

George Buchanan  
Masood Masoodian  
Sally Jo Cunningham (Eds.)

LNCS 5362

# Digital Libraries: Universal and Ubiquitous Access to Information

11th International Conference on  
Asian Digital Libraries, ICADL 2008  
Bali, Indonesia, December 2008, Proceedings

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Alfred Kobsa

*University of California, Irvine, CA, USA*

Friedemann Mattern

*ETH Zurich, Switzerland*

John C. Mitchell

*Stanford University, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

Oscar Nierstrasz

*University of Bern, Switzerland*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*University of Dortmund, Germany*

Madhu Sudan

*Massachusetts Institute of Technology, MA, USA*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max-Planck Institute of Computer Science, Saarbruecken, Germany*

George Buchanan Masood Masoodian  
Sally Jo Cunningham (Eds.)

# Digital Libraries: Universal and Ubiquitous Access to Information

11th International Conference on  
Asian Digital Libraries, ICADL 2008  
Bali, Indonesia, December 2-5, 2008  
Proceedings

Volume Editors

George Buchanan  
Swansea University, Department of Computer Science  
Singleton Park, Swansea SA2 8PP, UK  
E-mail: g.r.buchanan@swansea.ac.uk

Masood Masoodian  
Sally Jo Cunningham  
The University of Waikato, Department of Computer Science  
Hamilton, New Zealand  
E-mail: {m.masoodian, sallyjo}@cs.waikato.ac.nz

Library of Congress Control Number: 2008939868

CR Subject Classification (1998): H.3, H.2, H.4.3, H.5, J.7, D.2, J.1, I.7

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

ISSN 0302-9743  
ISBN-10 3-540-89532-9 Springer Berlin Heidelberg New York  
ISBN-13 978-3-540-89532-9 Springer Berlin Heidelberg New York

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

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2008  
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper SPIN: 12567447 06/3180 5 4 3 2 1 0



# Preface

This book constitutes the refereed proceedings of the 11th International Conference on Asian Digital Libraries (ICADL 2008) held in Bali, Indonesia, in December 2008. The objective of this conference series is to provide a forum for presentation of high-quality research in the field of digital libraries. ICADL 2008 provided an opportunity for digital libraries researchers and practitioners in the Asia Pacific area and beyond to gather to explore ideas, exchange and share experiences, and further build the research network in this region. ICADL 2008 was a truly international event, with presenters from 21 countries.

A total of 63 papers were accepted for inclusion in the proceedings: 30 full papers, 20 short papers, and extended abstracts of 13 posters. Submissions were subject to a rigorous, blind peer-review process. The research topics cover the spectrum of digital libraries, including multimedia digital libraries, usability and evaluation, information retrieval, ontologies, social tagging, metadata issues, multi- and cross-language retrieval, digital preservation, scholarly publishing and communities, and more. Additionally, three tutorials were offered in association with the conference by Andreas Rauber (Vienna University of Technology), David Bainbridge (University of Waikato), and George Buchanan (Swansea University).

December 2008

Sally Jo Cunningham  
Rolly Intan  
George Buchanan  
Edna Reid

# Organization

## Organizing Committee

Conference Chairs	Zainal A. Hasibuan
Vice Chairs	Liauw Toong Tjiek, Aditya Nugraha, Woro Titi Haryanti
Program Chairs	Sally Jo Cunningham, Rolly Intan, George Buchanan, Edna Reid
Tutorial Chair	Mohammad Aries
Publication Chair	Rolly Intan
Publicity Chairs	Henny Widyaningsih, Anita Nusantari
Local Arrangements Chairs	Kalarensi Naibaho, Adi Wibowo
Proceedings	Masood Masoodian

## Program Committee

Trond Aalberg	Norwegian University of Science and Technology
Maristella Agosti	University of Padua
Robert Allen	Drexel University
Thomas Baker	Kompetenzzentrum Interoperable Metadaten
José Borbinha	Instituto Superior Técnico / INESC-ID
Christine Borgman	UCLA
Pavel Braslavski	Yandex
George Buchanan	Swansea University
Tru Cao	Ho Chi Minh City University of Technology
Hsinchun Chen	University of Arizona
Hsin-His Chen	National Taiwan University
Pu-Jen Cheng	National Taiwan University
Youngok Choi	The Catholic University of America
Gobinda Chowdhury	University of Technology Sydney
Parisa Eslambolchilar	Swansea University
Schubert Foo	Nanyang Technological University
Edward A. Fox	Virginia Tech
Richard Furtuta	Texas A&M University
Dion Goh	Nanyang Technological University
Jane Greenberg	University of North Carolina at Chapel Hill
Preben Hansen	Swedish Institute of Computer Science
Md Maruf Hasan	Shinawatra University
Geneva Henry	Rice University
Jieh Hsiang	National Taiwan University

