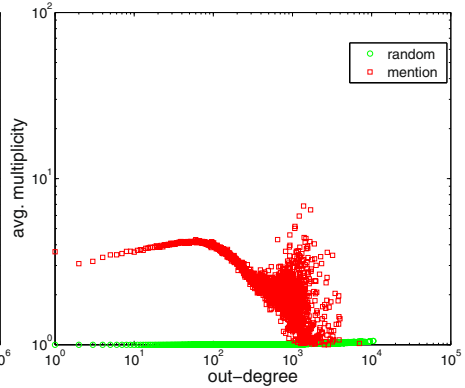


Fig. 18. Mentions out-degree

The results in subsection 4.3 revealed that there are positive correlations between the Gini coefficient and the degree for all the range of degree for Favorites network, but not so for Mentions network. This may suggest that Favorers in high out-degree tends to preferentially bookmark specific Favorees' tweets, and vice versa for Favorees in high in-degree.

6 Conclusion

With the final goal of constructing a new user preference model in daily activities in mind, we analyzed, from the network structure perspective, the similarity and difference in the user behavior of the three functions of Twitter: Favorite, Follow and Mention. User behavior is embedded in the logs that users carried out these functions, which are represented by directed graphs. Favorite function was analyzed using three different graph representations: a simple graph, a multigraph and a bipartite graph, Follow function by one graph representation:

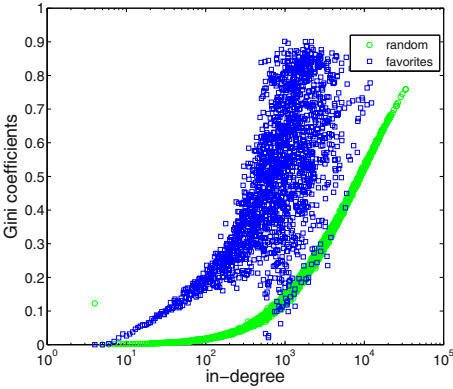


Fig. 19. Favorites in-degree

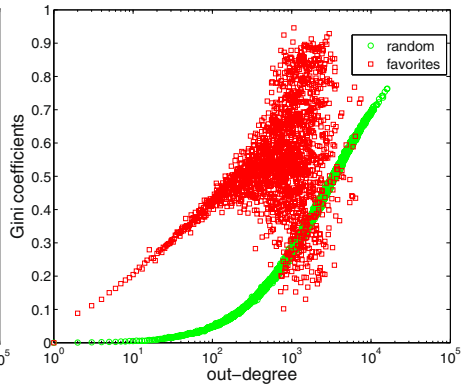


Fig. 20. Favorites out-degree

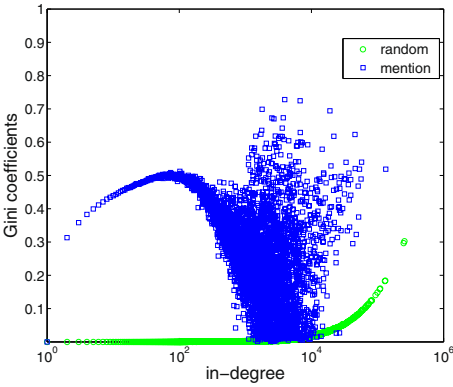


Fig. 21. Mentions in-degree

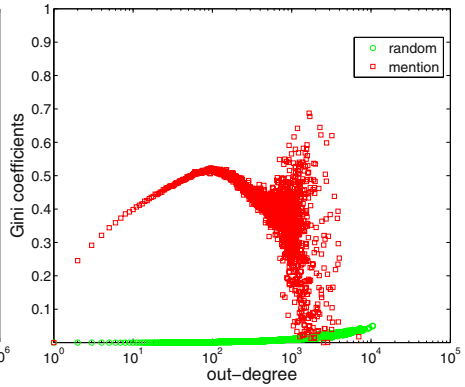


Fig. 22. Mentions out-degree

a simple graph, and Mention function by two graph representations: a simple graph and a multigraph. We used three real world Twitter logs to create these directed graphs and performed various kinds of analysis using several representative measures for characterizing structural properties of graphs, and obtained several salient features.

Major findings are that 1) Favorites and Followers networks have a very large connected component but Mentions network is not, 2) all the three networks (both simple and multiple) have the scale-free property in degree distribution, 3) all three networks (simple) have predominant three-node motifs having mutual links, 4) Favorites networks have freaks in Gini coefficient (one of the measures), 5) Favorites networks have two clusters of popular users, and 6) Favorites and Mentions networks differ in structure for high degree nodes in case of multigraph representation reflecting that Favorite operation is much easier than Mention operation although they are similar in case of simple graph representation.

As an immediate future work, we plan to obtain betweenness centrality, closeness centrality, or k-core percolation of Favorites network represented as a multi-graph to further characterize use behavior and hopefully to extract enough regularity to model user preference, and pursue the literature review and usefulness of the model.

Acknowledgments. This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-11-4111, JSPS Grant-in-Aid for Scientific Research (C) (No. 22500133).

References

1. Brandes, U., Erlebach, T. (eds.): Network Analysis. LNCS, vol. 3418, pp. 293–317. Springer, Heidelberg (2005)
2. Fushimi, T., Kubota, Y., Saito, K., Kimura, M., Ohara, K., Motoda, H.: Speeding Up Bipartite Graph Visualization Method. In: Wang, D., Reynolds, M. (eds.) AI 2011. LNCS (LNAI), vol. 7106, pp. 697–706. Springer, Heidelberg (2011)
3. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. *First Monday* 14(1) (2009)
4. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600 (2010)
5. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. *Science* 298, 824–827 (2002)
6. Newman, M.E.J.: The structure and function of complex networks. *SIAM Review* 45, 167–256 (2003)
7. Salganik, M.J., Dodds, P.S., Watts, D.J.: Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311, 854–856 (2006)
8. Toriumi, F., Shinoda, K., Kurihara, S., Sakaki, T., Kazama, K., Noda, I.: Disaster Changes Social Media. In: Proceedings of the 7th Conference of JWEIN (2011)

User-Oriented Product Search Based on Consumer Values and Lifestyles

Hesam Ziaei, Wayne Wobcke, and Anna Wong

School of Computer Science and Engineering
University of New South Wales
Sydney NSW 2052, Australia
{hzia568,wobcke,annawong}@cse.unsw.edu.au

Abstract. Product search engines are essentially unchanged since the inception of online shopping, providing basic browsing by category and “advanced” keyword search. This paper presents a *user-oriented* product search method based on consumer values and lifestyles that correspond to user purchasing criteria rather than technical specifications. The technique is suited to relatively infrequent purchases where users possess little domain or market knowledge and existing interfaces are difficult to use. We show how to construct a knowledge base to support a user-oriented product search engine without the need for a domain expert to manually label the items. We present *Lifestyle Car Finder*, a user-oriented product search system in the domain of new cars. The system incorporates various modes of navigation (search refinement, a new form of critiquing adaptive to the user’s query, and breadcrumb trails) and decision support (similar car comparison, explanations and technical specifications). We report on a user study showing that, broadly speaking, users were highly satisfied with the system and felt they were confident in their decisions.

1 User-Oriented Product Search

Most online shopping search interfaces provide only basic browsing and “advanced” keyword search using technical features. These interfaces are difficult to use for users without domain knowledge, especially when products are highly technical and/or where the market contains many items that are very similar and thus hard to differentiate. The typical difficulty faced by everyday users is to express their information needs in the manner expected by the search engine. For some systems, users have to know the brand or the technical name of the product (e.g. multi-function printer), however naive users often start with little or no domain knowledge and so need to *construct* their product and brand preferences through their interaction with the system, Bettman, Luce and Payne [2]. With text-based keyword search interfaces, the user has to enter sufficiently precise keywords for the system to return useful results, however naive users only have a rough idea of their high-level requirements, and do not know how to express their needs in technical terms. Adding to the problem, naive users are often casual users, so there is limited interaction history to tailor the interaction.

This motivates the need for systems that enable users to express their requirements using needs-based attributes that are meaningful to the user and in accord with the

naive user's level of expertise. As far as we are aware, Felix *et al.* [6] were the first to explicitly identify needs-based attributes as a means to support naive users in an online shopping environment, where a *needs-based attribute* is defined as an attribute that fits the purpose of the product rather than a technical feature. Their system used particular needs-based attributes of digital cameras defined in terms of technical attributes but had only limited search: it was cast as a "sales assistant" that asked questions to elicit user preferences about attributes of the desired product (based on needs or technical features) and provided a list of items for consideration (the top 3 of 25 cameras). Users were asked to compare the two types of interaction. Felix *et al.* [6] also suggested that users learn during the course of their interaction with the system. Stolze and Ströbel [14] apply this "teaching" metaphor to the online product shopping domain and propose an (unimplemented) system design that combines needs-based and technical attributes in a constraint satisfaction framework, aiming to support users as they gradually develop from being novices (requiring needs-based assistance) to experts (preferring technical information). Since this approach is hypothetical, aspects of how users navigate through such a search space have not been evaluated.

In this paper, we present *user-oriented product search*, applying needs-based attributes in a search engine over a realistic product database (the space of new cars). The basic idea of user-oriented search is to map search in the space of needs-based attributes to search in the space of technical features. We address the following issues: (i) which needs-based attributes to select in the design of a system, (ii) how to efficiently define those attributes in terms of technical features, and (iii) what support for navigation and decision support should be provided to the user.

To determine suitable needs-based attributes, values and lifestyles have been long-identified as determinants of purchasing intent. The work presented here in the new car domain is based on consumer values, though in other domains lifestyle may be more important. A standard marketing methodology elicits values from consumers which can then be clustered for market segmentation. Perhaps remarkably, the broad values underlying the purchase of a new car (for illustration but this also applies in other domains) are relatively stable over time, e.g. Vinson, Scott and Lamont [15] listed unleaded petrol, high speed, handling, quality workmanship, advanced engineering and low level pollution emission for one class of buyers, and smooth riding, luxurious interior, prestige, large size and spacious interior for another class of buyers. The concept of lifestyle is more difficult to pin down (see Zablocki and Kanter [16]) but specifically relates to a pattern of consumption reflecting individual choices about how to spend time and money. The VALS system of values and lifestyles has been used extensively for market segmentation, Holman [8]. The main point is that the consumer decision criteria on which to base a user-oriented product search engine are fairly consistent over the population and relatively easy to determine.

In user-oriented product search, attributes of items that relate to people's purchasing criteria (values and lifestyles) are mapped to the set of items to produce a ranking of each item along each high-level dimension. The information needs of the user are expressed through the selection of certain attributes or by using a sliding scale to indicate the preferred value of each attribute. The user can then browse through the space of items based on their needs-based attributes using a variety of navigation techniques,

such as search refinement, critiquing and breadcrumb trails, and come to a decision concerning the items for further consideration using a variety of types of decision support, such as explanations and technical specifications. The critical problem in any user-oriented product search system is to determine suitable needs-based attributes, then to define those attributes in terms of technical features. The size of the product database is also an issue: we give an efficient method to construct a knowledge base to support a user-oriented product search engine without the need for a domain expert to manually label all the items.

Due to the complexity of the search space, a user-oriented search system requires a variety of navigation modes. One mode of navigation used is critiquing [4], which is suitable for naive users to explore an unfamiliar space. In contrast to existing critiquing systems, our critiques are in the space of high-level needs-based attributes, and in order to provide useful critiques to casual users, we develop a new adaptive critiquing method based on the user's initial query, where the two attributes involved in a compound critique are drawn from those the user considers more important (as derived from their initial query). The aim is to "personalize" the critiques to the user's interests and to provide domain information (that two attributes are typically correlated).

We also present *Lifestyle Car Finder*, a user-oriented product search system for the domain of new cars. The system uses 11 high-level needs-based attributes, including 4 *primary* car types (family car, sports car, city car and off-roader) and 7 *secondary* attributes (performance, safety features, luxury features, fuel efficiency, eco-friendliness, towing capacity and car size), plus price. The primary high-level attributes are different classes of cars that may be overlapping, while the secondary attributes are understood with respect to the primary class chosen, (so that, for example, a large city car is similar in size to small family car). The new car market is highly dynamic (around 4–5 new cars introduced each week in Australia), highly sophisticated (with cars designed for many different market niches) and highly complex (with around 300 car makes and 2000 models/model variants on the market at any time). It is thus impractical for users to examine a complete list of new cars when making a decision, motivating the use of a hybrid approach to navigation and decision support that provides the user with multiple options at each step. Our system contrasts with the early *Car Navigator* of Hammond, Burke and Schmitt [7], which used a database of 600 cars manually labelled into 4 different classes, and where, although users could choose a critique (such as *sportier*) that modified several technical features simultaneously, the user is still presented with a single result along with its technical features and unit critiques along each technical dimension.

The contributions of this paper are to provide: (i) a method for building user-oriented product search engines using needs-based attributes, (ii) an efficient method of building a knowledge base to support such a search engine, involving a new navigation method, query-based adaptive critiquing, and (iii) a prototype system in the domain of new cars with a user study demonstrating the effectiveness of the system. The paper is organized as follows. In the next section, we describe how a knowledge base for use in a user-oriented search system is defined. The following two sections present a description and user study of the *Lifestyle Car Finder*, focusing on the use of the navigation methods and decision support.

2 User-Oriented Product Search Engine

In this section, we describe an efficient method of constructing a knowledge base suitable for a user-oriented search engine, illustrated with the example of the *Lifestyle Car Finder*. This requires a ranking of each item for each needs-based attribute, and a search and retrieval model for mapping user queries into database queries.

2.1 Ranking Functions for Needs-Based Attributes

For each needs-based attribute, a function defines a ranking of items in the database indicating how well the item scores on that dimension. Our system uses weighted sums of technical specifications to define these functions. For example, a *luxury* attribute might be defined using:

$$\text{luxury} = 0.2 * (\text{power}) + 0.4 * (\text{suspension}) + 0.2 * (\text{length}) + 0.2 * (\text{trimming type})$$

While information other than technical specifications could be used to define the ranking functions, such as subjective information extracted from user ratings or reviews, it is a key design criterion for our approach in the domain of new cars that the rankings be objectively determined from the properties of cars themselves, so that each ranking can be validated by a domain expert and in principle justified to the user (even though these justifications will not be presented). This sometimes produces some unexpected results, since some cars that users *think* have high values on certain attributes, probably as a result of advertising, actually have lower values. In this way, the system is “recommending” cars based on their rankings, which in turn partially reflect expert opinion, rather than just returning a list of search results.

The first task is to define the set of needs-based attributes. The aim is to have a collection of attributes that are meaningful to the user, cover the space of items adequately (so that all items can be retrieved by some query), differentiate the items sufficiently, and that can be defined in objective terms. Some attributes may overlap due to natural correlations in the domain (e.g. fuel efficiency correlates with eco-friendliness). Since a domain expert has extensive market knowledge and manufacturers tailor products to this market, finding suitable attributes in the new car domain has not been problematic, though not all consumer values can be used (for example, there is no data on the reliability and resale value of new cars).

The main challenge to building a knowledge base in support of user-oriented search is to decide what technical features to use in the definition of each ranking function and how much weight each feature should hold. We use ranking functions that are linear weighted sums of technical features. To define the functions, we use a variety of techniques: for attributes where sufficiently complete data is available, expert knowledge is used to construct a ranking function, whereas for those attributes for which data is inadequate, machine learning techniques are used to infer missing data and define the ranking function. The expert provided initial ranking functions using a small number of features and initial weights. To determine additional features, we used DBSCAN and EM clustering to discover clusters of cars based on their technical specifications. The weights were then modified to further differentiate each cluster. Clustering was also

used to find a threshold for class membership. For example, in the *sports car* class, we defined a threshold below which a car is not considered a sports car at all. To further refine the ranking functions and their weights, we used the Pearson correlation coefficient to determine the correlation of different attributes in each primary class.

For expert-defined ranking functions, it is impractical for the expert to manually rank all items for each needs-based attribute, since there are typically a large number of items (around 2000 new cars). We developed the following iterative refinement method. Initially, the domain expert decides which technical features should be present in each ranking function. For each function, the expert ranks a small subset of initial items (10 cars consisting of 5 near the top and 5 near the bottom of the ranking). A model is determined using multiple linear regression, and all items in the database are ranked using this model. The domain expert is then presented with these and a further set of items (10 other cars), chosen to give a uniform spread in terms of distance from one particular item that is itself around a distance of $MAD(X)$ from the median of the set of cars X (see below). The expert then re-ranks all the items. This process repeats until the expert judges that no further changes to the ranking function are required. In practice, the process converges after around three iterations for each function, depending on the expert's knowledge and the quality of the available data.

To minimize the work of the domain expert, the additional items presented for re-ranking are chosen from different parts of the ranked list, but centred on a region where items are likely to be densely distributed. A *measure of dispersion*, the *median absolute deviation*, is used to find an item whose distance from the median is around this median absolute deviation (this could be ranked higher or lower than the median item). Then for the set X of size n , every $(n/10)^{th}$ item in order of distance from this item is selected for presentation to the expert at the next stage. $MAD(X)$ is defined as follows:

$$MAD(X) = median_{x_i \in X} (|x_i - median(X)|)$$

where $median(X)$ is the median of the set X . Reasons to choose MAD over other dispersion methods such as standard deviation are its robustness and resistance towards outliers.

In other cases, the ranking functions and their weights are defined using machine learning algorithms. For *safety features*, we started with the ANCAP safety ratings based on crash testing¹. However, ANCAP ratings are limited to a small number of makes and models so do not include all cars in the database. A Decision Tree learner was used to classify the cars without assigned safety ratings into discrete bands. The ranking function for *eco-friendliness* was defined using multiple linear regression. A sample dataset was obtained from *Green Vehicle Guide*, a government website with ratings of cars in Australia². Again, not all makes, models and model variants in the database are given ratings. From this dataset, we were able to specify the attributes that should be used in the ranking function, which could then be defined using linear regression. All final scores are normalized so as to have the same minimum and maximum values across all attributes.

¹ <http://www.ancap.com.au/>

² <http://www.greenvehicleguide.gov.au/>

2.2 Search and Retrieval

After defining ranking functions to rank items according to each needs-based attribute, each item is represented in the form of a vector of its values across all the high-level dimensions. This enables the use of vector operations for *distance* and the *similarity* of items in the search space of high-level attributes. A user's query is taken to be a vector indicating the importance of each needs-based attribute. Since the cardinality of the user query is the same as the cardinality of the items in the database, Euclidean distance is the distance metric used for comparison of items.

To retrieve k objects that are the most similar to a user's query, it is impractical to calculate the distance between the user's query and every item in the database. A more practical approach is to use the threshold algorithm of Fagin, Lotem and Naor [5]. With this algorithm, a threshold for the minimum acceptable similarity is set in advance, allowing the algorithm to stop searching as soon as k instances with similarity values over the threshold are retrieved.

3 Lifestyle Car Finder

The *Lifestyle Car Finder* is a user-oriented product search system that uses needs-based attributes corresponding to consumer values in the domain of new cars. The *Lifestyle Car Finder* uses 11 needs-based attributes, including 4 *primary* car types (family car, sports car, city car and off-roader) and 7 *secondary* attributes (performance, safety features, luxury features, fuel efficiency, eco-friendliness, towing capacity and car size), plus price. These attributes were chosen by a domain expert with extensive knowledge of the market (including manufacturers' market knowledge) and of the new car domain, as attributes suitable to being defined in terms of technical specifications.