Rolly Intan	Petra Christian University, Indonesia
Peter Jacso	University of Hawaii
Ji-Hoon Kang	Chungnam National University
Dongwon Lee	The Pennsylvania State University
Ee-Peng Lim	Singapore Management University
Jyi-Shane Liu	National Chengchi University
Yan Quan Liu	Southern Connecticut State University
Fernando Loizides	Swansea University
Akira Maeda	Ritsumeikan University
András Micsik	MTA SZTAKI
Nobuko Miyairi	Thomson Reuters
Jin-Cheon Na	Nanyang Technological University
Takashi Nagatsuka	Tsurumi University
Ekawit Nantajeewarawat	SIIT, Thammasat University
Michael L. Nelson	Old Dominion University
Erich J. Neuhold	University of Vienna
Liddy Nevile	La Trobe University
Shiyan Ou	University of Wolverhampton
Pimrumpai Premssmit	Chulalongkorn University
Uta Priss	Napier University
Edie Rasmussen	University of British Columbia
Andreas Rauber	Vienna University of Technology
Robert Sanderson	University of Liverpool
Hideyasu Sasaki	Ritsumeikan University, New York State Bar
Rudi Schmiede	Darmstadt University of Technology
Frank Shipman	Texas A&M University
Praditta Siripan	National Science and Technology Development Agency
Shigeo Sugimoto	University of Tsukuba
Hussein Suleman	University of Cape Town
Taro Tezuka	Ritsumeikan University
Yin-Leng Theng	Nanyang Technological University
M. Felisa Verdejo	National Distance Learning University
Jenq-Haur Wang	National Taipei University of Technology
Chunxiao Xing	Tsinghua University
Christopher C. Yang	Drexel University
Marcia Lei Zeng	Kent State University
Justin Zhan	Carnegie Mellon University
Ming Zhang	Peking University

## Host Institutions

The National Library of Indonesia, Jakarta  
University of Indonesia, Jakarta  
Petra Christian University, Surabaya

# Table of Contents

## Full Papers

DL2Go: Editable Digital Libraries in the Pocket . . . . .	1
<i>Hyunyoung Kil, Wonhong Nam, and Dongwon Lee</i>	
Hierarchical Classification of Web Pages Using Support Vector Machine . . . . .	12
<i>Yi Wang and Zhiguo Gong</i>	
The Prevalence and Use of Web 2.0 in Libraries . . . . .	22
<i>Alton Yeow Kuan Chua, Dion Hoe-Lian Goh, and Chei Sian Lee</i>	
Usability of Digital Repository Software: A Study of DSpace Installation and Configuration . . . . .	31
<i>Nils Körber and Hussein Suleman</i>	
Developing a Traditional Mongolian Script Digital Library . . . . .	41
<i>Garmaabazar Khaltarkhuu and Akira Maeda</i>	
Weighing the Usefulness of Social Tags for Content Discovery . . . . .	51
<i>Khasfariyati Razikin, Dion Hoe-Lian Goh, Chei Sian Lee, and Alton Yeow Kuan Chua</i>	
A User Reputation Model for DLDE Learning 2.0 Community . . . . .	61
<i>Fusheng Jin, Zhendong Niu, Quanxin Zhang, Haiyang Lang, and Kai Qin</i>	
Query Relaxation Based on Users' Unconfidences on Query Terms and Web Knowledge Extraction . . . . .	71
<i>Yasufumi Kaneko, Satoshi Nakamura, Hiroaki Ohshima, and Katsumi Tanaka</i>	
A Query Language and Its Processing for Time-Series Document Clusters . . . . .	82
<i>Sophoin Khy, Yoshiharu Ishikawa, and Hiroyuki Kitagawa</i>	
Ontology Construction Based on Latent Topic Extraction in a Digital Library . . . . .	93
<i>Jian-hua Yeh and Naomi Yang</i>	
Towards Intelligent and Adaptive Digital Library Services . . . . .	104
<i>Md Maruf Hasan and Ekawit Nantajeewarawat</i>	

Searching for Illustrative Sentences for Multiword Expressions in a Research Paper Database .....	114
<i>Hidetsugu Nanba and Satoshi Morishita</i>	
Query Transformation by Visualizing and Utilizing Information about What Users Are or Are Not Searching .....	124
<i>Taiga Yoshida, Satoshi Nakamura, Satoshi Oyama, and Katsumi Tanaka</i>	
Language Independent Word Spotting in Scanned Documents .....	134
<i>Sargur N. Srihari and Gregory R. Ball</i>	
Focused Page Rank in Scientific Papers Ranking .....	144
<i>Mikalai Krapivin and Maurizio Marchese</i>	
Scientific Journals, Overlays and Repositories: A Case of Costs and Sustainability Issues .....	154
<i>Panayiota Polydoratou and Martin Moyle</i>	
A Collaborative Filtering Algorithm Based on Global and Domain Authorities .....	164
<i>Li Zhou, Yong Zhang, and Chun-Xiao Xing</i>	
Complex Data Transformations in Digital Libraries with Spatio-Temporal Information .....	174
<i>Bruno Martins, Nuno Freire, and José Borbinha</i>	
Sentiment Classification of Movie Reviews Using Multiple Perspectives .....	184
<i>Tun Thura Thet, Jin-Cheon Na, and Christopher S.G. Khoo</i>	
Scholarly Publishing in Australian Digital Libraries: An Overview .....	194
<i>Bhojaraju Gunjal, Hao Shi, and Shalini R. Urs</i>	
Utilizing Semantic, Syntactic, and Question Category Information for Automated Digital Reference Services .....	203
<i>Palakorn Achananuparp, Xiaohua Hu, Xiaohua Zhou, and Xiaodan Zhang</i>	
A Collaborative Approach to User Modeling for Personalized Content Recommendations .....	215
<i>Heung-Nam Kim, Inay Ha, Seung-Hoon Lee, and Geun-Sik Jo</i>	
Using a Grid for Digital Preservation .....	225
<i>José Barateiro, Gonçalo Antunes, Manuel Cabral, José Borbinha, and Rodrigo Rodrigues</i>	
A User-Oriented Approach to Scheduling Collection Building in Greenstone .....	236
<i>Wendy Osborn, David Bainbridge, and Ian H. Witten</i>	

LORE: A Compound Object Authoring and Publishing Tool for the Australian Literature Studies Community . . . . .	246
<i>Anna Gerber and Jane Hunter</i>	
Consolidation of References to Persons in Bibliographic Databases . . . . .	256
<i>Nuno Freire, José Borbinha, and Bruno Martins</i>	
On Visualizing Heterogeneous Semantic Networks from Multiple Data Sources . . . . .	266
<i>Maureen, Aixin Sun, Ee-Peng Lim, Anwitaman Datta, and Kuiyu Chang</i>	
Using Mutual Information Technique in Cross-Language Information Retrieval . . . . .	276
<i>Syandra Sari and Mirna Adriani</i>	
Exploring User Experiences with Digital Library Services: A Focus Group Approach . . . . .	285
<i>Kaur Kiran and Diljit Singh</i>	
Beyond the Client-Server Model: Self-contained Portable Digital Libraries . . . . .	294
<i>David Bainbridge, Steve Jones, Sam McIntosh, Ian H. Witten, and Matt Jones</i>	
<b>Short Papers</b>	
New Era New Development: An Overview of Digital Libraries in China . . . . .	304
<i>Guohui Li and Michael Bailou Huang</i>	
Browse&Read Picture Books in a Group on a Digital Table . . . . .	309
<i>Jia Liu, Keizo Sato, Makoto Nakashima, and Tetsuro Ito</i>	
Towards a Webpage-Based Bibliographic Manager . . . . .	313
<i>Dinh-Trung Dang, Yee Fan Tan, and Min-Yen Kan</i>	
Spacio-Temporal Analysis Using the Web Archive System Based on Ajax . . . . .	317
<i>Suguru Yoshioka, Masumi Morii, Shintaro Matsushima, and Seiichi Tani</i>	
Mining a Web2.0 Service for the Discovery of Semantically Similar Terms: A Case Study with Del.icio.us . . . . .	321
<i>Kwan Yi</i>	
Looking for Entities in Bibliographic Records . . . . .	327
<i>Trond Aalberg and Maja Žumer</i>	

Protecting Digital Library Collections with Collaborative Web Image Copy Detection . . . . .	331
<i>Jenq-Haur Wang, Hung-Chi Chang, and Jen-Hao Hsiao</i>	
Enhancing the Literature Review Using Author-Topic Profiling . . . . .	335
<i>Alisa Kongthon, Choochart Haruechaiyasak, and Santipong Thaiprayoon</i>	
Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools . . . . .	339
<i>Choochart Haruechaiyasak and Chaianun Damrongrat</i>	
On Developing Government Official Appointment and Dismissal Databank . . . . .	343
<i>Jyi-Shane Liu</i>	
An Integrated Approach for Smart Digital Preservation System Based on Web Service . . . . .	347
<i>Chao Li, Ningning Ma, Chun-Xiao Xing, and Aorong Jiang</i>	
Personalized Digital Library Framework Based on Service Oriented Architecture . . . . .	351
<i>Li Dong, Chun-Xiao Xing, Jin Lin, and Kehong Wang</i>	
Automatic Document Mapping and Relations Building Using Domain Ontology-Based Lexical Chains . . . . .	355
<i>Angrosh M.A. and Shalini R. Urs</i>	
A Paper Recommender for Scientific Literatures Based on Semantic Concept Similarity . . . . .	359
<i>Ming Zhang, Weichun Wang, and Xiaoming Li</i>	
Network of Scholarship: Uncovering the Structure of Digital Library Author Community . . . . .	363
<i>Monica Sharma and Shalini R. Urs</i>	
Understanding Collection Understanding with Collage . . . . .	367
<i>Sally Jo Cunningham and Erin Bennett</i>	
Person Specific Document Retrieval Using Face Biometrics . . . . .	371
<i>Vikram T.N., Shalini R. Urs, and K. Chidananda Gowda</i>	
The Potential of Collaborative Document Evaluation for Science . . . . .	375
<i>Jöran Beel and Béla Gipp</i>	
Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres . . . . .	379
<i>Robert B. Allen, Ilya Waldstein, and Weizhong Zhu</i>	

Arabic Manuscripts in a Digital Library Context . . . . .	387
<i>Sulieaman Salem Alshuhri</i>	