The *Lifestyle Car Finder* uses a variety of navigation modes and types of decision support. This is because the search space and intended usage of the system is too complex for single approaches to be adequate. Navigation methods include the familiar techniques of search refinement and breadcrumb trails allowing users to return to an earlier point in the search (on a search path) and the more unusual (for users) mode of critiquing. Decision support includes similar car comparison, explanation and technical specifications.


3.1 Search Interface


Figure 1 shows a screenshot of the *Lifestyle Car Finder* main page. The 4 primary car types are at the top, the 7 secondary attributes with sliding scales follow, and the price attribute is at the bottom, again with a sliding scale. Note that the 4 primary car types are overlapping, indeed manufacturers deliberately produce cars that are in more than one primary class. Nevertheless, users are asked to choose one such class. The reason for this is that the 7 secondary attributes are interpreted in the context of the initial choice of car type. That is, the *qualitative* values of the secondary attributes (as shown on the interface) are *relative* to the primary class, even though cars in the knowledge base are ranked along each high-level dimension on the same scale. For example, a *large city*


Lifestyle Car Finder

Choose the type of car you want:


Family Car


Sports Car


City Car


Off road Car

Choose the features you prefer:

Performance

Not important | Lower performance | Moderate performance | Higher performance

Safety Features

Not important | Basic features | Some features | Most features

Luxury Features

Not important | Basic features | Some features | Most features

Fuel Efficiency

Not important | Less fuel efficiency | More fuel efficiency | High fuel efficiency

Eco-friendly

Not important | Less eco-friendly | More eco-friendly | Very eco-friendly

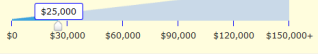
Towing Capacity

Not important | Minimum capacity | Moderate capacity | Maximum capacity

Car Size

Not important | Small | Medium | Large

About how much do you want to pay?



\$0 \$30,000 \$60,000 \$90,000 \$120,000 \$150,000+

Search For Cars

Fig. 1. Lifestyle Car Finder Search Interface

car is around the same size as a *small* family car. The reason for this design decision is so that the 5 point scale can be used to more finely differentiate cars within the same primary class.

Price is surprisingly difficult to model, since all the secondary attributes are of a type where “more” means “better” whereas for price the opposite is the case. In addition, some users may have a tight budget and treat price constraints as absolute, whereas others with a more flexible budget may treat price the same as any other attribute, and allow price tradeoffs with other attributes. Price is used as a filter in the *Lifestyle Car Finder* to remove cars that differ too much from the cars currently viewed by the user. Price is treated as a “semi-hard” constraint, so that some cars in the result set might be slightly over the user’s specified price in order to show a few “near misses” to the user’s query. However, the bounds of this price filter follow a function whose input is the user’s entered price. Where there is a high density of cars around a price, the price filter acts more like a hard constraint, but where there is lower density of cars, the filter acts more like a soft constraint. One reason this makes sense is that when manufacturers compete on price so that there are many cars with similar prices (e.g. with economy cars), people typically have tighter budgets.

Users choose qualitative values for the secondary attributes by moving the sliders. For instance, the user can choose on a 5 point scale from *basic luxury features* to *most luxury features* or a value of *not important*. At this point, deliberately, the meaning of

an attribute such as *luxury features* is somewhat vague for the user. Moreover, different users have different concepts of luxury (based on their differing experience). However, at this stage, our main aim is to present users with a reference point from which they can modify their search using refinement or critiquing according to their own preferences, and so calibrate their definitions with those of the system.

The user's initial search criteria are represented as a vector. By doing so, it is possible to calculate the difference between the user's search vector and vectors representing the cars in the knowledge base. The similarity of the user's query and database items is calculated in weighted manner so that a higher weight is given to the user's primary class.

3.2 Navigation

The *Lifestyle Car Finder* includes a variety of modes of navigation, including search refinement, critiquing and breadcrumb trails. There is an inherent design tradeoff in that multiple search modes make the interface more complicated and potentially more difficult to use, however without these modes the system would not be able to fulfil its intended purpose. This is particularly the case with critiquing, since only a small number of critiques can be presented with any given item. If the critiques do not correspond to the user's interests, another mode of navigation must be available to enable the search to proceed.

The system provides critiques that guarantee one or more results; this means that critiques must be different for each car. Since computing each critique requires, in effect, a query to the database, there is a computational tradeoff between the number of results presented and the number of critiques given per item. We determined that around 30 queries to the database could be performed without perceived loss of response time. Thus the options were to present 10 cars with 3 critiques, 7 cars with 4 critiques, or 5 cars with 5 or 6 critiques. It soon became clear when trialling the different approaches that 3 critiques were too few to be useful, and 5 cars were too few to provide sufficiently many relevant results. Hence the *Lifestyle Car Finder* provides 7 results with up to 4 critiques per item. This design decision partly reflects a domain-dependent property of the new car market in Australia, with 7 results covering all the major manufacturers. To ensure sufficient diversity of the result set, c.f. McCarthy *et al.* [12], Kelly and Bridge [10], model variants of the same car are filtered out: for each model, only the highest ranked variant is presented to the user in the search results page.

Search Refinement. Figure 2 shows the search results page corresponding to the query in Figure 1. At the top of the page, users can refine their initial search criteria by re-adjusting the sliders. This search refinement however, does not allow users to change their primary product class. To change the primary class, users need to start a new search.

Search refinement is useful when the user's perception of one of the primary classes is very different from the ones in the knowledge base. For example, if a user's idea of a *small car* is still smaller than the ones presented on the results page, the user can promptly adjust the slider to get to the desired size.



Fig. 2. Lifestyle Car Finder Results Page

Query-Based Adaptive Critiquing. The *Lifestyle Car Finder* includes a new type of critiquing, *query-based adaptive critiquing* (Figure 2). The system dynamically generates critiques for each item on the search results page, using as a heuristic the primary class, the most important criteria given in the user's query, price and size. The aim is to provide critiques that are relevant to the user's initial search query. Critiquing is dynamic, McCarthy *et al.* [11], in that the critiques are different for each car based on the particular related cars in the database. Each critique is shown on the interface as a button to the right of the car. In the *Lifestyle Car Finder*, critiques contain only two attributes, to simplify the critiques for the user and also to reduce the number of generated critiques. For example, in Figure 2 there are 4 alternatives to the Toyota Yaris. The first indicates an alternative car which is a better city car but less eco-friendly (the user chose *city car* with *very eco-friendly* in Figure 1). The Toyota Yaris and the alternative car(s) are similar (not necessarily the same) in all other aspects.

There are two main methods for generating dynamic compound critiques in the literature, Apriori and Multi-Attribute Utility Theory (MAUT). The Apriori algorithm is used to find association rules in the dataset, Agrawal and Srikant [1]. In this method, frequent itemsets are obtained from the whole dataset. At each cycle, a large number of compound critiques are generated by the algorithm for the reference product. The k best critiques are chosen based on the support value of the rule. Usually rules with lower support value are preferred since rules with lower support value will eliminate

more instances. In other words, the aim is to reduce the decision cycle significantly. However, choosing rules with lower support value also may eliminate items that may be of interest to the user.

The MAUT approach uses an additive utility function to represent the user's preference model, Keeney and Raiffa [9]. A multi-attribute utility function is defined by:

$$U(\langle x_1, \dots, x_n \rangle) = \sum_{i=1}^n w_i V_i(x_i)$$

where n is the number of attributes, x_i is the i^{th} attribute, w_i is the weight (importance) of the i^{th} attribute, and V_i is a value function of the i^{th} attribute. A detailed comparison of the two models is given in Zhang and Pu [17].

With dynamic compound critiquing, items in the neighbourhood of a presented item are examined to generate tradeoffs over more than one attribute that result in items related to the current item in the search. To simplify the critiques for the user and also to reduce the number of generated critiques, each critique in the *Lifestyle Car Finder* involves two needs-based attributes. These attributes are heuristically determined from the user's query based on the primary class chosen, the secondary attributes with highest importance, price and size, defined to guarantee that the critique produces results. The results of a critique are computed as follows. Critique generation in the *Lifestyle Car Finder* uses an additive utility function:

$$U(\langle x_1, \dots, x_n \rangle) = \sum_{i=1}^n w_i V_i(x_i)$$

where n is the number of attributes, x_i is the i^{th} attribute, w_i is the weight (importance) of the i^{th} attribute (from the user's initial query), and V_i is a value function of the i^{th} attribute. First, using a threshold, a subset of cars is chosen based on the difference between the value for the primary class chosen by the user and the corresponding values of cars in the database. This subset of items is defined by:

$$S = \{i \in I : |f_p(i) - V_p| < T\}$$

where I is the the set of all cars, V_p is the value of the primary class chosen by the user, f_p is the corresponding ranking function and T is a threshold. In other words, a subset of cars with values close to the user's chosen value is selected. The threshold value is chosen based on the data dispersion in such a way that by choosing a critique, the results will be "close" in distance to the active item. For example, if a user is viewing an item with low price and low value for *luxury features*, choosing a critique will not take a \$15,000 car to a \$200,000 car (which can occur if the critique concerns other attributes). The use of the threshold creates smaller jumps in the search space compared to search refinement using the sliders. Then a weighted utility function U is used to rank items i in this set according to a preference model derived from the user's query:

$$U(i) = \sum_M w_m |f_m(i) - V_m| - \sum_N w_n |f_n(i) - V_n|$$

where M is the set of the two attributes involved in the critique, N is the set of attributes chosen by the user as important but not involved in the critique, V_m and V_n are the attribute values of the item associated with the critique, and f_m and f_n are the ranking functions for m and n . The weights w_m and w_n are the importance of the attributes chosen by the user with the sliders in Figure 11. The weight of the primary class chosen by the user is always 1. The minus sign before the second summation means that attributes initially chosen by the user not present in the critique receive a negative weight, penalizing changes in these attributes. For example, in a critique such as *more luxury features, higher price*, changes are desired in the *luxury features* and *price* classes while the other attributes should change minimally. The needs-based attributes that are “not important” in the user’s query do not play any role in the utility function. Up to 4 critiques that return the highest utility items are presented to the user.

In the *Lifestyle Car Finder*, as shown in Figure 12, all cars in the results page have critiques presented. This makes it possible for the users to become more familiar with the domain and compare other cars based on their different critiques. The other property of this type of critiquing is that since it is produced based on the available data, it gives the user an idea about the correlations and tradeoffs of different product attributes in the database. If a user is presented with a critique such as *more luxury features, higher price* several times, this indicates that there is an intrinsic tradeoff between luxury and price in the current database. This serves as an implicit explanation of the product attributes and their relationships.

One limitation of the system is that users cannot refine their original search after following a critique. This is because after following a critique, the mapping back to the corresponding search query is lost, in particular as the jumps made by critiques are smaller than those made by refinement. The search history is provided so that the user can backtrack through the search when a critique does not give items preferred by the user.

Search History. The search history shows a “breadcrumb trail” of the user’s previous actions so that a user can go back to a previous page, either a search results page or the result of a critique. However, to limit the size of the history, which is displayed at the top of the page, the search history stores only one path from the most recent search query that includes all critiques and *Similar Cars* links followed.

3.3 Decision Support

Similar Cars. For each car, a *Similar Cars* button returns the most similar cars to the one chosen by the user. This refers to similarity on all high-level dimensions weighted equally. The intention was for users to look at similar cars when finalizing their decisions, though this feature also supports navigation. There are, however, no critiques available with similar cars, since the information used to formulate critiques is absent, so after following the *Similar Cars* results page, the user must return to a previous search or critiquing results page using the search history. It is also at this stage of the search that multiple variants of a single model can appear, whereas for simplification and coverage of the space, multiple model variants are suppressed from the search and

critiquing results pages. Users were expected to realize the importance of model variants and to incorporate them into their decision making.

Explanations. Since users are assumed to have limited knowledge of the car market, it is helpful to provide some kind of “explanation,” c.f. Bilgic and Mooney [3], McSherry [13], to assist users understand the differences in the cars being presented on the search or critiquing results page.

In the *Lifestyle Car Finder*, explanations are found by ranking the returned results along each high-level dimension: if one car is significantly better on one such attribute, this is highlighted in red below the image of the car (see Figure 2).

Technical Specifications. The *Lifestyle Car Finder* is designed for naive users, however it was thought that some users might want to check some technical specifications for some cars, so technical specifications were provided in a simplified form relating to the ranking functions. Users access this information by clicking on the picture or link from the car’s make/model. A question for the user study was the extent to which users made use of even the simplified technical specifications.

4 User Study

To evaluate the *Lifestyle Car Finder*, a small user study was conducted. As well as investigating how users interacted with the system and assessing the overall quality of the results provided by the system, the user study had several other goals. First, the extent to which users used critiquing and the effectiveness of the critiques was explored. Second, the manner and extent to which users made use of navigation techniques and decision support was also investigated.

4.1 Method

There were nine users in the study, 3 male and 6 female, ranging from 18 to 55 years old, with a variety of occupations. All users had previous experience purchasing cars, or were currently looking to buy a car. All users were asked to use the *Lifestyle Car Finder* to work through a series of 7 scenarios corresponding to lifestyles or life stages. An example scenario is as follows. *A family with one school-aged child will shortly be expecting the birth of a second child. The family currently owns a small city car, which needs to be upgraded to a larger car once their second child is born. The family is looking for a car that needs to be safe and large enough to accommodate the family, equipment for the baby such as a pram and taking one or two of the older child’s friends to school. The family’s budget is between \$20,000 and \$40,000, preferably towards the lower end of the range.* For each scenario, the user’s goal was to create a shortlist of one or more cars they would like to further investigate in real life, e.g. go to the car dealer and test drive the car. At the end of the session, users were given a questionnaire to complete, with questions focusing on the quality of the results provided to the user and their satisfaction with the overall performance of the system.

4.2 Results and Discussion

The first five columns of Table 1 give a summary of the number of times users made use of the different types of navigation methods, including *Similar Cars*. Across all scenarios, all users were able to find at least one car that they would be interested in further investigating in real life. The data shows that many users made use of all the main navigation options at least once during the user test (ignoring breadcrumbs, since it is preferable that users do not have to backtrack in their search). That said, users tended to favour one or two of the navigation options and used these continuously over the set of scenarios, rather than use all navigation options equally, or in some sort of order. The pattern regarding choice of search functionalities was dependent on the individual user. For example, User 9 tended to only use critiquing, whereas User 5 tended to use both *Similar Cars* and critiquing to refine their searches.

Inspection of user logs showed that users tended to perform “shallow” searches (up to depth 4 or 5), using critiquing mostly as a means to explore the search space. The last two columns of Table 1 show for each user, the number of scenarios containing cars on shortlists that could *only* have been found in the user’s search using critiquing (this is a conservative estimate of the use of critiquing since other cars appearing on search results pages may also have been found by the user with critiquing). This shows that critiquing was an essential mode of navigation in the system.

Table 1. Use of Navigation Methods and of Critiquing to Find Shortlisted Cars

User	Use of Navigation Method				Use of Critiquing	
	Sliders	Critiques	Similar Cars	Breadcrumbs	#Scenarios	#Shortlist
1	5	14	15	3	5	3
2	4	1	0	1	1	0
3	5	5	9	0	3	2
4	6	2	0	0	1	1
5	1	14	13	2	4	3
6	12	18	13	22	5	1
7	0	5	3	0	4	3
8	2	1	13	0	1	0
9	0	15	7	0	7	3

Users responded positively regarding usability. Users found the system easy to use (average rating 4.22 on a 5-point Likert scale, where 1 = very difficult, 5 = very easy) and were happy about the overall performance of the system (average rating 4.11 on a 5-point Likert scale, where 1 = very unhappy, 5 = very happy). Encouragingly, users also indicated they would use a product like the *Lifestyle Car Finder* in the future (average rating 3.89 on a 5-point Likert scale, where 1 = strongly disagree, 5 = strongly agree).

To assess users’ perceptions of the quality of suggestions produced by the system, users were asked about the relevance of the search results produced, and their confidence in the results produced and the search criteria provided. Users thought the results

generated were quite relevant and useful (average rating 4.11 on a 5-point Likert scale, where 1 = strongly disagree, 5 = strongly agree), felt sufficiently confident in the results generated to investigate the shortlisted models in real life (average rating 3.78 on a 5-point Likert scale, where 1 = not at all confident, 5 = very confident) and considered the search criteria provided to be sufficient and appropriate (average rating 3.78 on a 5-point Likert scale, where 1 = strongly disagree, 5 = strongly agree).

Qualitative feedback from users during the user study raised several interesting points. First, users generally did not find the technical specifications meaningful, even though this information was in a simplified form and related to the needs-based attributes. Consequently, users trusted the system results since they could not even try to validate the results using the technical specifications. When asked what kind of information they would prefer to see, nearly all users mentioned they would have preferred non-technical opinion-based information, such as owner reviews. This feedback indicates that the users in this study really were novice to the domain of cars and were not yet sufficiently expert to find the information meaningful, as they did not fully understand how the technical specifications related to their high-level needs. However, some scenarios did have specific requirements that required users to understand some technical aspects. For example, people who sail boats require a car with sufficient towing capacity and these users specifically sought information about a car's towing capacity. However, although understanding technical information about towing capacity, they still typically wanted non-technical information on other aspects of the cars.

5 Conclusion and Future Work

This paper proposed *user-oriented product search* as an alternative way of searching for items in complex product spaces suitable for naive users with little domain or market knowledge. User-oriented search makes use of functions mapping high-level product attributes that are relevant to a user's needs into the space of technical specifications, and uses a variety of navigation methods, including search refinement, a new form of query-based adaptive critiquing, and breadcrumb trails, and types of decision support such as explanations and technical specifications. Critiques are generated dynamically based on the user's initial query and results from the database. A key feature that distinguishes our work is the ability to handle large product databases without the need for a domain expert to manually label the items.

We described the *Lifestyle Car Finder*, a user-oriented product search system in the domain of new cars. A user study of the system showed that, broadly speaking, users were highly satisfied with the system and felt they were confident in their decisions. Future work for this system is to further assist users with decision making by providing opinion-based information such as owner reviews.

Further work is to apply user-oriented search in other domains involving high-risk purchases such as residential property. Since the markets for such products are highly dynamic, we also plan to investigate methods for maintaining consistent ranking functions over a period of time. Other work is to examine ways to extract the definitions of the ranking functions from domain-specific dictionaries or libraries. Automatic extraction of functions can be used in parallel with feedback from a domain expert.

Acknowledgements. This work was funded by Smart Services Cooperative Research Centre. We would like to thank our industry partners for the new car data set and their domain expertise.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th Conference on Very Large Data Bases, pp. 478–499 (1994)
2. Bettman, J.R., Luce, M.F., Payne, J.W.: Constructive Consumer Choice Processes. *Journal of Consumer Research* 25, 187–217 (1998)
3. Bilgic, M., Mooney, R.J.: Explaining Recommendations: Satisfaction vs. Promotion. In: Proceedings of the IUI 2005 Workshop: Beyond Personalization (2005)
4. Burke, R.D., Hammond, K.J., Young, B.C.: The FindMe Approach to Assisted Browsing. *IEEE Expert* 12(4), 32–40 (1997)
5. Fagin, R., Lotem, A., Naor, M.: Optimal Aggregation Algorithms for Middleware. *Journal of Computer and System Sciences* 66, 614–656 (2003)
6. Felix, D., Niederberger, C., Steiger, P., Stolze, M.: Feature-Oriented vs. Needs-Oriented Product Access for Non-Expert Online Shoppers. In: Proceedings of the IFIP Conference on Towards the E-Society: E-Commerce, E-Business, E-Government, pp. 399–406 (2001)
7. Hammond, K.J., Burke, R., Schmitt, K.: A Case-Based Approach to Knowledge Navigation. In: Proceedings of the AAAI 1994 Workshop on Knowledge Discovery in Databases, pp. 383–393 (1994)
8. Holman, R.H.: A Values and Lifestyles Perspective on Human Behavior. In: Pitts Jr., R.E., Woodside, A.G. (eds.) *Personal Values and Consumer Psychology*. Lexington Books, Lexington (1984)
9. Keeney, R.L., Raiffa, H.: *Decisions with Multiple Objectives: Preferences and Value Trade-offs*. John Wiley & Sons, New York (1976)
10. Kelly, J.P., Bridge, D.: Enhancing the Diversity of Conversational Collaborative Recommendations: A Comparison. *Artificial Intelligence Review* 25, 79–95 (2006)
11. McCarthy, K., McGinty, L., Smyth, B., Reilly, J.: On the Evaluation of Dynamic Critiquing: A Large-Scale User Study. In: Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI 2005), pp. 535–540 (2005)
12. McCarthy, K., Reilly, J., Smyth, B., McGinty, L.: Generating Diverse Compound Critiques. *Artificial Intelligence Review* 24, 339–357 (2005)
13. McSherry, D.: Explanation in Recommender Systems. *Artificial Intelligence Review* 24, 179–197 (2005)
14. Stolze, M., Ströbel, M.: Recommending as Personalized Teaching: Towards Credible Needs-Based eCommerce Recommender Systems. In: Karat, C.M., Blom, J., Karat, J. (eds.) *Designing Personalized User Experiences in eCommerce*. Kluwer Academic Publishers, Dordrecht (2004)
15. Vinson, D.E., Scott, J.E., Lamont, L.M.: The Role of Personal Values in Marketing and Consumer Behavior. *Journal of Marketing* 41(2), 44–50 (1977)
16. Zablocki, B.D., Kanter, R.M.: The Differentiation of Life-Styles. *Annual Review of Sociology* 2, 269–298 (1976)
17. Zhang, J., Pu, P.: A Comparative Study of Compound Critique Generation in Conversational Recommender Systems. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) *AH 2006*. LNCS, vol. 4018, pp. 234–243. Springer, Heidelberg (2006)

Extracting Communities in Networks Based on Functional Properties of Nodes

Takayasu Fushimi¹, Kazumi Saito¹, and Kazuhiro Kazama²

¹ Graduate School of Management and Information of Innovation,
University of Shizuoka 52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
{j11507,k-saito}@u-shizuoka-ken.ac.jp

² Nippon Telegraph and Telephone Corporation, Network Innovation Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo 180-8585
kazama@ingrid.org

Abstract. We address the problem of extracting the groups of functionally similar nodes from a network. As functional properties of nodes, we focus on hierarchical levels, relative locations and/or roles with respect to the other nodes. For this problem, we propose a novel method for extracting functional communities from a given network. In our experiments using several types of synthetic and real networks, we evaluate the characteristics of functional communities extracted by our proposed method. From our experimental results, we confirmed that our method can extract functional communities, each of which consists of nodes with functionally similar properties, and these communities are substantially different from those obtained by the Newman clustering method.

1 Introduction

Finding groups of functionally similar nodes in a social or information network can be a quite important research topic in various fields ranging from computer science to sociology. Hereafter, such a node group is simply referred to as a functional community. In fact, each node which typically corresponds to a person in a social network may have a wide variety of functional properties such as status, ranks, roles, and so forth. However, conventional methods for extracting communities as densely connected subnetworks, which include the Newman clustering method based on a modularity measure [1] and so forth cannot directly deal with such functional properties. Evidently, conventional notions of densely connected subnetworks such as k -core [2] and k -clique [3] cannot work for this purpose. Namely, it is naturally anticipated that these existing methods have an intrinsic limitation for extracting functional communities.

In this study, as typical functional properties of nodes, we especially focus on hierarchical levels, relative locations and/or roles with respect to the other nodes. This implies that there exist some functionally similar nodes even if they are not directly connected with each other. For instance, in case of a network of employees relationships in a company, we can naturally assume it to have a hierarchical property, where the top node corresponds to the president, and in

turn, the successive levels of nodes correspond to managers, section leaders, and so on. For example, our objective might be to extract a group of section leaders as a functional community in the network, even though they may not have direct connections with each other. Here we should emphasize that extracting these types of communities can be a quite tough problem for the conventional community extraction methods because these existing methods mainly focus on link densities among each subnetwork and between subnetworks.

In this paper, we propose a novel method for extracting functional communities from a given network. This algorithm consists of two steps: the method first assigns a feature vector to each node, which is assumed to be some functional properties, by using calculation steps of PageRank scores [4] for nodes from an initial score vector. Then, in a case that the supposed number of functional communities is K , the method divides all the node into K groups by using the K -medians clustering method based on the cosine similarity between a pair of the feature vectors. In our experiments using several types of synthetic and real networks, we evaluate the characteristics of functional communities extracted by our proposed method. To this end, we utilize the visualization result of each network where each functional community is indicated by a different color marker, and these results are contrasted to those obtained by the Newman clustering method [1].

2 Component Algorithms

2.1 PageRank Revisited

For a given Web hyperlink network (directed graph), we identify each node with a unique integer from 1 to $|V|$. Then we can define the adjacency matrix $\mathbf{A} \in \{0, 1\}^{|V| \times |V|}$ by setting $a(u, v) = 1$ if $(u, v) \in E$; otherwise $a(u, v) = 0$. A node can be self-looped, in which case $a(u, u) = 1$. For each node $v \in V$, let $F(v)$ and $B(v)$ denote the set of child nodes of v and the set of parent nodes of v , respectively, $F(v) = \{w \in V; (v, w) \in E\}$, $B(v) = \{u \in V; (u, v) \in E\}$. Note that $v \in F(v)$ and $v \in B(v)$ for node v with a self-loop.

Then we can consider the row-stochastic transition matrix \mathbf{P} , each element of which is defined by $p(u, v) = a(u, v)/|F(u)|$ if $|F(u)| > 0$; otherwise $p(u, v) = z(v)$, where \mathbf{z} is some probability distribution over nodes, i.e., $z(v) \geq 0$ and $\sum_{v \in V} z(v) = 1$. This model means that from dangling Web pages without out-links ($F(u) = \emptyset$), a random surfer jumps to page v with probability $z(v)$. The vector \mathbf{z} is referred to as a personalized vector because we can define \mathbf{z} according to user's preference.

Let \mathbf{y} denote a vector representing PageRank scores over nodes, where $y(v) \geq 0$ and $\sum_{v \in V} y(v) = 1$. Then using an iteration-step parameter s , PageRank vector \mathbf{y} is defined as a limiting solution of the following iterative process,

$$\mathbf{y}_s^T = \mathbf{y}_{s-1}^T ((1 - \alpha)\mathbf{P} + \alpha\mathbf{e}\mathbf{z}^T) = (1 - \alpha)\mathbf{y}_{s-1}^T \mathbf{P} + \alpha\mathbf{z}^T, \tag{1}$$

where \mathbf{a}^T stands for a transposed vector of \mathbf{a} and $\mathbf{e} = (1, \dots, 1)^T$. In the Equation (II), α is referred to as the uniform jump probability. This model means that with the probability α , a random surfer also jumps to some page according to the probability distribution \mathbf{z} . The matrix $((1 - \alpha)\mathbf{P} + \alpha\mathbf{e}\mathbf{z}^T)$ is referred to as a Google matrix. The standard PageRank method calculates its solution by directly iterating Equation (II), after initializing \mathbf{y}_0 adequately. One measure to evaluate its convergence is defined by $\|\mathbf{y}_s - \mathbf{y}_{s-1}\|_{L1} \equiv \sum_{v \in V} |y_s(v) - y_{s-1}(v)|$. Note that any initial vector \mathbf{y}_0 can give almost the same PageRank scores if it makes equation of convergence evaluation almost zero because the unique solution of Equation (II) is guaranteed.

2.2 K -medians Revisited

For a given set of objects (or nodes), denoted by $V = \{v, w, \dots\}$, the K -medians method first selects K representative objects $\mathcal{R} \subset V$ according to this objective function $f(\mathcal{R}) = \sum_{v \in V} \max_{r \in \mathcal{R}} \rho(v, r)$ to be maximized. Here $\rho(v, r)$ stands for a similarity measure between a pair of objects, v and r . Then, from the obtained K representative objects $\mathcal{R} = \{r_1, \dots, r_K\}$, the method determines the K clusters, $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$, by using this formula $\mathcal{C}_k = \{v \in V; r_k = \arg \max_{r \in \mathcal{R}} \rho(v, r)\}$. Finally, the method outputs $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$ as the result.

In order to maximize the objective function with respect to \mathcal{R} , due to simplicity we employ a greedy algorithm. Here note that in virtue of the submodularity of the objective function, we can obtain a unique greedy solution whose worst case quality is guaranteed [5].

3 Proposed Method

In this section, we describe our proposed method for extracting functional communities. Our method utilizes the PageRank score vectors at each iteration step s , i.e., $\{\mathbf{y}_1, \dots, \mathbf{y}_S\}$. Here, S stands for the final step when the PageRank algorithm converges. Then, for each node $v \in V$, we can consider an S -dimensional vector defined by $\mathbf{x}_v = (y_1(v), \dots, y_S(v))^T$, where $y_s(v)$ means the PageRank score of node v at iteration step s . In our method, \mathbf{x}_v is regarded as a functional property vector of node v .

Here we note a reason why we employ the vector described above. Basically, we assume that functional properties of nodes, such as hierarchical levels, relative locations and/or roles with respect to the other nodes are embedded into the network structure. On the other hand, the PageRank scores at each iteration step also reflect the network structure. Therefore, as an approximation, we can consider that functional properties are also represented by the vector \mathbf{x}_v .

In order to divide all nodes into the K groups, our method employs the K -medians algorithm described in the previous section. To this end, we need to define an adequate similarity $\rho(u, v)$ between the nodes u and v . In our proposed method, for each pair of functional property vectors, we employ this cosine similarity $\rho(u, v) = \frac{\mathbf{x}_u^T \mathbf{x}_v}{\|\mathbf{x}_u\| \|\mathbf{x}_v\|}$, where $\|\mathbf{x}_v\|$ stands for the standard L2 norm.

For a given network $G = (V, E)$ and the number K of functional communities, we summarize our proposed algorithm below.

1. Calculate the PageRank score vectors at each time step $\{\mathbf{y}_1, \dots, \mathbf{y}_S\}$;
2. Construct the functional property vector \mathbf{x}_v for each node $v \in V$;
3. Calculate the cosine similarity $\rho(u, v)$ of \mathbf{x}_u and \mathbf{x}_v for all node pair;
4. Divide all nodes into K clusters according to the similarity $\rho(u, v)$ by the K -medians method;
5. Output functional communities $\{\mathcal{C}_1, \dots, \mathcal{C}_K\}$;

4 Experimental Evaluation

In this section, using several types of synthetic and real networks, we experimentally evaluate the characteristics of functional communities extracted by our proposed method. For this purpose, we utilize the visualization result of each network where each functional community is indicated by a different color marker, and these results are contrasted to those obtained by the Newman method [1].

4.1 Network Data

We describe a detail of four networks used in our experiments.

First one is a synthetic network with a hierarchical property, just like an employee relationships or Web hyperlinks network. In this hierarchical network, we can assume two types of nodes, central (or high status) and peripheral (or low status) nodes. As shown later in Fig. 1, in terms of its basic network statistics, the central nodes are characterized by relatively high degree and low clustering coefficients, while the peripheral nodes by relatively low degree and high clustering coefficients. We generated this network according to Ravasz et.al. [6]. Hereafter, this network is referred to as Hierarchical network.

Second one is a two dimensional-grid network implemented as a set of 10×10 lattice points. Evidently, as shown later in Fig. 2, because of the regular structure, dividing this network into several portions does not make sense in the aspects of standard community extraction. Whereas, we can consider a functional property in terms of relative locations to other nodes, i.e., the relative closeness to the center position. Hereafter, this network is referred to as Lattice network.

Third one is a social network of people belonging to a karate circle, which has been widely used as a benchmark network. As shown later in Fig. 3, we see a number of hub nodes, which play an important role to connect other nodes. Namely, we can assume that some group of nodes has a similar role with respect to the other nodes. Hereafter, this network is referred to as Karate network [7].

Forth one is a hyperlink network of a Japanese university Web site, where we obtained this network by crawling the Web site as of Aug. 2010. As shown later in Fig. 4, there exist a number of unique characteristics in this network. Namely, we can assume that some group of Web pages has a similar topic specificity level. Hereafter, this network is referred to as Hosei network [4].

¹ The site name and its address are "Faculty of Computer and Information Sciences, Hosei University" and <http://cis.k.hosei.ac.jp/>, respectively.

4.2 Experimental Settings

We first explain the settings of our proposed algorithm. In order to calculate the PageRank score vectors, we set the initialized vector to $\mathbf{y}_0 = (1/|V|, \dots, 1/|V|)^T$, and the convergence criterion is implemented as $\|\mathbf{y}_s - \mathbf{y}_{s-1}\|_{L1} < 10^{-12}$. The number K of communities to be extracted is changed from $K = 2$ to 10.

As mentioned earlier, we attempt to clarify the characteristics of the functional communities extracted by our method, in comparison to standard communities extracted by the Newman clustering method [1]. Hereafter, such a standard community is simply referred to as a Newman community. The Newman method is basically designed to obtain densely connected subnetworks by maximizing a modularity measure.

Finally, we describe methods to visualize each network. In Hierarchical network, we employ nodes' positions as displayed by Ravasz et.al. [6]. As for Lattice network, we can regularly assign the positions to nodes. In cases of Karate and Hosei networks, the cross-entropy embedding method [8] is used to determine the positions of nodes.

4.3 Experimental Results

We show the experimental results of Hierarchical network at $K = 5$ in Fig. 1. Here note that this network consists of five portions of densely connected subnetworks, as observed in Fig. 1. Thus, as an example, we selected this number, $K = 5$. As expected, from Fig. 1(a), we see that our method could extract reasonable functional communities, each of which consists of nodes with the similar hierarchical levels, just like employees with same position such as the president, managers, or general staffs. On the other hand, from Fig. 1(b), we see that the Newman method extracted standard communities, each of which is characterized as a densely connected subnetwork, just like employees belonging to the same department or section.

We show the experimental results of Lattice network at $K = 3$ in Fig. 2. Here recall that in the aspects of standard community extraction, dividing this network into several portions does not make sense. Thus, as an example, we selected this relatively small number, $K = 3$. From Fig. 2(a), we see that our method could extract reasonable functional communities, each of which consists of nodes with the similar relative locations, i.e., the relative closeness to the center position. On the other hand, as shown in Fig. 2(b), we can hardly make sense to the communities extracted by the Newman method.

We show the experimental results of Karate network at $K = 2$ in Fig. 3. Here note that this network consists of two portions of densely connected subnetworks, as observed in Fig. 3. Thus, as an example, we selected this number, $K = 2$. From Fig. 3(a), we see that our method could extract reasonable functional communities, each of which consists of nodes with different roles with respect to the other nodes, i.e., groups of hub nodes and the other nodes. On the other hand, from Fig. 3(b), we see that the Newman method extracted standard communities, each of which is characterized as a densely connected subnetwork.

We show the experimental results of Hoseni network at $K = 10$ in Fig. 4. Here note that this network consists of several portions of characteristically connected subnetworks, as observed in Fig. 4. Thus, as an example, we selected this relatively large number, $K = 10$. From Fig. 4(a), we see that our method extracted several communities, each of which consists of nodes with similar connection patterns. In order to more closely investigate these extracted communities, we focused on a particular community indicated by small blue squares surrounding with large transparent squares in Fig. 4(a). From our examination of these Web pages belonging to this community, we realized that these Web pages correspond to annual reports of each year produced by faculty members. Namely, it is assumed that these Web pages in this community have a similar topic specificity level. Thus, we can consider that our method could extract a piece of reasonable functional communities in the sense described above. On the other hand, from Fig. 4(b), we see that the Newman method divided the functional community focused above into several communities.

From our experimental results using these networks with different characteristics, we confirmed that our method could extract functional communities, each of which consists of nodes with similar functional properties such as hierarchical levels, relative locations and/or roles with respect to the other nodes. These results indicate that our method is promising for tasks of extracting functional communities with these properties. On the other hand, the Newman method extracted standard communities characterized by densely connected subnetworks.

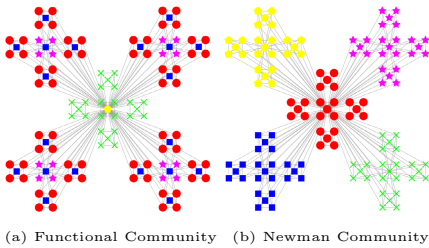


Fig. 1. Hierarchical Network ($K = 5$)

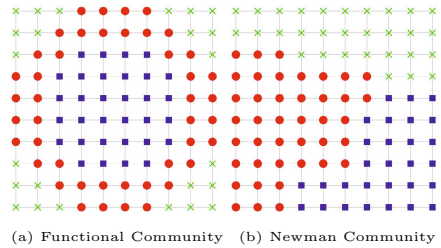


Fig. 2. Lattice Network ($K = 3$)

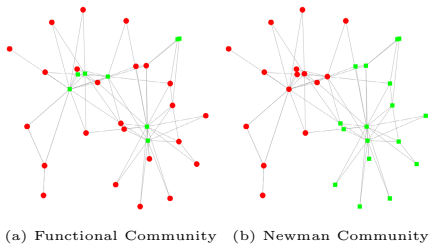


Fig. 3. Karate Network ($K = 2$)

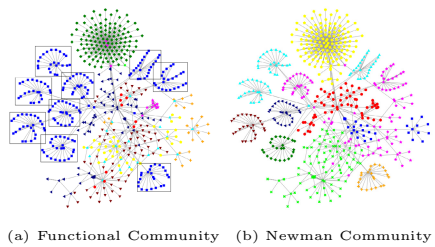


Fig. 4. Hoseni Network ($K = 10$)

From these results, we see that these functional communities extracted by our method are substantially different from those obtained by the Newman method.

5 Conclusion

We addressed the problem of extracting the groups of functionally similar nodes from a network. In this paper, such a node group was simply referred to as a functional community. As functional properties of nodes, we focused on hierarchical levels, relative locations and/or roles with respect to the other nodes, and proposed a novel method for extracting functional communities from a given network. In our experiments using several types of synthetic and real networks, we evaluated the characteristics of functional communities extracted by our proposed method. From our experimental results, we confirmed that our method could extract functional communities, each of which consists of nodes with functionally similar properties, and these communities were substantially different from those obtained by the Newman clustering method. In future, we plan to evaluate our method using various networks.

Acknowledgments. This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-11-4111, JSPS Grant-in-Aid for Scientific Research (C) (No. 23500128), and NTT Network Innovation Laboratories.