## Posters

Discovering Early Europe in Australia: The <i>Europa Inventa</i> Resource Discovery Service . . . . .	394
<i>Toby Burrows</i>	
Mapping the Question Answering Domain . . . . .	396
<i>Mohan John Blooma, Alton Yeow Kuan Chua, and Dion Hoe-Lian Goh</i>	
A Scavenger Grid for Intranet Indexing . . . . .	398
<i>Ndapandula Nakashole and Hussein Suleman</i>	
A Study of Web Preservation for DMP, NDAP, and TELDAP, Taiwan . . . . .	400
<i>Shu-Ting Tsai and Kuan-Hua Huang</i>	
Measuring Public Accessibility of Australian Government Web Pages . . .	402
<i>Yang Sok Kim, Byeong Ho Kang, and Raymond Williams</i>	
Named Entity Recognition for Improving Retrieval and Translation of Chinese Documents . . . . .	404
<i>Rohini K. Srihari and Erik Peterson</i>	
Current Approaches in Arabic IR: A Survey . . . . .	406
<i>Mohammed Mustafa, Hisham AbdAlla, and Hussein Suleman</i>	
A Bilingual Information Retrieval Thesaurus: Design and Value Addition with Online Lexical Tools . . . . .	408
<i>K.S. Raghavan and A. Neelameghan</i>	
Entity-Based Classification of Web Page in Search Engine . . . . .	410
<i>Yicen Liu, Mingrong Liu, Liang Xiang, and Qing Yang</i>	
MobiTOP: Accessing Hierarchically Organized Georeferenced Multimedia Annotations . . . . .	412
<i>Thi Nhu Quynh Kim, Khasfariyati Razikin, Dion Hoe-Lian Goh, Quang Minh Nguyen, Yin Leng Theng, Ee-Peng Lim, Aixun Sun, Chew Hung Chang, and Kalyani Chatterjea</i>	
Social Tagging in Digital Archives . . . . .	414
<i>Shih-Yuarn Chen, Yu-Ying Teng, and Hao-Ren Ke</i>	
Editor Networks and Making of a Science: A Social Network Analysis of Digital Libraries Journals . . . . .	416
<i>Monica Sharma and Shalini R. Urs</i>	



Empowering Doctors through Information and Knowledge.....	418
<i>Anjana Chattopadhyay</i>	
<b>Author Index</b> .....	421

# DL2Go: Editable Digital Libraries in the Pocket

Hyunyoung Kil, Wonhong Nam, and Dongwon Lee\*

The Pennsylvania State University  
University Park, PA 16802, USA  
{hkil,wnam,dongwon}@psu.edu

**Abstract.** A preliminary framework, termed as DL2Go, that enables *editable* and *portable* personal digital libraries is presented. For mobile offline users of digital libraries, DL2Go can: (1) package digital libraries into mobile storage devices such as flash drives, along with needed application softwares (e.g., wiki and DBMS), (2) (de-)compress contents of digital libraries to address storage constraints of mobile users when needed, (3) enables users to add, delete, and update entities of digital libraries using wiki framework, and (4) share/sync edited contents with other DL2Go users and the server using web services and RSS framework.

## 1 Introduction

Due to the significant improvement in underlying technologies, strong support from governments (e.g., NSF NSDL, DELOS), and rapid increase in its user base, *Digital Libraries* (DLs hereafter) have become a norm, reaching out a large number of audiences in cyberspace. For instance, there are currently about 930 DLs at NSDL<sup>1</sup> for diverse audiences, ranging from kids to K-12 educators to academic scholars. Although abundant, however, in this paper, we claim that the current support of DLs is not without a problem. In particular, we believe that current DLs are not well designed for *nomadic users* who want to use DLs in *offline environment*.

**Example 1.** Consider the following motivating scenarios: (1) Together with students, a K-12 teacher “John” leads a bird watching trip to a remote park. He prepares his laptop computer with a CD-ROM encyclopedia including full of photos and descriptions of various birds. However, when they want to add their comments about birds spotted, they have to write on papers because the encyclopedia does not allow them to write comments. Upon returning from the trip, “John” realizes that it is hard to gather students’ comments and share them with other K-12 educators; (2) A computer science professor “Sue” needs to finish out the draft of her paper while she flies from Los Angeles to New York. However, in the airplane, she realizes that she does not have access to literatures that she needs to read and cite; (3) An officer “Kim” at a non-profit NGO travels through backcountries in Africa to distribute a new version of the

---

\* Partially supported by IBM and Microsoft gifts.

<sup>1</sup> <http://crs.nsdll.org/collection/>

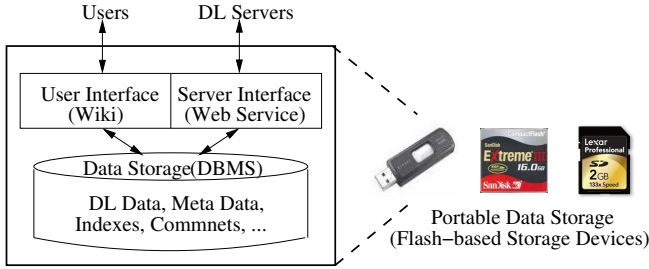


Fig. 1. The overview of DL2Go framework

children’s book of Encyclopedia to children who use laptops from “One Laptop per Child”<sup>2</sup>. However, she cannot carry her laptop computer nor external devices because of too many baggages. Moreover, due to the weak network infrastructure and lack of CD-ROM drives, the only option is to store the online book in a USB flash drive. □

Tasks sketched in these examples cannot be resolved easily in conventional DL frameworks where users are assumed to have stable network access to DLs and majority of users’ operations is only “reading” the contents of DLs. To address these new needs, therefore, we have developed a novel framework and prototypes, termed as DL2Go, with the following desiderata in mind: (1) DL2Go is easy to carry around for nomadic users, (2) DL2Go is cheap to produce and distribute, (3) DL2Go supports direct update on the contents of DLs in both personal and collaborative environments, (4) DL2Go supports the sharing of contents of DLs with other users, and (5) (in addition) DL2Go supports all regular features of DLs such as archiving, indexing, searching, and browsing of contents.

## 2 The DL2Go Framework

Figure 1 illustrates the preliminary design of DL2Go. Consider a DL  $X$  that users may have access to (e.g., arXiv [6]). Initially, users download the current snapshot of  $X$  and package it into a *portable data storage*. Once building a personal copy of  $X$ , users can browse, search, read, and even add, delete, and update the contents of  $X$  freely via the *editable user interface* of DL2Go. Users can *sync* the contents of  $X$  in a portable storage with ones in the server. Next, we present various issues that we have considered in implementing the prototypes of DL2Go framework.

### 2.1 Portable Data Storage

We envision that users of DL2Go have some forms of electronic devices (e.g., laptop, PDA, smartphone) with limited space in the primary storage area. Since

<sup>2</sup> <http://laptop.org/>

DL2Go aims to be a portable DL in a mostly offline environment, we consider portable devices to host contents of DLs. As a secondary data storage, popular choices include external HDDs, Solid-State Drives (SSDs), CDs/DVDs, and flash-based storage devices. External HDDs or more recently-developed SSDs can store a large amount of data for relatively low price, but are bulkier to carry around for DL2Go users. Although CDs/DVDs are affordable and a popular device in many settings, they are not appropriate for data with frequent updates. Even if there are writable CDs/DVDs, their usage is still limited. Furthermore, not all machines that we have in mind come with a reading device for CDs/DVDs. Finally, flash-based storage devices are small and supported in many electronic devices. Furthermore, with the rapid advancement in flash technologies, more and more affordable flash-based devices with an ample space become available. These days, it is not difficult to find a USB flash drive with 32GB space for less than \$100. Among all the candidates, therefore, we decide to use flash-based storage devices. Figure 1 shows three popular commercial examples – USB Flash Drive, CompactFlash, and Secure Digital. Even though the I/O speed of current flash-based storage devices tends to be slower than that of HDDs (e.g., 30MB/sec vs. 100MB/sec for read/write), we believe that running DLs in flash-based devices is still viable.

## 2.2 Data Management

DLs can contain various types of data as their contents – text/multimedia data objects, metadata, index, comments, etc. Depending on the characteristics of the DL data, different data management system can be selected. For DLs with a relatively small amount of data (e.g., DLs with textual data only), one may use the file system with the support of OS to manage the contents of DLs – i.e., data are stored as files in the hierarchy of folders. Although simple and appropriate for small DLs, however, this approach does not work well for DLs since a large amount of storage space is wasted. For instance, MS Windows file system allocates at least 4KB for a file even if it includes only 1 character. If we were to store all 1 million paper metadata information in DBLP as files, then it would require roughly 4GB just for the metadata information without the PDF copies of papers (see Table 2), making it inapplicable for DLs with a large amount of (multimedia) data. Furthermore, file systems support only limited capability for indexing and searching data. Therefore, in DL2Go, we decide to use a RDBMS as the underlying data management system. Most commercial RDBMS can optimize space and fragmentation issues by using sophisticated schemes, provide a standardized way to query, and is equipped with efficient query optimization techniques. In particular, in building prototypes, we use the MySQL open-source RDBMS, modified to run on flash devices.

In addition, to address the issue of large-size DLs further, one can also use various *compression* techniques, including database compression [5], PDF file compression [18], and general-purpose file compression [25,19]. To select one among the compression techniques above, we carry out preliminary experiment. We have tested 1,000 PDF files (247 MB) randomly selected from arXiv

**Table 1.** Experimental result for PDF file compression

Tool	PDF compression			General purpose compression		
	Verypdf	CVista	Magic	WinZip	WinRK	7Zip
Compression ratio	0.99	1.01	0.98	1.18	1.23	1.24
Avg cmp time (in sec.)	0.48	4.77	1.25	0.39	0.67	0.37

**gr-qc** with three PDF file compression tools (Verypdf Advanced PDF Tools 2.0 [22], CVista PdfCompressor 4.0 [7] and Magic PDF compressor 2.2 [17]) and three general purpose compression tools (WinZip Pro 11.1 [25], WinRK 3.0.3b Normal compression [16] and 7-Zip 4.57 [19]). Table 1 presents the compression ratio and the average compression time in second for each tool, where  $\text{the compression ratio} = \frac{\text{the uncompressed size}}{\text{the compressed size}}$ . In our experiment, PDF compression tools did not achieved better results than general purpose compression tools (even two of them increased the total size). With this result, we believe that the PDF compression tools are not appropriate for research papers. 7-Zip finally showed the best performance (the compression ratio is 1.24—it has saved 19.54% for space). Therefore, in our prototypes, we used 7-ZIP as the compression method. We leave tighter and more efficient integration of the compression scheme in the data management layer of DL2Go to the future work.