References

1. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. *Physical Review E* 69(6), 066133 (2004)
2. Seidman, S.B.: Network structure and minimum degree. *Social Networks* 5(3), 269–287 (1983)
3. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818 (2005)
4. Langville, A.N., Meyer, C.D.: Deeper inside pagerank. *Internet Mathematics* 1(3), 335–380 (2004)
5. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions. *Mathematical Programming* 14, 265–294 (1978)
6. Ravasz, E., Barabási, A.L.: Hierarchical organization in complex networks. *Physical Review E* 67(2), 026112 (2003)
7. Zachary, W.W.: An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33, 452–473 (1977)
8. Yamada, T., Saito, K., Ueda, N.: Cross-entropy directed embedding of network data. In: *Proceedings of the 20th International Conference on Machine Learning (ICML 2003)*, pp. 832–839 (2003)

Revealing and Trending the Social Knowledge Studies

Maria R. Lee¹ and Tsung Teng Chen²

¹Department of Information Technology and Management, Shih Chien University, Taiwan
maria.lee@mail.usc.edu.tw

²Graduate Institute of Information Management, National Taipei University, Taipei, Taiwan
misttc@mail.ntpu.edu.tw

Abstract. In recent years, the study of social knowledge has become so extensive that ample amounts of literature are now available. However, it is a challenging task to search, review and analyze such vast literatures effectively. The primary goal of the paper is to provide an intellectual structure to facilitate the understanding of the rapidly evolving field of social knowledge. This article explores and maps the research themes and trends of social knowledge studies. Document co-citation analysis, pathfinder network and strategic diagram techniques are applied to provide a dynamic view of social knowledge research themes. The top four themes that emerged in the social knowledge study are: situated learning, social choice, social capital, and perception behavior. This study may be used by novice researchers to gain useful insights about the themes of this emerging and fast-growing field.

Keywords: Social Knowledge, Research Trends, Literature Review, Intellectual Structure.

1 Introduction

Social knowledge is an exciting area to investigate due to its dynamic and exploding growth. It is interesting to foresee the impact in the development of this field. There is a wealth of literature now available in the field. It is challenging to search, review, and analyze such vast literatures effectively. In addition, it is difficult to provide a systematic and structured approach to explore and prepare for the development of research in the future [1].

The abilities of visualizing the growth of the domain knowledge and revealing the evolution of research themes are two essential points in the studies of a new topic domain. Visualization can facilitate the understanding of the structures of a collection of documents that are related to each other by links, such as citations in formal publications. The primary goal of the paper is to provide an intellectual structure to facilitate the understanding of the rapidly evolving field of social knowledge. The objective of this study is to explore and map the research themes and trends of social knowledge studies. We aim to map the intellectual structure of the social knowledge and to identify the inter-relationship between themes, which enable better understanding the development of the field.

The paper is organized as follows. First, we review relevant literature on social knowledge. Second, we briefly describe the research methodology applied. Third, we present the results of social knowledge analysis. Fourth, we conclude with a general discussion on the results and suggestions regarding future research.

2 Social Knowledge Definition

What is social knowledge? Different perspectives or context have been used to define social knowledge. [2] discovers the roots of social knowledge in two distinct sorts: spatial and social knowledge. Spatial is based on sense perceptions, gives rise to exact science whereas social is based on the "mental-social complex", a working agreement and the accumulation of social knowledge.

From a philosophical point of view, [3] studies on the development of social knowledge, focuses on morality and convention. [4] develops the augments for human science as social construction. The descriptions of human action can neither be derived from nor corrected by scientific observation. [4] provides a bold interdisciplinary challenge to traditional views, thus clearing the way for significant alterations in scientific practice. [5] argues that objectivity is most appropriately regarded as a feature of the scientific community rather than as a characteristic of individuals (because individuals are susceptible to such a wide variety of idiosyncratic influences).

Scientists, such as evolutionary biologists, pursue the social knowledge idea as solving social problems has driven the evolution of intelligence, not only in humans but also in other big-brained species. [6] presents the so-called social intelligence hypothesis which resulted in the observations of primates (geladas) that exhibit social knowledge of individuals they remember in their surroundings.

From medical sociology perspective, [7] wrote that the relationships between medical research agencies and the sociological community do not have a sophisticated framework. [8] reveals the medial sociological debate by considering the recent development in sociological of the body, the concept of risk in society and process of globalization.

From the Artificial Intelligence (AI) perspective, [9] studies some of the social aspects of knowledge and action relevant to thinking in AI, and in particular the basic experience of multiple perspectives and integrating different kinds of local knowledge. [10] addresses the problems of efficient representation, maintenance and exploration of social knowledge enabling task decomposition, organization of negotiations, responsibility delegation and other ways of agents' social reasoning. The social knowledge is kept separated from both the problem solving knowledge and the agents' specific internal intelligence, and organized and administered in the acquaintance models located in the agents' wrappers.

Recent, social knowledge is growing at an explosive rate, a number of social aspects to be considered. [11] harvests social knowledge from folksonomies, collaborative tagging systems, or have the potential of becoming technological infrastructure to support knowledge management activities in an organization or a society.

3 Methodology

We aim to reveal and identify the subfields characterized by the intellectual nature of specialties and the main trends within social knowledge. We apply document co-citation analysis and pathfinder network (PFNET) [12] to meet our goals. A factor analysis technique is applied as a data reduction and structure detection method. The Pearson correlation coefficients between items (papers) are used as the basis for Pathfinder Network (PFNET) scaling. The major research themes and their interrelationships can thus be easily identified via the intellectual structure map. The proposed methodologies have been applied to many research fields such as knowledge management, ubiquitous computing, etc. [13]. The proposed research provides a powerful tool for understanding the epistemology of a field as it evolves.

An online citation database, Microsoft Academic Search (MAS) [14] is used to construct a full citation graph. The reasons for applying the MAS database is not only it is a free search engine for academic research papers and resources, but also the database contains broad domains with citation contexts [13]. The MAS database includes the bibliographic information (metadata) for research articles published in journals, conference's proceedings, and books. The key phrases "social knowledge" is used to query the database. The documents retrieved by the query are then used as the initial seed set to search for papers that cite them or for papers that are cited by them [15].

4 Results

Following the steps described in [15], we review, analyze and synthesize the papers retrieved. Table 1 shows a list of social knowledge studies and themes. Factors with variance over one are listed in the table, which includes factor numbers, research themes and percentage of variance. We will briefly explain the content of factor 1 to 10.

Factor 1 represents research in social behavior, social anxiety or social interaction. Researcher studies the role of the amygdala in animals that display a level of social sophistication, which can help in understanding human social behavior and social anxiety. Social anxiety is an enduring and lasting condition that is characterized by a fear of social interaction. Emotional expressivity is the tendency to express emotions nonverbally, such as facial expressions, tone of voice and body languages, etc.

Factor 2 represents studies on social computing. Social computing concerns the intersection between social behavior and computational systems. It is based on creating or recreating social conventions and social contexts through the use of software and technology.

Factor 3 represents studies on social cognition. Gallese *et al* provide a unifying neural hypothesis on how individuals understand the actions and emotions of others. Moral cognitive neuroscience focuses on the neural basis of uniquely human, forms of social cognition and behavior. J. Moll propose a cognitive neuroscience view of how cultural and context-dependent knowledge, semantic social knowledge, and motivational states can be integrated to explain complex aspects of human moral cognition.

Factor 4 represents studies on situated learning/community of practice. A community of practice (CoP) is a group of people who share a craft and/or a profession. Situated learning is a model of learning in a community of practice that takes place in the same context in which it is applied.

Factor 5 represents studies on social choice. Social choice theory is a theoretical framework for measuring individual interests, values, or welfares as an aggregate towards a *collective decision*. K. Arrow present a theorem of social ethics and voting theory with an economic flavor.

Table 1. List of Major Social Knowledge Studies and Themes

Factor 1. Social Behavior (10.758)	Factor 2. Social Computing (7.667)
Factor 3. Social Cognition (7.667)	Factor 4. Situated Learning (5.757)
Factor 5. Social Choice (5.635)	Factor 6. Social Captial (4799)
Factor 7. Perception Behavior (4.01)	Factor 8. Congestion Game (3.321)
Factor 9. Intelligent Agent (3.253)	Factor 10. Theory of Mind (3.146)
Factor 11. Knowledge Level (2.911)	Factor 12. Implicit Association Test (IAT) (2.437)
Factor 13. Face Perception (2.124)	Factor 14. Social Judgment (1.842)
Factor 15. Swarm Behavior (1.776)	Factor 16. Social Schema (1.596)
Factor 17. Social Network Analysis (1.443)	Factor 18. --- (1.435)
Factor 19. Social Software App. (1.346)	Factor 20. Natural Selection (1.239)

Factor 6 represents studies on social capital. Social capital refers to the value of social relations and the role of cooperation and confidence to get collective or economic results. Nahapiet and Ghoshal incorporates social capital, intellectual capital and organizational advantage and examines the role of social capital in the creation of intellectual capital. They suggested that social capital should be considered in terms of three clusters: structural, relational, and cognitive. J. Pfeffer presents sixteen practices of competitive advantage through people.

Factor 7 represents studies on perception behavior. Bruner presents perception readiness and summarizes that perception depends not only on what's out there but also on the momentary activation of concepts such as a person's needs, goals or attitudes. Chartrand and Bargh study the chameleon effect, which refers to non-conscious mimicry of the postures, mannerisms, facial expressions, and other behaviors of a person's interaction in their current social environment.

Factor 8 represents studies on congestion games. R. Rosenthal proposes congestion games, which define players and resources, where the payoff of each player depends on the resources he/she chooses and the number of players choosing the same resource. Beier *et al* investigates algorithmic questions concerning a basic microeconomic congestion game in which there is a single provider that offers a service to a set of potential customers. Koutsopias *et al* studies N numbers of selfish agents or players, each are having a load, who want to place their loads to one of two bins. The agents are having partial knowledge, which is an incomplete picture of the world.

Factor 9 represents studies on intelligent agents. I. Havel studies the theory of consciousness by applying Artificial Intelligence (AI) and Connectionism (Neural Network Modeling) to our understanding of the human mind and brain. Kumar *et al* proposes the adaptive agent architecture (AAA) for achieving fault-tolerance using persistent broker teams. The AAA multi-agent system can maintain a specified number of functional brokers in the system despite broker failures, thus effectively becoming a self-healing system.

Factor 10 represents studies on theory of mind. Theory of mind is the ability to attribute mental states—beliefs, intents, desires and knowledge, etc. C.M. Heyes studies theory of mind in nonhuman primates, which is different from the research on the development of theory of mind in childhood. Vogeley *et al* addresses the issue of whether taking the self-perspective or modeling the mind of someone else employ the same or different neural mechanisms.

4.1 Visualization of Intellectual Structure in Social Knowledge

When the factor analysis is applied, Pearson correlation coefficients between items (papers) are also calculated. Pathfinder network (PFNET) scaling is used to derive the

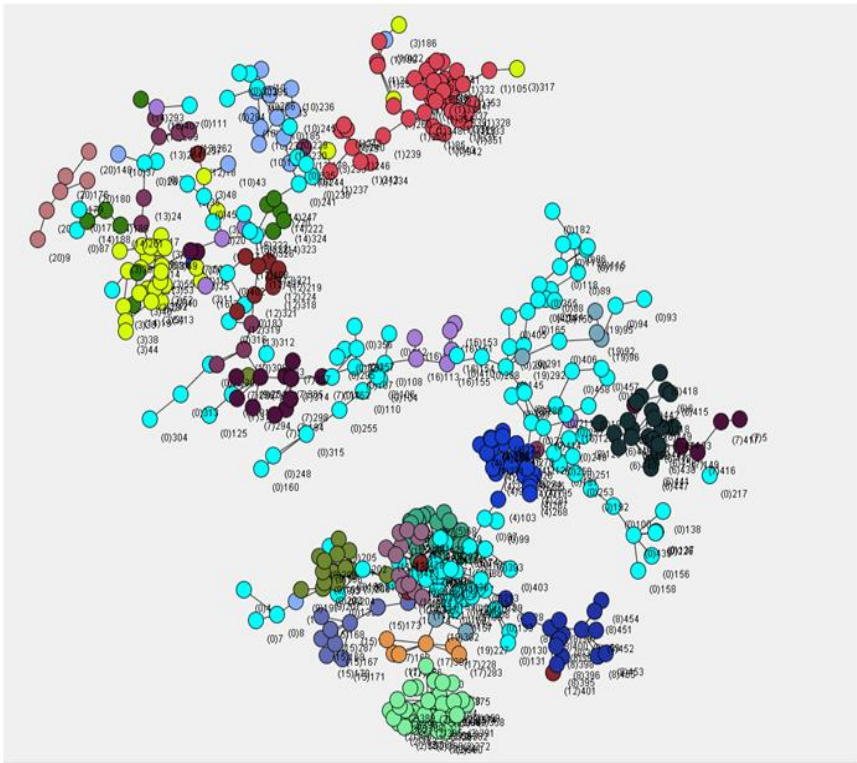


Fig. 1. The PFNET graph of social knowledge

essential relationships from the correlation matrix [16]. The value of the Pearson correlation coefficient falls between -1 and 1. Two items (papers) correlate completely when the coefficient is approaching 1. Items that are closely related representing the fact that the papers are highly correlated and should be placed closely spatially. The distance between items is inversely proportional to the correlation coefficient, which depicts less correlated items apart and highly correlated items close together.

The nodes located close to the center of a PFNET graph represent papers contributing to a fundamental concept and in the mainstream of a research domain. Figure 1 shows PFNET scaling of social knowledge drawn from Microsoft Academic Search. Articles under the same factor are painted with the same color. The number in the parentheses is the factor number that the article belongs to. Cyan nodes with (0) represent articles that are not assigned to any factor.

4.2 Interrelationships between Research Themes in Social Knowledge

The PFNET graph of social knowledge shown in figure 2 facilitates the identification of interrelationships between research themes in social knowledge. To explicitly visualize the relationships between themes in the PFNET graph, figure 2 is derived from figure 1 by consolidating nodes in the same factor into a node block to highlight the intellectual structure relationships between themes. The nodes located close to the center of the graph represent papers that contribute to a core concept. The lines shown in figure 2 indicate a connection relationship. The number inside the box refers to the factor number described in table 1.

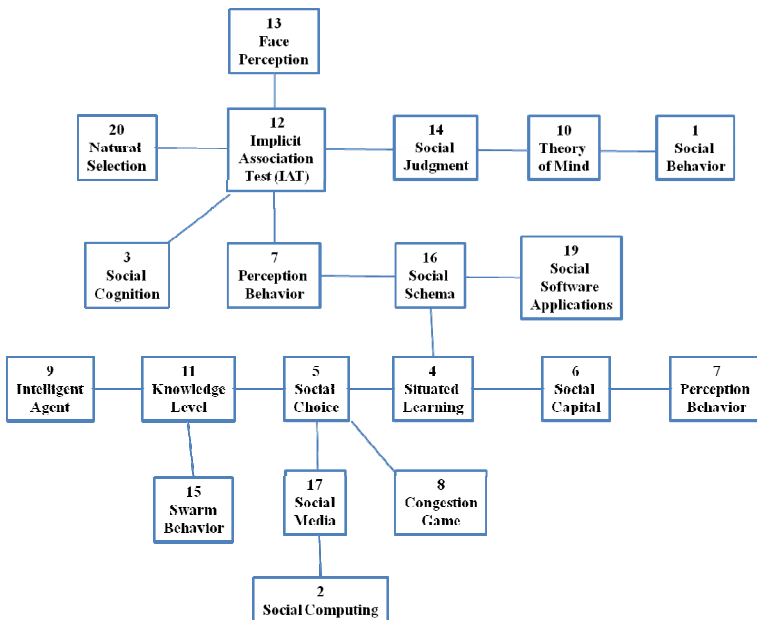


Fig. 2. The intellectual structure relationships between Social Knowledge themes

For example, the inter-relationships between research themes in figure 2, knowledge level (factor 11) has inter-relationships with intelligent agent (factor 9), social behavior (factor 1), and social choice (factor 5). This implies those themes are more or less rationalized agent's behavior.

5 Conclusion

Although this review does not claim to be exhaustive, it does provide a reasonable amount of insight into social knowledge related researches. The motivation behind our investigation is to adopt a past-future orientation by analyzing the past to prepare for the future [Webster and Watson 2002]. We have presented an intellectual structure based literature review and analysis of social knowledge studies. The results presented in this paper have several important implications:

Social knowledge research is a cross-disciplined study that spanned across a broad spectrum of disciplines. Apart from computer science, social science, business, others include such as language and social knowledge, social education, etc. There is no doubt that social knowledge research will increase significantly in future. Social ethics or cultural issues have not been explored much in social knowledge research.

Top four emerged researches in the social knowledge study are: situated learning, social choice, social capital, and perception behavior. There are three implications for future research emergent from this analysis and the conclusions drawn from it. There is a need to revisit the analysis to determine if the research themes in social knowledge continue to evolve in a seemingly new direction. Although the study on social behavior, social computing, and social cognition have high variances, they are not main research stream topics yet. The evolution of these research themes need to be further investigated. The dependency of the interrelation intellectual structure shown in figure 2 can be further analyzed.

Acknowledgments. This work is supported in part by the Shih Chien University (USC grant 100-08-01004).

References

1. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: Writing a literature review. *Management Information Systems Quarterly* 26, 8–13 (2002)
2. Cooley, C.H.: The roots of social knowledge. *American Journal of Sociology*, 59–79 (1926)
3. Turiel, E.: The development of social knowledge: Morality and convention. Cambridge Univ. Press (1983)
4. Gergen, K.J.: *Toward transformation in social knowledge*. Sage Thousand Oaks (1994)
5. Longino, H.E.: *Science as social knowledge: values and objectivity in scientific inquiry*. Princeton Univ. Press (1990)
6. Bergman, T.J.: Experimental evidence for limited vocal recognition in a wild primate: implications for the social complexity hypothesis. *Proceedings of the Royal Society B: Biological Sciences* 277(1696), 3045–3053 (2010)

7. Turner, B.S.: Professions, knowledge and power. *Medical Power and Social Knowledge*, 131–156 (1987)
8. Turner, B.S., Samson, C.: *Medical power and social knowledge*. Sage Publications (1995)
9. Gasser, L.: *Social knowledge and social action: Heterogeneity in practice*. Lawence Erlbaum Associate (1993)
10. Marík, V., Pechoucek, M., Štěpánková, O.: Social knowledge in multi-agent systems. *Multi-Agent Systems and Applications*, 211–245 (2006)
11. Wu, H., Zubair, M., Maly, K.: Harvesting social knowledge from folksonomies. *ACM* (2006)
12. Schvaneveldt, R.W.: *Pathfinder associative networks: studies in knowledge organizations*. Ablex series in computer science, Norwood, p. 315 (1990)
13. Lee, M.R., Chen, T.T.: Revealing research themes and trends in knowledge management: From 1995 to 2010. *Knowledge-Based Systems*, 47–58 (2012)
14. Carlson, S.: Challenging Google, Microsoft Unveils a Search Tool for Scholarly Articles. *Chronicle of Higher Education* 52, 1 (2006)
15. Chen, T.T.: The development and empirical study of a literature review aiding system. *Scientometrics*, 1–12 (2012)
16. Chen, C., Paul, R.J.: Visualizing a Knowledge Domain's Intellectual Structure. *Computer* 34(3), 65–71 (2001)

Workflow Knowledge Sharing through Social Networks

Peter Busch and Amireh Amirmazaheri

Department of Computing, Macquarie University, NSW 2109, Australia
peter.busch@mq.edu.au

Abstract. Social Network Analysis provides the researcher a means to interpret relationships and in turn the flow of information. What has been less well researched is how such analysis may be used to interpret workflows or business processes and potentially lead to their improvement. This research-in-progress paper explores the comparatively fine-grained social interactions of employees in a company in Sydney. One particular workplace process, namely the use of Oracle Primavera P6 is examined. Following such an approach, attention is turned to exploring the working relationships of employees through Social Network Analysis to determine the closeness of fit of employees against the workflows officially established by management. Such a combination of techniques extends the toolkit of methodological approaches in this space, and from a more practical viewpoint permits the process improvement workflow manager to consider how well matched employees are to their workflows.

Keywords: Social Network Analysis, Business Process Management, Business Process Improvement.

1 Introduction

Knowledge Acquisition (KA) in the broadest sense of the term encompasses approaches that exist in a number of other disciplines. Social Network Analysis (SNA) is one such example that is used in other areas and KA does occasionally use SNA to explore the interrelationships of individuals and how knowledge can be transferred. Examples of such overlapping research include semantic analysis of learner discourse in online forums [10] or the use of SNA in the creation of social capital [7]. Another area falling under the purview of KA in Business Process Management (BPM), is how social networks influence workflow patterns and are in turn influenced by them through the question: “how extensive is the degree of overlap between an employee’s work processes and their organisational social network?” While management may have fixed ideas on how employees conduct their work, in fact the peer networks of employees may produce alternate workflow patterns compared to configurations illustrated in structure charts or workflow models [11]. This short paper provides an exploratory case study of a BPM process in a Sydney based company, examined through the ‘lens’ of SNA.

2 Background

At an underlying level organisations operate successfully through relatively efficient communication flows amongst employees. Early research work at the Xerox Palo Alto Research Center (PARC) led in some ways to the founding of Knowledge Management as a discipline, with ‘war stories’ amongst photocopier repairers setting the scene for improvements in work processes [2]. In more recent times PARC has been at the forefront managing practical or workplace forms of knowledge. The benefits for an organisation better managing its work processes are many and include improvements to organisational learning [3]; gaining improved clarity in intra-organisational workflows [8], reduced time wastage amongst knowledge workers [5]; and in MNCs, formalising the knowledge sharing process as a means of reducing employee cultural hurdles [1].

One advantage of applying SNA specifically to the business is to provide more seamless customer service, for customer calls and emails can be routed more effectively with an internal social network in place. Other studies [1; 4] have comprehensively examined the flows of soft knowledge through organizations and how they can be mapped with SNA to determine the likelihood of soft knowledge transfer or perhaps knowledge ‘bottlenecking’. Through injecting already existing workflow logs in to SNA we build a more complete organisational workflow with regard to how work is actually being undertaken as opposed to how the ‘system’ thinks it is [6 13]; for workflow systems do not actually represent work practices in a pragmatic sense through a completely valid association between work items and knowledge workers [12].

SNA can aid in providing a number of measures, approaches or techniques for evaluating alternative process designs [6]. A brief explanation of some of these measures is provided here. In directed networks with asymmetric relationships between actors or individuals such measures include centrality and prominence where the former refers to the number of information flows going out from one actor or individual to another while prestige or prominence relate to ties or information flows coming in from another actor [14]. In undirected networks such as the one presented in figure 1, the edge strength is the only means of determining actor relationships. Typically edge thickness is used to measure either quality of contact or frequency of contact. Quality of contact is typified by the type of media actors may use to pass information or knowledge to one another, where typically electronic means of communication are considered to lead to poorer information transfer while face-to-face visual means of communicating are said to lead to richer knowledge exchange [1]. Our edges in figure 1 use the latter approach illustrating frequency of contact as a means of determining how often actors interact and thereby exchange their knowledge. Another SNA measures is that of density, which refers to the number of relationships that exist between individuals. For example if 4 actors are all joined directly by way of 6 edges in the graph, the network density would be 1.0 or 100%. If there were only 4 edges joining the 4 actors, rather than 6, then the network density would be 0.7 rather than 1.0. SNA software will also perform Multi-Dimensional Scaling (MDS) where nodes are positioned in the resultant graph (figure 1) to illustrate the closeness of one actor to

another, based upon questionnaire parameters that have been fed in to the software. Where some actors are not as close according to their questionnaire results, this may be visualized in the resultant output graph. Arguably the best means of conducting research involving the exploration of concepts in a practical setting is through case study, where employee interactions can be explored and potentially generalised.

3 An Australian Example

LCPL is an existing mining and construction company headquartered in Sydney, employing in excess of 5,000 employees across Australia. Business processes in LCPL are organised into three areas; they are *winning*, *delivering* and *supporting* work processes.

3.1 The Process: Oracle Primavera P6

Within the *supporting* work processes section of LCPL is the Control and Planning (CP) division. In turn CP has different groups who provide the standards and policies for other project groups in the company. One such group is Group Operational Services (GOS), responsible for supporting other business units in developing and maintaining control and planning standards and methodologies. For the purpose of this research-in-progress case study, the use by GOS employees of Oracle Primavera P6 will be examined. P6 is an example of project portfolio management software. The purpose of the case study was to compare a real business process workflow that is already applied in the company against the informal social networks that exist between employees. Officially all divisions within LCPL should use P6 for their project planning, update and control purposes. To support P6 users, the GOS cooperates with LCPL's IT group to integrate the support process with the IT service desk for the company, where IT Service Management (ITSM) help-desk software is used to log requests, changes and incidents.

To gather data from the relevant P6 process employees, open-question interviews were conducted with 12 staff members at LCPL between April and May of 2011. The interviewees were chosen because (a) they formed a group that used the P6 process intensively, (b) they provided a blend of occupation types, and (c) were close colleagues of the second named author. The roles of the interviewees were optimization manager, three senior planners, three planners (not senior), a scheduler, a planning manager, a site engineer, a systems analyst and a business analyst. In these same interviews, interviewees were also asked to nominate colleagues they networked with and how often. These last details were then fed in to SNA software so their working relationships could then be mapped the results of which appear in figure 1.

3.2 The P6 Group at LCPL through SNA

Figure 1 explores the interaction of the staff introduced above, through the lens of SNA. Merely for the reader's interest, squares represent males, circles represent

females and colours of the nodes relate to the ages of the employees (red = 40 years old; green = 36 years old etc). The edges in the graph reveal the strength of information flow, given that relations among actors determine access to information resources [4 14]. Edge-strength in descending order is as follows: a value of 6.0 on the edge of the graph represents hourly contact; a value of 5.0 represents contact every few hours; a value of 4.0 indicates daily contact between colleagues; a value of 3.0 indicates staff interact once every couple of days; finally a value of 2.0 infers no more than weekly contact between colleagues.

Some aspects are immediately apparent from figure 1. The senior planner (bottom left) is a relative P6-process isolate, who works on the process only through the senior planner to his top right. Only two colleagues: the business analyst (light blue in the centre of the graph) and her planner colleague to her bottom right use Lync (formerly Office Communicator); the business analyst's communication flow with this colleague is particularly strong as she communicates with him hourly. Interestingly the only staff involved in the P6 process who communicate via the IT help-desk's ITSM software are the senior planner (bottom right), the 'star' business analyst in the centre, the site engineer (in green - right centre), as well as the planning manager (in grey - centre left). Management was surprised at the paucity of staff utilising ITSM as their means of communication.

The remaining communication flows with regard to working on P6 related issues were by way of verbal communication. Note how often employees see one another; our bottom-left senior planner collaborates with his senior planner colleague only daily which may suffice, but other employees involved in the P6 process such as the P6 optimisation manager (top), meets with a senior planner (top) only weekly. Another planner (bottom centre) is also a relative isolate in the P6 process and has no other connection with her employees other than through the (high-centrality ranked) business analyst in the middle, and then only on a weekly basis. Are the above points of any relevance? In fact such communication patterns for dealing with P6 issues led to workflow re-engineering at LCPL.

4 Results

From an analysis of a multitude of sources including blogs, internal company literature and also the SNA results (figure 1), differences exist between the 'officially' documented P6 support process and the actual work-flow in LCPL. Users raise their P6 issues through phone and email or by approaching the IT help-desk personally, rather than through the official channel of ITSM. In addition, some P6-related staff use Lync to 'chat' with the support team, which was never intended to be part of the official P6 support process. As a result of this last discovery, management is now seeking to change processes to officially incorporate Lync.

The above analysis reveals that users raise their issues first with their colleagues and then more broadly with the P6 support team in a tacit knowledge sense. There exists substantial evidence that tacit knowledge flows underpin most knowledge understanding and transferral processes [1 9]. As such it is hardly surprising that

employees approach their colleagues first as a means of resolving issues, particularly if the strength of workgroup ties is strong [4].

However such extra steps in the P6 support process were never envisaged by LCPL management. The analysis reveals varying severities of issues received by the support team that differ markedly amongst diverse LCPL groups. For example employees with P6-savvy colleagues raise fewer issues, the ramifications of which are that the training support team in the company can obtain a better idea of the level of instruction. Even such superficial findings have nonetheless provided management at LCPL with ideas for process improvement.

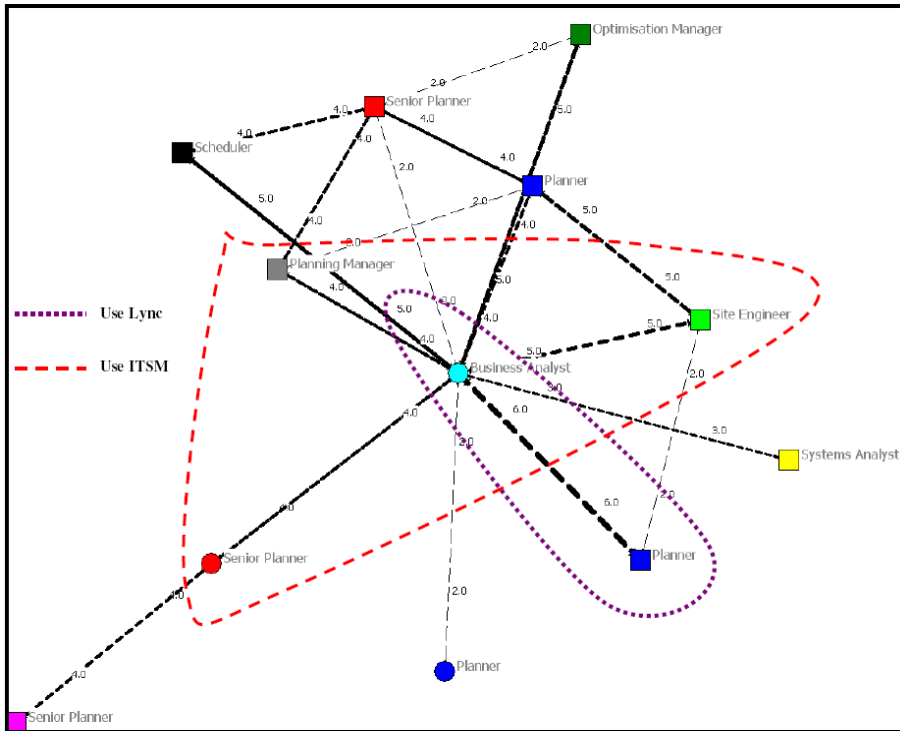


Fig. 1. illustrating the Social Workplace Network in LCPL

4.1 Need for Change

Business Process Management provides a consistent strategy to develop and implement organisational operational changes [15]. As an example here, it transpires that the documented P6 process is in fact not run by users and requires some modification to be more effective. The status quo at LCPL is as follows: 1) all support services should be performed through ITSM, which is the central support management system in the company; 2) users are distributed across the country and do not have access to their local domain constantly; 3) P6 is the enterprise application selected for all

projects, and GOS can receive requests from a project's client users; 4) there are some urgent requests that should be fixed with the highest priority.

Management have now come to acknowledge the following options at LCPL: 1) users can raise their requests by sending an email to the service desk and cc-ing their request to the P6 Support email address, so that the service desk can create a ticket while administration is fixing the issue; 2) that the ITSM software now incorporates a module helping supporters to capture calls as an 'incident'; this call-capture function will also be useful for training purposes later.

4.2 Options for the Firm

Arising out of the above changes to the P6 business process were two options for LCPL. Option 1 does not add costs to the business and the user support process will be more efficient and reliable. The benefits of these changes are: a) GOS can provide monthly support reports from their use of ITSM to plan their short-term and long-term support strategies; b) GOS can plan required training courses based on the level of issues received by the system; c) GOS can use ITSM for other types of services, such as the Process Support Service that supports project team members to the standards issued by the CP division; d) the GOS manager will be better able to calculate the work hours of GOS employees.

With option 2: e) all requests will be logged in to the system and the user need have no interaction with the ITSM software. However implementation of this recommendation produces a cost for LCPL in that some system settings will require re-adjustment, as it now needs to purchase this module as well as produce an internal staff costing. f) Additionally some modifications will be required to the P6 process for reporting purposes. Through the solution mentioned the business process was modified to fill the gap between the official process and the informal social network identified among employees via SNA.

Currently GOS is able to control all requests and can evaluate the work loads of its support team monthly. Furthermore, users are confident that they will receive their service at the right time, since their requests are in the queue and the support team cannot miss such requests by mistake. LCPL is now further currently establishing which documented business processes lack integration with social networks within the firm. Finally, in all businesses there are sets of hidden factors such as human social concepts that have significant impacts on business process management.

5 Conclusion

Knowledge Acquisition in its broadest sense adopts approaches used in other disciplines as a way of obtaining information critical to understanding how knowledge may be obtained and transferred. This research-in-progress paper has explored a means by which a comparison may be made between what management envisages as a formal approach to conducting a business process, and how such a process takes place in reality. In order to add a degree of ecological validity to the research, an actual Oracle Primavera P6 example was investigated which then led to an examination of relevant P6 team relationships through the 'prism' of SNA. Staff involved with the

P6 process, their attributes and their relationships were investigated and comparisons made against official management practice of the P6 process. The answer to “how extensive is the degree of overlap between an employee’s work processes and their organisational social network?” was ‘not as well as it could be’.

References

1. Busch, P.: Tacit Knowledge in Organizational Learning. IGI-Global, Hershey (2008)
2. Cox, A.: Reproducing knowledge: Xerox and the story of knowledge management. *Knowledge Management Research and Practice* 5(1), 3–12 (2007)
3. Guzman, G., Wilson, J.: The “soft” dimension of organizational knowledge transfer. *Journal of Knowledge Management* 9(2), 59–74 (2005)
4. Hansen, M.: The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organization Subunits. *Administrative Science Quarterly* 44(1), 82–111 (1999)
5. Harrison-Broninski, K.: Dealing with Human-Driven Processes. In: Rosemann, J.V. (ed.) *Handbook on Business Process Management* 2, pp. 443–461. Springer, Heidelberg (2010)
6. Hassan, N.: Using Social Network Analysis to Measure IT-Enabled Business Process Performance. *Information Systems Management* 26, 61–76 (2009)
7. Jiang, H., Carroll, J.: Social Capital, Social Network and Identity Bonds: A Reconceptualization. In: *C&T 2009*, June 25–27, pp. 51–60 (2009)
8. Liebowitz, J.: Linking Social Network Analysis with the Analytic Hierarchy Process for Knowledge Mapping in Organizations. *Journal of Knowledge Management* 9(1), 76–86 (2005)
9. Sternberg, R., Forsythe, B., Hedlund, J., Horvath, J., Snook, S., Williams, W., Wagner, R., Grigorenko, E.: *Practical Intelligence in Everyday Life*. Cambridge University Press, New York (2000)
10. Teplovs, C., Fujita, N., Vatrappu, R.: Generating Predictive Models of Learner Community Dynamics. In: *LAK 2011*, February 27–March 1, pp. 147–152 (2011)
11. Tichy, N., Tushman, M., Fombrun, C.: Social Network Analysis for Organizations. *Academy of Management Review* 4(4), 507–519 (1979)
12. van der Aalst, W., Kumar, A.: A Reference Model for Team enabled Workflow Management Systems. *Data and Knowledge Engineering* 38, 335–363 (2001)
13. van der Aalst, W., Reijers, H., Song, M.: Discovering Social Networks from Event Logs. *Computer Supported Cooperative Work* 14, 549–593 (2005)
14. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge (1994)
15. Weske, M.: *Business Process Management, Concepts, Languages, Architectures*. Springer, Heidelberg (2007)

Identifying Characteristics of Seaports for Environmental Benchmarks Based on Meta-learning

Ana Ximena Halabi Echeverry¹, Deborah Richards¹, and Ayse Bilgin²

¹Department of Computing, ²Department of Statistics
Macquarie University
NSW 2109, Australia

{ana.halabiecheverry, deborah.richards, ayse.bilgin}@mq.edu.au

Abstract. In this paper we discuss a model which classifies any seaport in the context of environmental management system standards as leader, follower and average user. Identification of this status can assist Port Authorities (PAs) in making decisions concerned with finding collaborating seaport partners using clear environmental benchmarks. This paper demonstrates the suitability of meta-learning for small datasets to assist pre-selection of base-algorithms and automatic parameterization. The method is suitable for small number of observations with many attributes closely related with potential issues concerning environmental management programs on seaports. The variables in our dataset cover main aspects such as reducing air emissions, improving water quality and minimizing impacts of growth. We consider this model will be suitable for Port authorities (PAs) interested in effective and efficient methods of knowledge discovery to be able to gain the maximum advantage of benchmarking processes within partner ports. As well as for practitioners and non-expert users who want to construct a reliable classification process and reduce the evaluation time of data processing for environmental benchmarking.

Keywords: classification in small datasets, meta-learning, environmental benchmarks, seaports.

1 Introduction

In the case of seaports, environmental benchmarks clearly identify competitive advantages and offer better risk management performance (Green Port Portal). “In general, benchmarking is the process of comparing individual objects which compete in a specific field of activity [1, p. xv]”. According to Bichou [2] the literature on performance measurement and benchmarking in seaports can be grouped into four broad categories: economic-impact studies, performance metrics and productivity index methods, frontier methods and process approaches. Port economics impacts directly or indirectly on social development, urban planning and environmental economics. For example Lee et al. [3] connect the economic and spatial dimensions of the pair city-port and Georgakaki et al. [4] construct relations between air pollutants such as diesel emissions and the operation of maritime transportation in the international trade. Performance metrics usually refer to input and output measures accounted for

efficient operational management of ports. Productivity indexes compare ports for example, on market price basis. The frontier benchmark concept denotes competitive and incentive-based productivity. Finally, the process approaches measure methods, methodologies and information flows. We take advantage of this distinction to categorize our framework of analysis into a hybrid approach of these categories which allows us to explore systemically the nature of decisions based on environmental benchmarking approaches.

Port authorities (PAs) are interested in effective and efficient methods of knowledge discovery to be able to gain the maximum advantage of benchmarking processes within partner ports. As decision-makers, they are requested to survey complicated models with the ease of the available computer software. Decisive factors emerged for management with the use of data mining and its related techniques in the process of making even the simplest decision. This also lead them to find patterns among ports that enable the identification of a partner port in an environmental benchmark context, desirable requirement for the port's sustainable development. Moreover, with the goal of finding key focus of collaboration for the innovation of port governance, knowledge discovery is highly beneficial condition to identify which port is more suitable for future connections (transportation of goods between ports). This paper documents a process to identify the characteristics of seaports which are closely related with potential issues concerning environmental management programs.