### 2.3 Editable User Interface

The role of users in DLs has changed from the passive read-only users to active participatory ones of Web 2.0 era. These days, many DLs not only allow users to customize the look and feel of interfaces, but also encourage users’ input such as commenting on contents, correcting errors, and adding missing information. Moreover, due to its pervasive usage, most DLs are designed to have web interface of its contents. That is, using hyperlinks and images on web pages, users can search and browse the contents of DLs. To accommodate such needs, via the *wiki* system, DL2Go is designed to have *editable* web interface to all contents of DLs. As stated [3] as: “A wiki is software that allows users to create, edit, and link web pages easily. Wikis are often used to create collaborative websites and to power community websites.”, the wiki system is primarily used to create and edit web pages for multi-user web environment. In DL2Go, however, we adapt this wiki system as the user interface of DL2Go running on flash devices for mainly personal DL users. However, when needs arise, DL2Go can still handle multi-user collaborative tasks. For instance, when multiple computers share one server machine or storage device in a underdeveloped country, one can plug DL2Go into a USB port of the server and do collaborative editing on some topics. In addition, we take advantage of built-in index and search capability of the modern wiki system (although we allow direct manipulation of stored data objects in the underlying RDBMS).

<sup>3</sup> <http://en.wikipedia.org/wiki/Wiki>

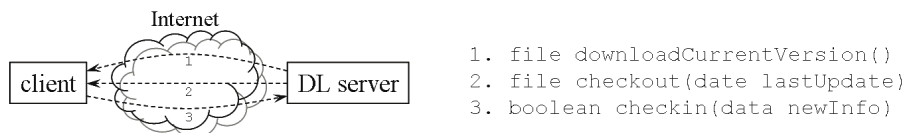


Fig. 2. Interface with a DL server by web services

## 2.4 Interface with Server

Our DL2Go requires two kinds of interactions with the DL server; (1) *Downward interface*: Ideally, the server provides standard-based interfaces for users to download data – e.g., OAI-PMH<sup>4</sup> for metadata harvesting, OAI-ORE<sup>5</sup> for exchanging information about digital objects in DLs, RSS-based update syndication, and FTP-based direct download; and (2) *Upward interface*: When users make updates on contents in DL2Go and need to upload them to the server (for archiving or sharing), updated information can be uploaded to the server and merged with the rest of contents uniformly. Such a case can occur when an archeologist comes back from the field trip with full of comments recorded in her DL2Go based DL, and wants to share her updated DL contents with other archeologists. Although the current support of DLs for downward interface is abundant, we believe that one for upward interface is rather limited. To address this issue, in the design of DL2Go, we propose to use the framework of web services [23] to implement the upward channel.

*Web services* are software systems designed to support machine-to-machine interoperation over the Internet. By using the standard XML-based web services framework, we can have a programmable interface to a DL server so that not only human users but also software agents can build DL2Go easily. Figure 2 illustrates an example for the interface with a DL server by web services, which we propose. For the interface, three web service operations are provided as follows: `downloadCurrentVersion` allows clients to download the current snapshot of the DL, and by `checkout` operation the client is able to obtain new data since the last update. Moreover, `checkin` operation provides the upward interface by which users can upload the information they want to share.

## 3 Case Studies

To demonstrate the soundness of our proposed idea, we have built two prototypes of DL2Go based DLs in the domain of bibliographic DLs—“DL2Go for arXiv” and “DL2Go for DBLP.” Table 2 shows the detailed statistics of two DLs that we used. The subset of arXiv [6] data set we used includes the information about 78,146 papers and 35,803 authors in Physics discipline. Often, the venue information of papers is missing in the arXiv data set. It contains PDF files for

<sup>4</sup> <http://www.openarchives.org/pmh/>

<sup>5</sup> <http://www.openarchives.org/ore/>

**Table 2.** Statistics of two data sets

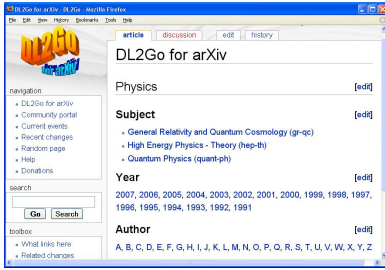
DL	papers	authors	conf./jour.	tuples in RDBMS	total size (in MB)
arXiv	78,146	35,803	N/A	114,729	234
DBLP	925,494	569,227	9,331/41,808	1,558,625	2007

most papers. On the other hand, the DBLP [15] data set includes the information about 925,494 papers (580,509 conference papers and 344,985 journal papers) and 569,227 authors in mainly Computer Science discipline. Among venues, there are 9,331 conferences and 41,808 journals. We have separately counted each year of conferences and each volume(number) of journals. Note that DBLP data set contains only bibliographic metadata without actual PDF or PS files. As the software and hardware systems, we have used the WOS Portable[4] package (that includes Apache 2.2.4, MySQL 5.0.41, and PHP 5.2.3), MediaWiki 1.10.1, and a USB flash drive with 4GB space. All features, except the upward channel using web services, are currently implemented in the prototypes.