1.1 Overview of the Environmental Management System Standards

Many local and global agencies have previously determined whether a port has the minimum competencies to operate; whilst new endeavors are focused on more aggressive strategies towards radical actions to protect the environment from long-term PAs decisions. Fig. 1 summarises the main initiatives which we discuss here. The Port Safety and Environmental Protection Management Code (IPSEM) is the minimum merit for a port to operate under environmental management system standards (EMS) conditions. Other schemes are the result of award societies' programs such as The Environmental Improvement Annual Program Winners proposed by The American Association of Port Authorities (AAPA). Selection and completion procedures to hold the prize are highly demanding, ports are voluntarily called to register and do not need financial and technological commitment with national and international governments in order to comply with improved policies and regulations. Among the sector initiatives is also The EMS Primer for Ports, which is a formal system for proactively promoting awareness of the environmental footprint in a port. Over the last few years the Environmental Protection Agency (EPA) has been involved in the follow-up of this process, but conducting improvements on ports has been delegated to the leadership of PAs. Finally a stricter EMS framework is presented through international schemes such as The International Organization for Standardization -ISO14001. Ports who wish to benefit from this high recognition need to comply with complicated guidelines and conduct internal and external audits. This systematic program implies financial and technological contributions that in certain cases are not met by small but efficient ports. Therefore, other qualifications are available such as The World Port

Climate Initiative (WPCI) membership, for whom the commitment is based on compliance with international reducing greenhouse gas emissions (GHGs) standards and The Climate Registry Association which seeks collaboration among states with the initiative of port members to set consistent and transparent standards to calculate, verify, and report GHGs. Other initiatives include European and Asia ports such as The European Eco-Management and Audit Scheme (EMAS), specifically with its Port Safety and Health and Environmental Protection Management Code (PSHEM); The Partnerships in Environmental Management for the Seas of East Asia (PEMSEA); EcoPorts Foundation (EPF) with the Self Diagnosis Method (SDM) and the Port Environmental System (PERS) restricted to European Maritime States Members (ESPO).

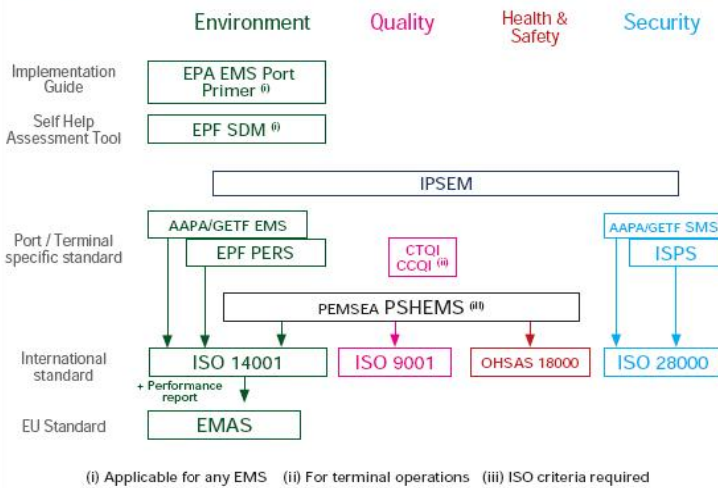


Fig. 1. Overview of standards, schemes and tools for EMS’s Seaports: Taken from Green Port Portal (www.greenport.com)

In order to facilitate benchmarking around environmental standards, we introduce the notion of a seaport leader (L), follower (F) and average user (A) of EMS. We want to identify the characteristics of each of these three user types. We have also identified a class variable named *StatusEMS* as the target variable for classification of the status dimension held by any seaport in the context of EMS. We extracted its concept from our understanding of the above overview of standards, schemes and tools for EMS’s seaport protocols.

1.2 Using Meta-learning for Small Datasets

In this paper we provide a framework for problems represented by small datasets. “...metaknowledge generated from a set of datasets is only useful in practice if they are representative of real world problems” [5, p.500], many of which are restricted by the number of instances. Carrying out experimental studies is the only viable

approach to obtain empirical knowledge. It is possible to say that, in practice, this approach guides the experimental process in a data mining application [6]. Moreover, there is not a specific interest in every single seaport but the representative ones for benchmarking purposes. This is the approach followed in this work.

An important issue in knowledge discovery regards to finding the optimal classifier. Recently, the literature points out meta-learning as a solution for the automatic prediction of the best classifier [7]. Meta-learning is usually compared with base-learning methods. Giraud-Carrier [8] states that whilst base-learning is concerned with gathering experience on a specific learning task, learning at the metalevel is concerned with gathering experience on the performance of a learning system. One of the important assumptions in meta-learning is that there exists an optimal learner algorithm a for each problem p . Among the several advantages of using the meta-learning approach is the increasing number of models and techniques that take advantage of its dynamic features on issues such as model selection and method combination. “The performance of learning algorithms is determined by data set characteristics and algorithms, this is common knowledge [1, p. 5]”. We use meta-learning taking advantage of its main properties [9, 10, 11]:

1. Preselection of base-algorithms
2. Automatic parameterization
3. Suitability to handle datasets consisting of a small number of cases
4. Preservation of classification accuracy

We also consider this model suitable for practitioners and non-expert users who want to construct a reliable classification process and reduce the evaluation time of data processing for environmental benchmarking. In the next section, data gathering and characterisation for an exemplar benchmark experiment is discussed and followed by its meta-learning application.

2 Data Gathering and Characterisation for the Problem Domain

This section characterises the problem domain and explains the data gathering process. The data gathering process involves selection of variables closely related with environmental issues covering three main challenges: reducing air emissions, improving water quality and minimizing impacts of growth (for more detail see Ng & Song [6], Kruse [12] and APPA [13]). Figure 2 displays an example



Fig. 2. An EMS program addressing main challenges. SOURCE: *Port Environmental Management Tools* [13].

of an environmental program which addresses most of the main challenges mentioned in the context of PA's long-term EMSs.

Reducing Air Emissions: Seaports are urged to join and ratify the International Convention for the Prevention of Marine Pollution (MARPOL) annexes in order to demonstrate willingness to operate and compete internationally. Vessels, cargo-handling equipment, trucks, and trains all contribute to air emissions at ports. *Facilities*, *oils* and *chemical* variables help to determine pollutant levels based on MARPOL annexes. Whereas, naturally weather marine conditions are affected by depositions of air pollutants. Common air pollutants include the variables nitrogen oxides (*NOx*), and sulfur oxides (*SOx*), carbon dioxides (*CO2*) and ozone (*O3*).

Improving Water Quality: Dredging activity place an important role on EMS programs because it involves water sediments and endangered species if habitat creation is damaged. Also, ballast water of onboard vessels is typically released in different maritime areas than where it was taken in, resulting in the introduction of non-native or invasive species. We use the variables *dredgeOcean* and national marine sanctuaries (*NMS*) to observe such impacts. On the other hand, most large ports have a high number of acres of paved waterfront for cargo handling, therefore; stormwater runoff picks up various pollutants before entering waterways. Some of the variables representing this phenomenon are *runoff* and need of water treatment (*needWtTreat*).

Table 1. Exemplar dataset for environmental challenge

Reducing Air Emissions	<i>Facilities, Inadequacies, CO2, O3comply, O3, SO2, NOx, Scientist</i>
Improving Water Quality	<i>needWtTreat, Facilities, oils, chemicals, Inadequacies, Scientist, runoff, NMS, dredgeOcean</i>
Minimizing Impacts of Growth	<i>CRP, LandFarms, Scientist, MarketVal, GAPStatus1, GAPStatus2, GAPStatus3, GAPStatus4, CountyArea, LeaseNum, LeaseAcres</i>

Minimizing Impacts of Growth: Generally, surrounding communities are increasingly interested in the impacts of port expansion. Congestion, safety, and other issues are derived from port growth. Variables such as the county area (*CountyArea*) represent notable boundaries for each seaport, for instance, changes in waterfront

areas and voting precinct lines that are now included in the current county administration. The list of variables for each of the environmental challenges introduced above can be found in Table 1.

Table 2. Number of attributes at sites and dataset assembled

Site/Repository	Num Attrib
site (BOERMRE)	4
site (EPA)	6
site (IMO)	4
site (USDA)	4
site (USGS)	5
site (USCOUNTY)	1
site (USAGE)	1
Vertical-Ensemble	25

Data have to be gathered from multiple available online repositories to cover the scope of the issue. The repositories used are the Bureau of Ocean Energy Management, Regulation and Enforcement (BOERMRE), The Environmental Protection Agency (EPA), the International Maritime Organization (IMO), United States

Geological Survey (USGS), United States Department of Agriculture (USDA) and the US Army Corps of Engineers (USACE). Appendix II indicates for each variable its respective association with the repository.

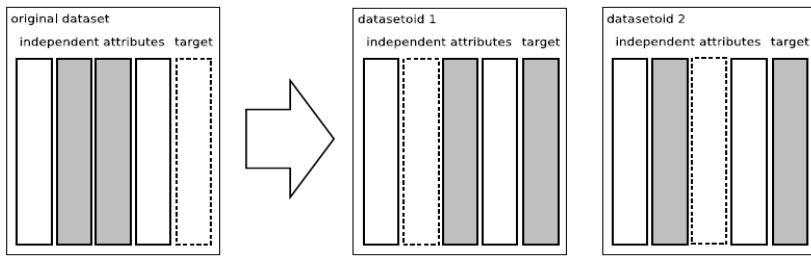


Fig. 3. Graphical representation the assembled dataset and the datasets in each site. Taken from [16, p.502]

We acknowledge that small datasets, collected in multiple locations, sites or repositories, impose a different approach to model construction (Skillicorn & McConnell) [14]). As this is our approach, we display in Table 2 the number of attributes collected per site, and in Figure 3 the graphical representation of the relationship between the assembled dataset and the datasets in each site.

We have also attempted to demonstrate that 44 observations with variables obtained from 7 different sites improve the opportunity to develop the dataset analysis [15]. Moreover, using meta-learning we have compared the classification accuracies per site in each subset of variables and the one after the dataset has been assembled and this work will be reported elsewhere.

Table 3. Description of the characteristics or meta-features used for the EMS seaport problem domain. Based on [1] and Souto et al. [16] approaches.

Chr/Meta-f	Description – Dataset assembled	Dataset value
obs.n	number of observations	44
var.n	number of variables	27
nvar.n	number of nominal variables	6
nvar.bin	number of binary variables	3
cvar.n	number of continuous variables	18
resp.cl	number of response classes	3
class.d	class distribution	A/F/L= 0.64/0.20/0.16
LgE	Log10 of the number of examples (raw indication of the available amount of training data)	0.608
LgREA	Log10 of the ratio: number of examples/number of attributes (indication of the number of examples available to the number of attributes.	1.629
PMV	Percentage of missing values (indication of the quality of the data)	0.037
PFA	Percentage of the attributes that were kept after the application of the remove correlation operation.	0.926

To facilitate the characterisation of the assembled dataset it is useful to find a standard set of meta-features that might be used in this and other benchmarking activities. A basic description of this characterisation is presented in Table 3 based on [1] and Souto

et al. [16] approaches. Thirdly, meta-learning can be used for preselection of base-algorithms. With meta-learning a number of algorithms can be executed simultaneously (which is possible in Rapid Miner 5.0®), while tracking their performance order. This information helps us to understand the relationship between the dataset and the performance of algorithms, resulting in better model construction and application.

3 Method of Analysis

We documented in section 2 our data gathering process over a number of data sites and the characterisation for the assembled dataset, which constitutes the meta-features for the model construction. Figure 4 displays a clear representation of the meta-learning process we have thoroughly followed: starting with the selection of repositories of data, following by the characterization of the dataset and the pre-selection of base-algorithms. In this section we explain the method of analysis and in section 4 we make recommendations based on the performance of the models (algorithms). The most likely classifiers are ranked by predicted accuracies along with their root mean square error (RMSE) values.

As shown in Table 3, the EMS seaport dataset includes 44 seaports, 3 classes on response (with a class distribution of 63.7% for average users, 20.4% for followers and 15.9% for leaders), 25 relevant attributes and 2 irrelevant variables according to a remove correlation operation (i.e. GAP Status 2 and Lease Acres). As previously mentioned, the most likely classifiers are ranked by prediction accuracies and root mean square errors (RMSE) as displayed in Table 4. The candidate classifiers used are neural networks (nnet), one specific library of support vector machines (svm), random forests weka algorithm (rf-w), k-nearest neighbor classifier (knn), random forests (rf), naïve bayes (nb) and the learner supervised rules – OneR (or). Using the automatic system construction wizard in Rapid Miner 5.0®, the meta-learning classification is a straightforward process. This wizard also aids evaluating each classifier and finding an optimal parameterisation for the dataset at hand.

Another issue of this model is the selection variables which are defined as “dataset characteristics representing the performance of simple learners on this dataset [5, p.1]”. This is why we seek the smallest group of learners likely to cover the learning space [6]. Performance results on these samples serve to characterize tasks. In meta-learning these are focused on accuracy measures which have been justified by theoretical and pragmatic studies. Three main reasons are outlined by [8].

1. No single learning algorithm will construct hypothesis of high accuracy on all problems.
2. Predictive accuracy is virtually impossible to forecast. (i.e. to know how accurate a hypothesis will be, the hypothesis has to be induced) by the selected learning model and tested on unseen data.
3. Predictive accuracy is easy to quantify. It is not subjective and induces a total order on the set of all hypotheses. Given an application, it is straightforward, through experimentation, to find which of a number of available models produces the most accurate hypothesis.

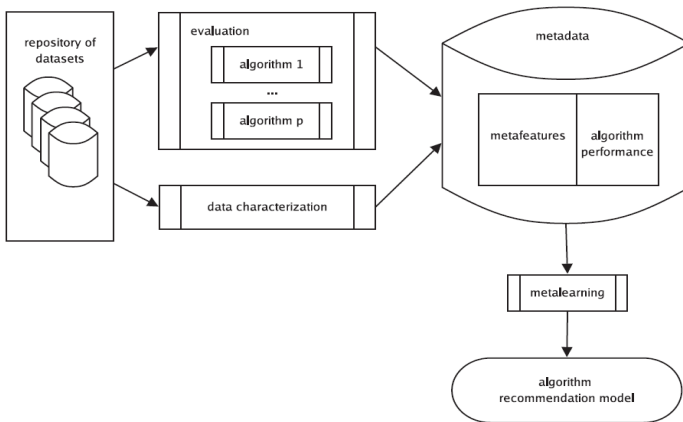


Fig. 4. Meta-learning process for the model (algorithm) selection. SOURCE: [9, p. 13].

We used the parameterization observed in the automatic model construction of Rapid Miner 5.0®, which returned a prediction accuracy of 68.2% and a RMSE of 8.4% for Random Forest. Results provided by [7] also corroborate that Random Forest has the lowest RMSE and highest confidence of prediction among different classifiers for several problems they considered. Table 4 shows the corresponding classifier performances in decreasing order.

Table 4. Meta-learning for classification of the StatusEMS class

Classifier	RMSE	Accuracy
RandomForest (rf)	0.084	0.682
LibSVM (svm)	0.068	0.659
NeuralNetImproved(nnet)	0.078	0.636
RandomForest-Weka (rf-w)	0.052	0.636
NearestNeighbors (knn)	0.056	0.636
OneR (or)	0.083	0.614
NaiveBayes (nb)	0.083	0.295

4 Exemplar Benchmark Experiment

This section introduces an exemplar benchmark experiment built in order to discover the capabilities of the previous learning method in the classification of the *StatusEMS* label. The class label has been created numerically using a weighting criterion of 5, 3 or 1 according to the level of compliance with the overviewed standards reported by the literature. The weighting factor corresponds to 5 when the port holds a high recognition of an EMS framework such as ISO14001, WPCI and Climate-Registry. The weighting criterion corresponds to 3 when the port competes in schemes, such as AAPA, GETF EMS and EPA EMS Port Primer, and receives an award. In this case, the factor is multiplied by the number of years in which the port holds the position.

Finally, a weight of 1 is given to each port who complies with the minimum merits of operation such as the IPSEM. For further clarification, two steps in the process were necessary; create a numerical variable with the sum of the weights and a class distribution based on its frequency. Figure 5 provides the distribution of weights for each class. See Appendix I for port’s weighting detail. Because Random Forest (RF) is a collection of tree classifiers, we compared the best performance of the experiment with the number of trees necessary to produce it. We found 3 trees as the full forest. A low number of variables were used in comparison with the initial ones. This indicates the algorithm identifies the most salient variables. These are; *GAPStatus1*, *dredgeOcean*, *Facilities* and *CountyArea*. They also indicate the presence of at least one variable per environmental challenge according to the problem characterisation. Implications are presented in Section 5.

During the model implementation, the best classification results are achieved when we used the following parameters for the Random Forest: *Number of trees=3*, *criterion=gain ratio*, *minimal size for split=4*, *minimal leaf size=2*, *minimal gain=0.1*, *maximal depth=20*, *confidence level used for the pessimistic error calculation of pruning= 0.45* and *number of alternative nodes tried when prepruning=3*.

We have attained 70.45% instead of 68.2% as an overall accuracy in this experiment by just changing the recommended parameter number of trees from 4 to 3. This reflects the prediction accuracy for the largest cohort of average users (A) of EMS framework (see Table 5). However, the best prediction is given for the followers, which in this case corresponds with 2 out of 9 ports. We select this model considering the limitations of the label class which has been numerically assigned. The RF trees are presented in Figure 6. To illustrate how the classification can be used to describe the distinct groups of ports, in the last section we explain and incorporate each of the predictor variables for the explanation of the profiles for each group of seaports.