### 3.1 DL2Go for arXiv

The arXiv [6] is one of the most popular scientific DLs in the fields of physics, mathematics, computer science, and so on. Although the number of archived publications in the arXiv is smaller than that of other scientific DLs, since it is self-archived by authors, the quality of associated metadata in the arXiv is excellent. In addition, since it often contains publication files in various formats (e.g., PS, PDF, DOC, TeX), the amount of the required storage for data and metadata is substantial. Since the arXiv does not provide the snapshot of their collection for download, we use the subset of arXiv data downloaded from a repository by Karczmarek [11]. This subset is created to be useful for anyone who wants to have a local copy of the arXiv, and contains the collections of the PDF-formatted papers and index files in three fields of Physics (i.e., General Relativity and Quantum Cosmology (**gr-qc**), Quantum Physics (**quant-ph**) and High Energy Physics - Theory (**hep-th**)) from 1991 to 2007.

By these index files, we reorganized the paper metadata, and created the main wiki page of DL2Go for arXiv (see Figure 3(a)) where users could browse through papers based on fields, years, and authors, mimicking the interface of the original arXiv web site. Although stored as tuples of various tables in the underlying RDBMS, conceptually, DL2Go for arXiv is comprised of a hierarchy of wiki pages. There is a wiki page for each paper at the bottom of the hierarchy while various kinds of intermediate internal paths in the hierarchy are also mapped to wiki pages. By following different internal paths, for instance, users can browse papers based on fields (i.e., **gr-qc**, **quant-ph** and **hep-th**), year (i.e., 1991-2007) or author names. An example of a wiki page for an internal path on author names is shown in Figure 3(b) whose role is to guide users to the correct wiki page. When a field is selected on the main wiki page, users need to choose a specific year



(a) Main page



(b) Internal path page on author

**Fig. 3.** Screen-shots of “DL2Go for arXiv”

and month. Then, the wiki page containing the paper list of the field submitted in the period shows up. Through the path by year, DL2Go displays the paper list submitted in the month of the year selected, regardless of fields. Through the path by author, users can find out author names in alphabetic order, and finally we can obtain the list of papers written by the author.

A wiki page for each paper at the bottom of the hierarchy includes metadata of the paper (e.g., title, authors, subject, abstract, etc.), and links to the paper file and to user’s comment (see Figure 4). Through this page, users can read the paper using a PDF viewer, edit the metadata of the paper information (e.g., correcting a typo in author names), add comments, and search other papers by keywords. Note that searching by keywords is available in any wiki pages. In addition to (or in stead of) the PDF versions of paper files, DL2Go may also provide text files, which contain the contents extracted from the original PDF paper files, for space and user flexibility. In general, the size of text files is much smaller than that of the original PDF files—e.g., 79,673 text and PDF files consume about 2.8GB and 17.5GB, respectively. In DL2Go, users can choose which versions of files to package, according to their preference.

One of unique features of DL2Go is that users can leave their comments for any papers (or any wiki pages). While reading a paper, for instance, users can add their own review for the paper in the corresponding wiki page. In addition to user’s comments, every data including paper context and metadata in our DL2Go are editable by users. While we read papers or books, it is not unusual to find out wrong or missing data as well as a simple typo, but it is not easy for readers to correct data and share it with other users in conventional DLs. Since our DL2Go adopts the wiki system as a user interface, users can edit all pages easily. In addition, through the server interface, the updated information can be reported to DL servers easily.

### 3.2 DL2Go for DBLP

The DBLP (Digital Bibliography & Library Project) [15] provides bibliographic information on major Computer Science journals and proceedings. Our DL2Go



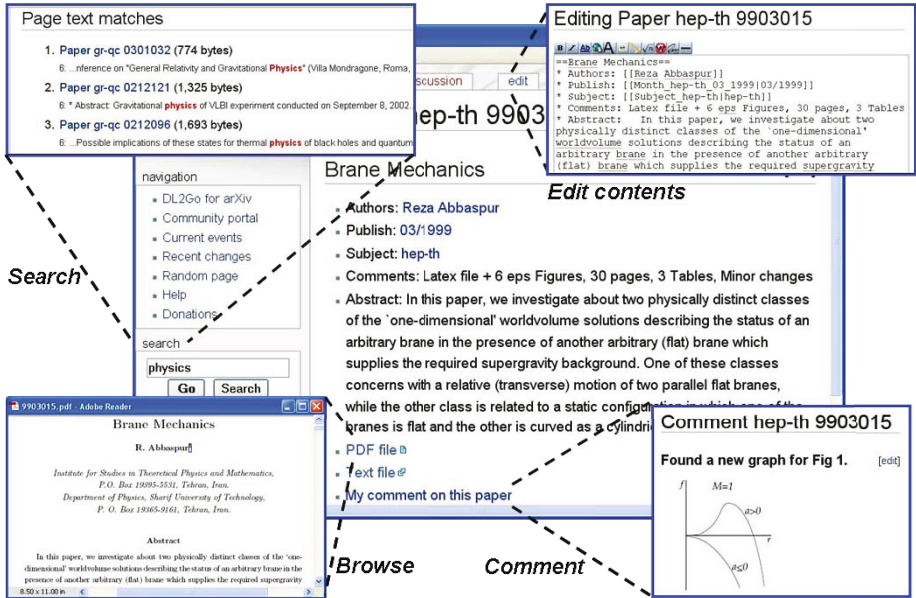
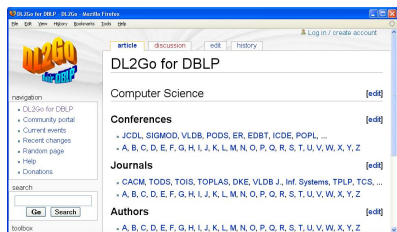