Table 5. Confusion matrix for a seaport EMS classification problem

		Actual Class			Prediction
		Class=A	Class=F	Class=L	
Predicted Class	Class=A	26	6	4	72.22%
	Class=F	0	2	0	100%
	Class=L	2	1	3	50%
Recall		92.86%	22.22%	42.86%	

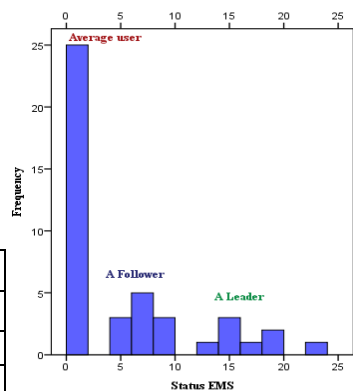
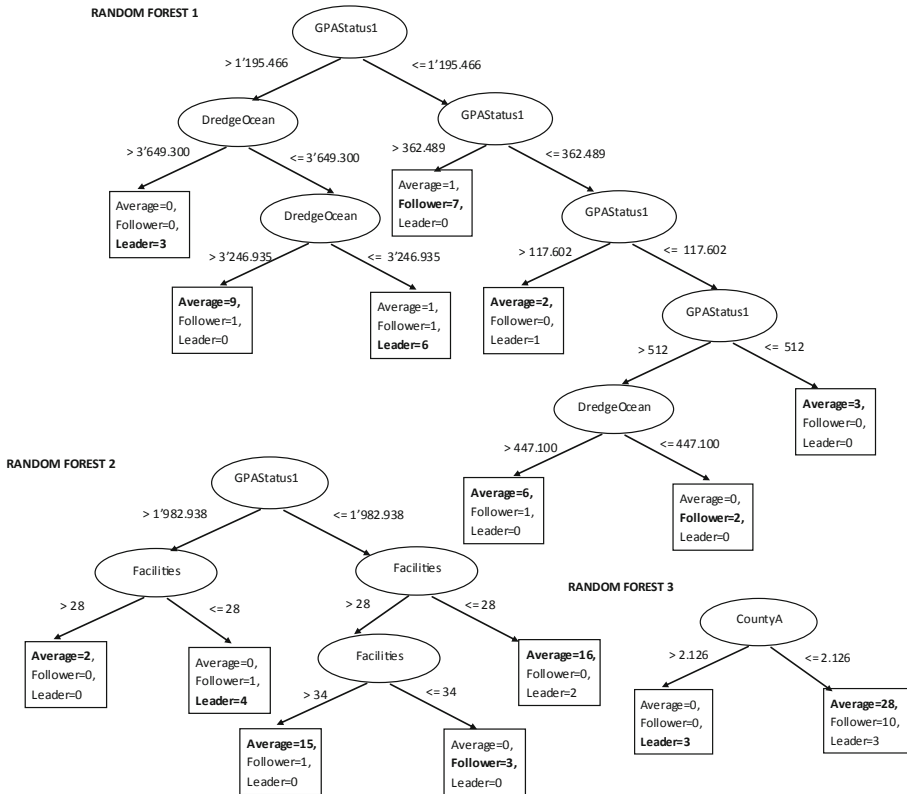


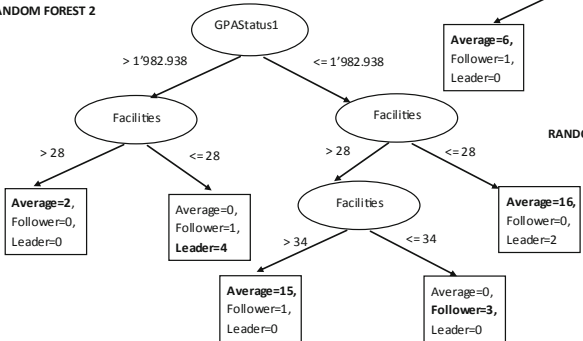
Fig. 5. StatusEMS Label Class Assignment

5 Practical Implications

The key objective of this model is to give an indication of the relationship between the Port’s EMS current framework and the future environmental challenges that PAs may face. This section allows the reader to understand the importance of such relationships. The regulatory function of the port has led the PAs to face high pressures to become accredited and internationally recognized. However, a number of environmental measures, produced by agencies and local administrative authorities, are difficult with respect to decision-making and as a result with defining strategies to understand the consequences of collaboration between seaports using clear environmental benchmarks. We believe the following classifications would help PAs to become more aware of the principles behind environmental benchmarking. In our analyses, three major profiles emerged. These are: 1) who the port leader might be, 2) the follower and 3) the average user of EMS programs. These profiles are based on the concepts of the variables identified by the best model (RF):



RANDOM FOREST 2



RANDOM FOREST 3

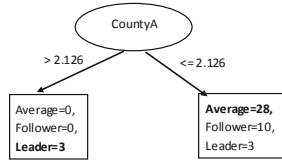


Fig. 6. Random Forest Classification

GAPStatus1. *current area of a state in acres in which the port is located (georeferencial measure) having permanent protection of natural land and water cover and a mandated management plan in operation to maintain a natural state within which disturbance events (of natural type, frequency, intensity, and legacy) are allowed to proceed without interference or are mimicked through management according with the USGS.*

DredgeOcean. *current amount of dredged ocean material disposed in cubic yards by state according with the US Army Corps.*

Facilities. *current number of port reception facilities (be they fixed, floating or mobile) in which final disposal of MARPOL residues/wastes occurs in a manner that protects the environment, the health and safety of workers and general population according with IMO*

CountyArea. *current county land area in square miles*

Profile of a Port Leader of EMS Programs. Characterised by having nearly more than two million acres of protected natural land and water cover and a mandated management plan to maintain this state. With less or around three million cubic yards of dredged ocean material disposed of annually they become one of the ports which regulate this activity. This kind of port accounts for less than 28 final disposal facilities which indicates a low demand of MARPOL disposal residues. Its location is better associated with county area of more than around two thousand square miles.

Profile of a Port Follower of EMS Programs. Characterised by having between three hundred and two million acres of protected natural land and water cover and a mandated management plan to maintain this state. The port conducts no defined dredged ocean material disposal activity. This kind of port accounts for 28 to 34 final disposal facilities which indicates a higher demand of MARPOL disposal residues compared with their leader counterparts. Its location is better associated with county areas of less than around two thousand square miles.

Profile of a Port Average User of EMS Programs. Characterised by having between around one hundred and 2 million acres of protected natural land and water cover and a mandated management plan to maintain this state, this kind of port overlaps with its follower counterpart's designation. However with more than around three millions of cubic yards of dredged ocean material annually disposed, it does not give evidence of regulation of this activity. Moreover, it accounts for more than 34 final disposal facilities which indicate a high demand of MARPOL disposal residues. Its location is better associated with county areas of less than around 2 thousand square miles.

6 Conclusions and Future Work

New avenues of future work on meta-learning are essential especially those obtained from empirical studies. We have taken an important step towards this aim. Because our dataset is based on a real application, we have applied our meta-learning analyses to an assembled dataset in order to assess accuracy of the method. We found that Random Forest (RF) performs well for small datasets. In the same way, we have provided a reference for problems with a limited amount of data and tested how

meta-learning can be used for pre-selection of base-algorithms. The key objective of this model is accomplished in the sense that the model has allowed us to give an indication of the relationship between the Port's EMS current framework and the future environmental challenges of PAs. Knowledge discovery and computer-aided strategic decision making are new endeavours in the seaport domain and thus experimental data and results for comparison purposes are not available. However, as future work we will be further evaluating our approach with datasets for other seaports and countries. Other directions include collection and decision making using time-series data and decision making in areas beyond environmental benchmarking.

References

4. Eugster, M.J.A.: Benchmark experiments: A tool for analysing statistical learning algorithms (Doctoral dissertation), http://edoc.ub.uni-muenchen.de/12990/1/Eugster_Manuel_J_A.pdf
5. Bichou, K.: A benchmarking study of the impacts of security regulations on container port efficiency. Centre for Transport Studies (Doctoral dissertation). Department of Civil and Environmental Engineering, Imperial College London (2008)
6. Lee, S., Song, D., Ducruet, C.: A tale of Asia's world ports: The spatial evolution in global hub port cities. *Geoforum* 39, 372–385 (2008)
7. Georgakaki, A., Coffey, R.A., et al.: Transport and Environment Database System (TRENDS): Maritime air pollutant emission modelling. *Atmospheric Environment* 39, 2357–2365 (2005)
8. Soares, C.: UCI++: Improved Support for Algorithm Selection Using Datasetoids. In: Theeramunkong, T., Kijssirikul, B., Cercone, N., Ho, T.-B. (eds.) PAKDD 2009. LNCS, vol. 5476, pp. 499–506. Springer, Heidelberg (2009)
9. Ng, A.K.Y., Song, S.: The environmental impacts of pollutants generated by routine shipping operations on ports. *Ocean & Coastal Management* 53, 301–311 (2010)
10. Abdelmessih, S.D., Shafait, F., Reif, M., Goldstein, M.: Landmarking for Meta-Learning using RapidMiner. In: German Research Center for Artificial Intelligence, Germany (2010), <http://www.mendeley.com/research/landmarking-metalearning-using-rapidminer>
11. Giraud-Carrier, C.: Toward a Justification of Meta-learning: Is the No Free Lunch Theorem a Show-stopper? In: Proceedings of the International Conference on Machine Learning, Workshop on Metalearning, pp. 9–16 (2005)
12. Brazdil, P., Leite, R.: Determining the Best Classification Algorithm with Recourse to Sampling and Metalearning. *Advances in Machine Learning* 1262, 173–188 (2010)
13. Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, R.: Metalearning for Algorithm Recommendation: an Introduction. *Metalearning*, 11–29 (2009)
14. Brazdil, P.B., Soares, C., Da Costa, J.P.: Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning* 50, 251–277 (2003)
15. Kruse, C.J.: Environmental Management Systems at Ports - A new initiative. In: Proceedings of the 14th Biennial Coastal Zone Conference (2005)
16. American Association of Port Authorities AAPA, Environmental Management Handbook (1998)

17. Skillicorn, D.B., McConnell, S.M.: Distributed prediction from vertically partitioned data. *Journal of Parallel and Distributed Computing* 68, 16–36 (2008)
18. Chang, F.M.M.: Characteristics analysis for small data set learning and the comparison of classification methods. In: *Advances on Artificial Intelligence, Knowledge Engineering and Data Bases, Proceedings of the 7th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED 2008)*. University of Cambridge, Cambridge (2008)
19. de Souto, M.C.P., Prudencio, R.B.C., Soares, R.G.F., Araujo, D.A.S., Costa, I.G., Luder-mir, T.B., Schliep, A.: Ranking and selecting clustering algorithms using a meta-learning approach. In: *IJCNN - International Joint Conference on Neural Networks* (2008)

Appendix I. Standards used to create labels for the seaports as follower, average and leader

IdPort	Locode	Stabbr	PortName	ISO14001	AAPA / GETF EMS		EPA EMS Port Primer	IP-SEM	WPCI	Climate-Registry	Status-EMS	Status-EMSLabel
					High=5	Medium=3 (per year)						
C0128	USPWM	ME	Portland ME					X			1	Average
C0149	USBOS	MA	Boston		X09		X	X			7	Follower
C0297	USCHT	PA	Chester					X			1	Average
C0398	USNYC	NY	New York	X		X		X	X	X	19	Leader
C0552	USPHL	PA	Philadelphia					X			1	Average
C0554	USILG	DE	Wilmington DE					X			1	Average
C0700	USBAL	MD	Baltimore		X10	Xd		X			7	Follower
C0766	USILM	NC	Wilmington NC					X			1	Average
C0773	USCHS	SC	Charleston		X09			X			4	Average
C0776	USSAV	GA	Savannah					X			1	Average
C0780	USSSI	GA	Brunswick					X			1	Average
C2004	USPGL	MS	Pascagoula					X			1	Average
C2005	USMOB	AL	Mobile		X09			X			4	Average
C2012	USHOU	TX	Houston	X		X		X	X		14	Leader
C2016	USPFN	FL	Panama City					X			1	Average
C2017	USJAX	FL	Jacksonville					X			1	Average
C2021	USTPA	FL	Tampa					X			1	Average
C2083	USGPT	MS	Gulfport					X			1	Average
C2162	USPAB	FL	Palm Beach					X			1	Average
C2163	USPVS	FL	Port Everglades	X		Xd		X			9	Follower
C2164	USMIA	FL	Miami					X			1	Average
C2251	USMSY	LA	New Orleans	X		Xd		X			9	Follower
C2254	USLCH	LA	Lake Charles					X			1	Average
C2395	USBPT	TX	Beaumont					X			1	Average
C2404	USTXT	TX	Texas City					X			1	Average
C2408	USFPO	TX	Freeport		X08	Xd		X			7	Follower
C2416	USPOA	TX	Port Arthur		X10			X			4	Average
C2417	USGLS	TX	Galveston					X			1	Average
C2423	USCRP	TX	Copus Christi	X			X	X			9	Follower
C4100	USSAN	CA	San Diego					X		X	6	Follower
C4110	USLGB	CA	Long Beach		X11,10	Xd		X	X		15	Leader
C4120	USLAX	CA	Los Angeles	X		X		X	X	X	19	Leader
C4150	USNTD	CA	Port Hueneme					X			1	Average
C4335	USFO	CA	San Francisco					X			1	Average
C4345	USOAK	CA	Oakland			Xd		X	X	X	14	Leader
C4350	USRCH	CA	Richmond CA					X			1	Average
C4420	USHNL	HI	Honolulu					X			1	Average
C4644	USPDX	OR	Portland OR	X	X08	X		X		X	17	Leader
C4708	NA	WA	Port Angeles					X			1	Average
C4720	USTIW	WA	Tacoma		X10,08			X			7	Follower
C4722	USSEA	WA	Seattle		X11,10,08	X		X	X	X	23	Leader
C4816	USVDZ	AK	Valdez					X			1	Average
C4820	USANC	AK	Anchorage					X			1	Average
C5735	USORF	VA	Norfolk	X	X11	X		X			12	Follower

Appendix II. Definition of the variables and their sources

Variable Name	Definition	Variable source
Lease Area	Named area where oil and gas leasing activity take place within the port location	Bureau of Ocean Energy Management, Regulation and Enforcement BOERMRE http://www.boemre.gov/ld/PDFs/OCSstatusMap8e3.pdf
Need W/Treat	Updated needs of water treatment by state in billions of dollars	Environmental Protection Agency "Clean Watersheds needs survey" (CWNS) 2008. http://water.epa.gov/scitech/datait/databases/cwns/upload/apex-2.pdf http://waters.geo.epa.gov/mwm/ layer=LEGACY_WBD&feature=03160205&extraLayers=null
facilities	Updated number of port reception facilities (be they fixed, floating or mobile) in which final disposal of MARPOL residues/wastes occurs in a manner that protects the environment, the health and safety of workers and general nonpollution	IMO International Maritime Organization - https://gis.imo.org/Public/PRF/Default.aspx
oils	Updated maximum discharge (m3) at port allowed and generated on board ships as oily waste, oily mixtures, oily bilge water, slops, sludge, oily tank washings, oily cargo residues, ballast water containing oily mixtures as defined in MARPOL annex I	IMO International Maritime Organization - https://gis.imo.org/Public/PRF/Default.aspx
chemicals	Updated maximum discharge (m3) at port allowed and generated on board ships as tank washings and cargo residues containing noxious liquid substances (NSL) as defined in MARPOL annex II	IMO International Maritime Organization - https://gis.imo.org/Public/PRF/Default.aspx
Inadequacies	Yes/No historical problems encountered at the port reception facility and informed to IMO	IMO International Maritime Organization - https://gis.imo.org/Public/PRF/Default.aspx
CO2	Tons of CO2 (carbon dioxide) calculated based on emissions per hour and operating time for the hour, measured in 2008 and by State in which the EPA facility is located	Environmental Protection Agency http://camdataandmaps.epa.gov/gdm/index.cfm
O3 comply	Compliance with the CASTNET standards up to 75 ppb (parts per billion) of fourth-highest daily maximum 8-hour average of ozone concentrations (O3 -air pollutants)	Environmental Protection Agency "Clear Air Status and Trends Network (CASNET) 2008 Annual Report" http://epa.gov/casnet/jav/aweb/docs/annual_report_2008.pdf
O3	Fourth-highest daily maximum 8-hour average of ozone concentrations (O3 -air pollutants) measured in ppb (parts per billion) within an area of a state by analysts of CASTNET (EPA) in 2008 (nominal)	Environmental Protection Agency "Clear Air Status and Trends Network (CASNET) 2008 Annual Report" http://epa.gov/casnet/jav/aweb/docs/annual_report_2008.pdf
O3cont	Fourth-highest daily maximum 8-hour average of ozone concentrations (O3 -air pollutants) measured in ppb (parts per billion) within an area of a state by analysts of CASTNET (EPA) in 2008 (continuous)	Environmental Protection Agency "Clear Air Status and Trends Network (CASNET) 2008 Annual Report" http://epa.gov/casnet/jav/aweb/docs/annual_report_2008.pdf
SO2	Significant deposition of sulphur dioxide (SO2) in the production of acid rain in tons measured in 2008 within an area of a state.	Environmental Protection Agency http://camdataandmaps.epa.gov/gdm/index.cfm
NOx	Significant deposition of nitrogen oxides (NOx) in the production of acid rain in tons measured in 2008 within an area of a state.	Environmental Protection Agency http://camdataandmaps.epa.gov/gdm/index.cfm
CRP	Average rental payment per acre in US dollars FY 2008 of lands enrolled in the Conservation Reserve Program (CRM)	Data.gov Department of Agriculture http://explore.data.gov/Agriculture/Conservation-Reserve-Program-Average-Payments-by-S/7w2u-44pa
Land-Farms	Land in farms given in acres by state and county where the port is located or closely located.	Data.gov Department of Agriculture http://explore.data.gov/Agriculture/Census-of-Agriculture-Race-Ethnicity-and-Gender-Privdnn-fk45
Scientist	Rank of scientists FTEs (full-time equivalents) based on National Institute of Food and Agriculture (NIFA) funding in 2008 counted by state	USDA United States Department of Agriculture http://www.reis.usda.gov/portal/page?_pageid=193.1&_dad=portal&_schema=PORTAL
Market-Val	Ranked market value of all farm products by state	USDA United States Department of Agriculture http://www.reis.usda.gov/portal/page?_pageid=193.1&_dad=portal&_schema=PORTAL
GAP-Status1	Current GIS acres calculated for a state under the protection laws from conversion of natural land and water cover according with the GAP Analysis Program in the US Geological Survey (USGS) and coded 1 for a mandated management plan in operation to maintain a natural state within which disturbance events (of natural type, frequency, intensity, and legacy) are allowed to proceed without interference or are mimicked through management	United States Geological Survey http://gapanalysis.usgs.gov/protected-area-statistics-by-state/
GAP-Status2	Current GIS acres calculated for a state under the protection laws from conversion of natural land and water cover according with the GAP Analysis Program in the US Geological Survey (USGS) and coded 2 for a mandated management plan in operation to maintain a primarily natural state, but which may receive uses or management practices that degrade the quality of existing natural communities, including suppression of natural disturbance	United States Geological Survey http://gapanalysis.usgs.gov/protected-area-statistics-by-state/
GAP-Status3	Current GIS acres calculated for a state under the protection laws from conversion of natural land and water cover for the majority of area. According with the GAP Analysis Program in the US Geological Survey (USGS) and coded 3. Subject to extractive uses of either broad, low-intensity type (e.g., Logging) or localized intense type (e.g. Mining). Confers protection to federally listed endangered and threatened species throughout the area	United States Geological Survey http://gapanalysis.usgs.gov/protected-area-statistics-by-state/
GAP-Status4	Current GIS acres calculated for a state under the protection laws according with the GAP Analysis Program in the US Geological Survey (USGS) and coded 4 for unprotected areas or unknown protection. The remaining area (land and water) of a state (not designated as GAP Status 1-3) is classified as GAP Status 4. Status 4 areas are primarily private lands. They have no known public/private institutional mandates or legally recognized easements.	United States Geological Survey http://gapanalysis.usgs.gov/protected-area-statistics-by-state/
runoff	Runoff given in mm by water-year 2008 and state. "Runoff is the water flow that occurs when soil is infiltrated to full capacity and excess water from rain, melt water or other sources flows over the land" (Wikipedia)	United States Geological Survey http://waterwatch.usgs.gov/new/index.php?r=a&id=statsum
County Area	Updated county land area in square miles	USCounty. Org http://uscounty.org/us-counties-descending-by-population.htm
Lease-Num	Existing number of leases since 2008 for oil and gas production within a leasing area where the port is located	Bureau of Ocean Energy Management, Regulation and Enforcement BOERMRE http://www.boemre.gov/ld/PDFs/OCSstatusMap8e3.pdf
Lease Acres	Existing number of land acres since 2008 for oil and gas production within a leasing area where the port is located	Bureau of Ocean Energy Management, Regulation and Enforcement BOERMRE http://www.boemre.gov/ld/PDFs/OCSstatusMap8e3.pdf
NMS	Yes/no existence of national marine sanctuaries (NMS) where the port is relatively close located	Bureau of Ocean Energy Management, Regulation and Enforcement BOERMRE http://www.boemre.gov/ld/PDFs/OCSstatusMap8e3.pdf
Dredge Ocean	Updated amount of dredged material ocean disposed in cubic yards by state according with the US Army Corps	US Army Corps of Engineers http://el.erdc.usace.army.mil/odd/SiteQuery.asp

A Situated Experiential Learning System Based on a Real-Time 3D Virtual Studio

Mihye Kim¹, Ji-Seong Jeong², Chan Park², Rae-Hyun Jang³, and Kwan-Hee Yoo^{2,*}

¹ Department of Computer Science Education, Catholic University of Daegu, South Korea
mihyekim@cu.ac.kr

² Department of Information Industrial Engineering and Department of Computer Education,
Chungbuk National University, South Korea
{farland83, szell, khyoo}@chungbuk.ac.kr

³ Virtual Reality Business Team, Korea Internet Software Corporation, South Korea
jrh@kis21.com

Abstract. Current educational systems are creating new teaching methodologies that can produce various types of experiential learning opportunities owing to the convergence of information technology with graphics, multimedia and virtual reality (VR) technologies. This paper introduces a situated experiential learning system based on a real-time three-dimensional (3D) virtual studio, developed by improving upon several technical shortcomings inherent in the previously developed virtual education system, KIS-VR 1010. The main improvements include system performance, the quality of 3D background templates and 3D background effects, and the control functions of the user interface. A remote control feature and a speech-recognition function are also incorporated into the new system. The proposed system was successfully tested in English classes at schools and language institutes that provide an enhanced learning environment. The visualization of a real-time 3D virtual space, the interactive user interface, and the speech-recognition feature were found to form the basis for an effective learning environment, facilitating self-regulated study by enhancing student engagement in the learning process.

Keywords: Situated experiential learning system, Real-time 3D virtual studio.

1 Introduction

With the advent of an era of rapidly changing multimedia, traditional classroom-oriented teaching and learning systems are giving way to web-based systems with multimedia content. Recently, more active, self-directed cyber learning systems have been developed by leveraging new state-of-the-art technologies, such as digital media, real-time virtual reality (VR), and augmented reality (AR) technologies. Such VR-based teaching and learning systems can provide new experiential learning opportunities by eliminating the simplicity and boredom of traditional textbook-oriented or two-dimensional (2D) multimedia content-based learning systems. This type of

* Corresponding author.

system can also enhance the self-regulated learning abilities of students by stimulating their motivation to learn. In accordance with these educational trends, various types of virtual education systems have been developed and commercialized in Korea [1-4]. However, many technical shortcomings remain that require improvement and further development, especially with regard to system performance, graphic resolution, and the user interface. Furthermore, most VR-based educational systems are expensive to install because they are usually hardware based.

This paper proposes a situated experiential learning system that employs chroma keying and real-time three-dimensional (3D) VR techniques, and is software rather than hardware based, thereby enabling high usability at low cost. The proposed system is rooted in KIS-VR 1010 [5] and improves upon several of that system's technical shortcomings while incorporating some additional features. KIS-VR 1010 is a learning system based on multimedia and 3D VR technologies developed by the Korea Internet Software Company [1]. The main improvements of the proposed system are to system performance, the quality of the 3D background templates and 3D background effects, and the control functions of the user interface. The new features include a remote controller and a speech-recognition function. The goal of developing this system was to provide a more advanced situated experiential learning environment in a 3D virtual studio that would heighten students' motivation and engagement in learning. We also aimed to provide optimal multimedia content to facilitate self-directed study through the visualization of real-time 3D graphics and voice recognition. Moreover, the system was intended to provide a more dynamic interactive environment between students and teachers, allowing them to learn together in the same virtual studio space. Note that the present paper is an extended version of a previously published paper [6], and thus includes some identical content and figures.

2 Related Work

Current VR technologies are evolving from single- to multi-user experience-centered techniques and from simulations for research verification to remote educational systems for public use. Applications of these technologies are also evolving from use in specific fields, such as medicine, to more general use, such as for education, games, shopping, and web-based broadcasts. Moreover, the technological environment has shifted from high-performance workstation-based systems to PC-based systems with network connections. Typical applications of VR are physical rehabilitation training [7], entrepreneurial training [8], virtual simulation education [9], design education [10], and language education [11].

Several companies in Korea have also developed and commercialized various types of virtual English learning systems based on the needs of specialized educational institutes. Representative developers of virtual learning systems in Korea include VR Media [2], INTOSYSTEM [3], and TAMTUS [4]. VR Media and INTOSYSTEM developed a virtual English educational system for situated experiential learning, and TAMTUS introduced the Magic VR-UCC Studio, which enables the construction of English-dedicated classrooms for experiential learning and allows students to have real-time interactive English conversations. Language education institutes such as YBM Sisa.com, Jung-Cheol Cyber Institute, and Sam-Yuk Language School are

utilizing this type of virtual learning system in their language classes. YBM Sisa.com uses the e-Speaking Learning System, which was developed by combining a customized learning management system with a sound-recognition solution in English language courses [12]. The Jung-Cheol Cyber Institute [13] provides a sound-recognition learning program that allows students to practice English expressions via online lectures. The Sam-Yuk Language School offers a “12-step” English learning program based on interactive voice-recognition solutions and communicative language teaching [14]. The 12-step program provides video lectures, virtual 3D images, and native-speaker voice training by installing a voice-recognition program on a computer.

However, most domestic systems continue to be hampered by several technical shortcomings, especially with regard to the quality of the 3D graphic images. Furthermore, most VR products are costly to install. This study has improved upon several of the technical shortcomings inherent in existing systems by developing a software-based product that operates in a PC environment.

3 Experimental Learning System

The proposed system is based on KIS-VR 1010, a multimedia learning system developed by the Korea Internet Software Co. and sold by UNIWIDE Technologies, Inc. (<http://www.uniwide.co.kr>). The objective was to improve the quality of situated experiential learning by upgrading the techniques used in KIS-VR 1010 and enhancing its effect on virtual experiential learning. The main improvements are to system performance and speed, the quality of the 3D background templates used for situated experiential learning, the quality of the 3D background special effects, and the control functions of the user interface. New features include a remote controller technique and a speech-recognition function.

3.1 Overview of the System

The situated experiential learning system introduced in this paper is based on a real-time 3D virtual studio that allows images to be created *in situ* by compositing real and virtual sets using stereoscopic 3D computer graphics. The system removes a specific single color or range of colors (such as blue or green) from a photographic image on a bluescreen using a chroma keying technique [5] (which is widely used to separate background colors and objects from actual images). The system then composes a virtual 3D environment by synthesizing the resulting image with a 3D background template prepared for a given learning situation. This type of composited virtual 3D environment enables students to engage in experiential learning, including scenarios anchored in reality, and to practice situated learning in specific places. Teachers can formulate appropriate learning content without spatial limitations by composing a 3D virtual environment to match the situation.

3.2 Main Improvements

System Performance. The speed of the system was enhanced by improving the video compression and decompression features, and by adding an alternative logic to handle

the delay due to system overhead in storing videos. The left screen of Fig.1 shows the existing and improved compression methods. Furthermore, several current multimedia technologies, including texture and feature extraction from video files, video streaming transmission, real-time 3D rendering, 3D graphics, and chroma keying techniques, were also used in the proposed system, resulting in more dynamic and realistic 3D virtual learning spaces.

Background Template. The quality of the background templates was enhanced by incorporating stereoscopic 3D VR techniques and interactive 3D animations into the learning content, resulting in improved situated experiential learning. In other words, the virtual studio development environment was improved. Not only images and photos, but also digital video disc (DVD) movies and animations can be utilized as background objects for situated experiential learning. Third-party content and user-created content (UCC) can be also employed as background templates. Thus, various virtual learning studios can be created from various types of image objects using 3D authoring techniques. This improvement can increase the motivation and understanding of students, thereby enhancing the learning effect. The right screen of Fig. 1 shows the various types of content that can be used for background templates in the system.

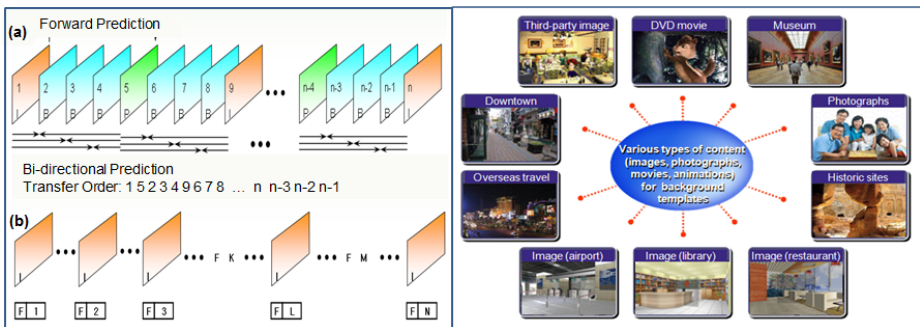


Fig. 1. The existing compression method (a), the improved compression method (b), and the various types of background template (right)

Special Effects of 3D Backgrounds. In the proposed system, several special animation effects have been strengthened by optimizing the equations of brightness, reflection, and color, and by adjusting the shading formula to improve the quality of the 3D background effects. In addition, the resolution of the rendering window displayed on the screen was upgraded by using stereoscopic 3D images. The quality of the 3D backgrounds was also enhanced by synchronizing external images with the image output on the screen. Furthermore, the synthesis and relocation technique for 3D images and the real-time texture-mapping technique for extracted images in a virtual space were upgraded. Other new development techniques for special effects include: a technique that converts external images on a bluescreen into coordinate information, and then transmits the information to the system; a technique that implements and plays synchronized data; a technique that creates smooth animations in a virtual space; a color-key keying technique; and a zooming technique that optimizes subject input as video signals in a 3D background.

User Interface. The control functions were improved by creating an interactive user interface. Users can immediately input any desired information to the system via the interface, which offers features such as selecting the location and direction of a virtual camera and selecting a 3D virtual space. Users can open and control (execute/manipulate/save) template files on a control monitor. The final result of these user activities is then displayed on a preview monitor. Fig. 2 shows the user interface of the system (left) and the configuration of the dual monitors (right).

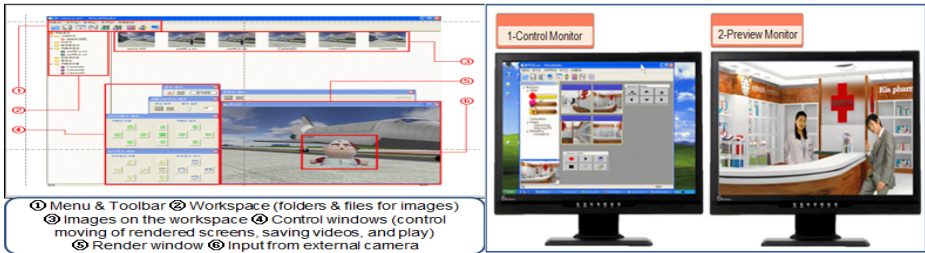


Fig. 2. User interface of the system (left) and configuration of the dual monitors (right)

Remote Control Technology. Remote control technology was incorporated into the system. Users can operate an English experiential learning system from a distance using a dedicated remote controller, as shown in the left screen of Fig. 3.

Speech Recognition. In the proposed system, we developed a sound-recognition feature that performs a specific motion in response to voice order data acquired by analyzing the characteristics of sounds input to a computer through a microphone or headset. This feature enables students to practice and correct their pronunciation via one-to-one conversations with a computerized native speaker in an English class, without requiring the presence of a real native speaker. The right screen of Fig. 3 shows the speech-recognition process used in the proposed system.

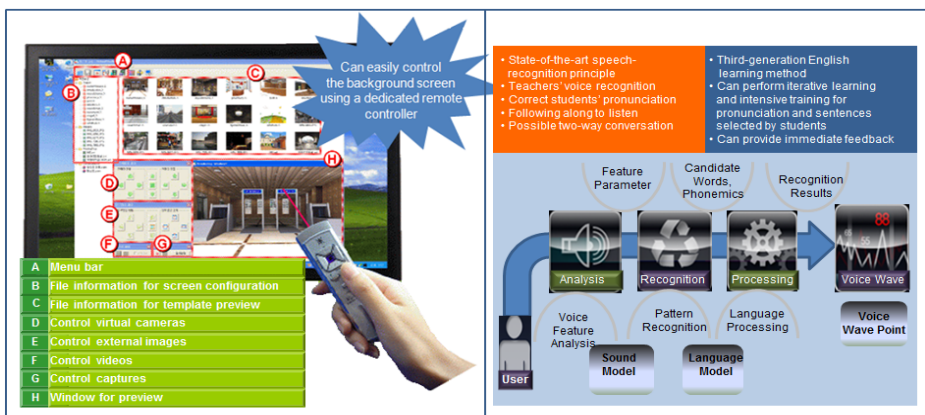


Fig. 3. Operating the learning system with a remote controller (left) and speech recognition process used in the system (right)

With these improvements, the proposed system offers several advantages over previously developed virtual learning systems. Firstly, the proposed system can produce more vibrant and realistic learning spaces because it uses real-time 3D rendering and stereoscopic 3D graphic techniques to visualize virtual learning spaces. Secondly, the interactive interface allows users to immediately input information to the system. Thirdly, student engagement in the learning process is heightened by enabling real-time compositing of user motions, actions, and sounds in a virtual space using a chroma keying technique. Another important feature is the enhanced interactive capabilities between users. The system allows users to carry on one-to-one or *n-to-n* English conversations in a virtual space. In other words, the system provides a learning environment in which several students can learn together.

4 Experimental and Conclusion

We have introduced a situated experiential 3D virtual learning system that improves upon several technical shortcomings inherent in previously developed virtual education systems, resulting in the creation of an enhanced learning environment. The left screen of Fig. 4 shows the software product, hardware specifications, and peripherals of the system. The right screen of Fig. 4 shows an overall illustration of a classroom equipped with the real-time VR studio developed for situated experiential learning.

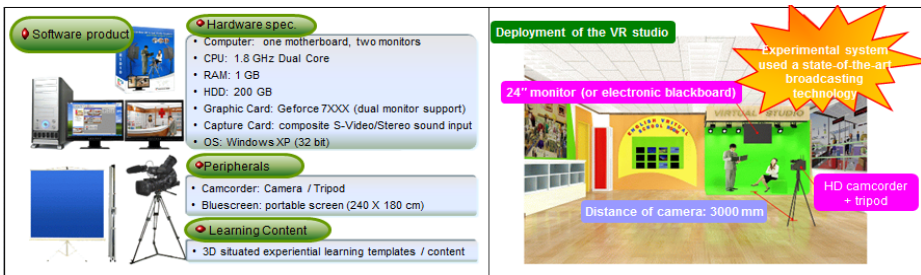


Fig. 4. Specifications of the System (left) and Deployment of the VR studio (right)



Fig. 5. Examples of English classes

Fig. 5 shows an example of how the proposed system was used for situated experiential learning in real English classes. While a teacher and students converse in front of the bluescreen, video images are created in real time by synthesizing the bluescreen images with appropriate 3D images previously loaded into the system. The resulting images are shown on a television and the dual monitors.

The proposed system has been tested in English classes at several schools in Korea and it is successfully using in English classes in about 400 elementary and secondary schools and language institutes across Korea, encompassing about 70% of the market share. We anticipate that the demand for this type of education will increase annually, and thus that the utilization of the proposed system will be high. Observations of actual use revealed that students were generally very interested in participating in the learning process. The visualization of real-time stereoscopic 3D virtual spaces, the interactive user interface, interactions among multiple users, and the speech-recognition feature were also found to provide the basis for an effective learning environment, enabling more efficient self-directed learning by enhancing student engagement in the educational process. For further development of the system, a wide range of case studies will be necessary to identify additional requirements and improvements from the users' point of view.

Acknowledgments. This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST)(No. 2012-0000479).

References

1. Korean Internet Software Co. English experiential learning system, <http://www.kisvr.com/>
2. VR Media Co., Virtual on air, <http://www.vrmedia.co.kr>
3. INTOSYSTEM Co., VR English, <http://www.vrenglish.co.kr>
4. TAMTUS (Total Unique IT Solution) Co., Virtual experiential learning system: Magic VR-UCC Studio, <http://www.tamtus.com/default>
5. Usage Guidance for KIS-VR 1010, Korea Internet Software, http://www.uniwide.co.kr/download/english/KIS_VR_m1.pdf
6. Jeong, J.-S., Park, C., Han, J.-J., Im, M.-S., Jang, R.-H., Kim, M., Yoo, K.-H.: Development of a 3D Virtual Studio System for Experiential Learning. In: Kim, T.-h., Adeli, H., Robles, R.J., Balitanas, M. (eds.) AST 2011. CCIS, vol. 195, pp. 78–87. Springer, Heidelberg (2011)
7. Keshner, E.A.: Virtual reality and physical rehabilitation: a new toy or a new research and rehabilitation tool. *Journal of NeuroEngineering and Rehabilitation* 1, 8 (2004)
8. Stieglitz, S., Lattermann, C., Kallschnigg, M.: Experiential Learning in Virtual Worlds – A Case Study for Entrepreneurial Training. In: Proceedings of the 16th Americas Conference on Information Systems (AMCIS), Paper 352, pp. 1–11 (2010)
9. Klaassens, J.B., Honderd, G., Azzouzi, A.E., Cheok, K.C., Smid, G.E.: 3D Modeling Visualization for Studying Controls of the Jumbo Container Crane. In: Proceedings of the American Control Conference 1999, pp. 1745–1758 (1999)

10. Gul, L.F., Gu, N., Williams, A.: Virtual Worlds as a Constructive Learning Platform: Evaluating of 3D Virtual Worlds on Design Teaching and Learning. *Journal of Information Technology in Construction (ITcon)* 13, 578–593 (2008)
11. Vickers, H., Languages, A.: VirtualQuests: Dialogic Language with 3D Virtual Worlds. *Computer Resources for Language Learning* 3, 75–81 (2010)
12. YBM Sisa.com Co., eSLS (e-Speaking Learning System), <http://ybmsisa.com>
13. Jung-Chul Cyber Institute, <http://cyber.jungchul.com>
14. Sam-Yuk Language Schoo, <http://www.sda.co.kr>

Author Index

- Abu Bakar, Zainab 106
Ahmad, Mohd Sharifuddin 221
Al-Sewari, AbdulRahman A. 1
Amirmazaheri, Amireh 343
An, Fengqi 28
Ang, Karen 195
Araki, Kenji 251
Atif, Amara 229
- Beydoun, Ghassan 12
Bilgin, Ayse 350
Boulemden, Ahmed 140
Busch, Peter 229, 343
- Cai, Xiongcai 28
Cao, Tru H. 94
Chen, Tsung Teng 335
Chen, Zhao 58
Chua Jr., Roland Christian 244
Chung, Hyunsuk 272
Compton, Paul 118, 160, 175, 258
- Dazeley, Richard 147, 188
- Farhan, Syed Ahmad 106
Fushimi, Takayasu 298, 328
- Galgani, Filippo 118
- Halabi Echeverry, Ana Ximena 73, 350
Han, Soyeon Caren 258, 272
Hanna, Nader 43, 209
Ho, Dung T. 94
Hoffmann, Achim 118
- Ikeda, Mitsuru 251
Ishikawa, Kyohei 133
Ismail, Shahrinaz 221
- Jacobson, Michael J. 43
Jang, Rae-Hyun 364
Jelinek, Herbert 147
Jeong, Ji-Seong 364
- Kang, Byeong Ho 58, 175, 258
Kato, Shoko 298
- Kazama, Kazuhiro 328
Kelarev, Andrei 147
Kim, Mihye 364
Kim, Myung Hee 160
Kim, Yang Sok 175, 258
Koide, Akihiro 298
Kuah, Chuen Tse 86
- Lee, Maria R. 335
- Maruatona, Omaru 188
Menzies, Tim 258
Motoda, Hiroshi 298
- Ogawa, Taisuke 251
Ohwada, Hayato 133, 285
Ong, Ethel 195, 244
Othman, Siti Hajar 12
Oxley, Alan 106
- Park, Chan 364
- Richards, Deborah 43, 73, 209, 229, 350
Ryu, Seung Hwan 258
- Saito, Kazumi 298, 328
Shahriar, Md. Sumon 58
Sowmya, Arcot 28
Stranieri, Andrew 147
Suzuki, Muneou 251
- Tlili, Yamina 140
- Vamplew, Peter 188
- Wobcke, Wayne 313
Wong, Anna 313
Wong, Kuan Yew 86
- Yearwood, John 147
Yoo, Kwan-Hee 364
Yoshida, Yutaka 285
- Zamin, Norshuhani 106
Zamli, Kamal Z. 1
Ziaei, Hesam 313