Fig. 4. Functions on paper page of “DL2Go for arXiv”

for DBLP is built with all metadata for conferences, conference papers, journal papers in XML files downloaded from the DBLP web site [15]. Similar to the implementation of DL2Go for arXiv, we reorganize the parsed data and generate wiki pages for the main, three internal paths (i.e., by conferences, journals, and authors), and papers. Figure 5 illustrates screen-shots for the main page and an internal path on conference. On the main page (Figure 5(a)), a paper can be easily searched by conference, journal, and author names. When users choose a conference/journal name and a specific year/volume(number), a wiki page for the selected conference/journal in that year/volume is displayed. The page includes the detailed conference/journal information (e.g., its full name, the location and date, the homepage URL) and a link to the list of the papers published in the conference/journal (see Figure 5(b)). Then, users can browse a page containing the paper list via the link, and finally obtain a paper page from the list. As DL2Go for arXiv, users are able to find author names in alphabetic order, and obtain the list of papers written by the author. A wiki page for each paper includes its title, authors, conference/journal information, and links to an electronic file for the paper and to user’s comment. Since DBLP does not provide PDF nor PS files for papers, DL2Go for DBLP does not support them either. However, DL2Go has a link to a paper file as a blank so that users can add the files once they obtain files from other resources. Every data in DL2Go for DBLP is editable as DL2Go for arXiv.



(a) Main page



(b) Internal path page on conference

Fig. 5. Screen-shots of “DL2Go for DBLP”

## 4 Related Work

Several personal DL systems are related to our DL2Go framework. Both of [26] and [13] are open-source projects for extensible digital library platforms. The former project provides a new way to build digital library collections and to publish them on the Internet or CD-ROMs, and the latter aims to create flexible, collaborative digital libraries based on the previous Fedora project [14]. In the Personal Libraries system of Berkeley Digital Library [24] which is largely premised on a distinction between collection (a specification of some set of documents) and repository (a means of storing and retrieving individual documents), users can create their own collections which are built from materials extracted from a very large document collection. The UpLib system [10] is specially designed for secure use by a single individual and is extended for collaborative operation of multiple UpLib repositories [9]. In the Salticus system [3], each user can build a personal digital library by collecting documents and generalizing user’s choice based on a user’s interest. They, however, do not support editable user interfaces and/or consider portability issues. [12] discusses user interface issues to browse, edit, store, and annotate DL resources but do not consider the wiki system.

A few recent studies [2,8,20,21,11] consider portability issue for DLs. Some [2,8] of them present DL systems where users can access through wireless network and/or mobile phones. [2] explores how users can be given access to digital information while they are mobile. Mobile G-Portal study [21] shows that a group of mobile devices can be used as learning assistant tools to support collaborative sharing and learning for geography fieldwork. [20] has developed indexing technology to search Wikipedia on a conventional CD. Bainbridge et al. have built a self-contained digital library on an iPod [1].

A number of efforts [18,19] are made to save storage space for (PDF document) data files and database compression [5]. MaximumCompression [18] has studied PDF file compression, and compressed 1 file (4,526,946 bytes) with a number of different compressors/archivers (195). The result is that WinRK [16] performs best with 21.60% saving for space. In addition, we have experimented on a huge number of PDF files (1,000 files, 247 MB) using several compression tools. 7-Zip [19] shows the best performance with 19.54% saving.

## 5 Conclusion and Future Work

The novel framework, termed as DL2Go, that supports *editable* and *portable* DLs for nomadic users on the road is presented. Since the whole DL2Go system and the contents of a DL are stored in a tiny “in-the-pocket” flash-based storage device, our proposal can be useful for users who have to work in an offline environment where the network access is scarce and storage space is limited (e.g., scientists in the remote fields).

Ample directions are ahead for future work. We are working on implementing an efficient upward interface to the DL server using Web services. We believe that this interface deserves further exploration for interactive DLs. For data (de)compression, we need to build complete integration of the compression scheme in the data management layer of DL2Go.

## References

1. Bainbridge, D., Jones, S., McIntosh, S., Jones, M., Witten, I.H.: Portable digital libraries on an iPod. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 333–336 (2008)
2. Bhargava, B., Annamalai, M., Pitoura, E.: Digital library services in mobile computing. ACM SIGMOD Record 24(4), 34–39 (1995)
3. Burke, R.D.: Salticus: guided crawling for personal digital libraries. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 88–89 (2001)
4. Software, C.H.: Webserver On Stick, <http://www.chsoftware.net/>
5. Chen, S., Nucci, A.: Nonuniform compression in databases with haar wavelet. In: Data Compression Conference (DCC), pp. 223–232 (2007)
6. Cornell. arXiv.org, <http://www.arxiv.org/>
7. CVISION Technologies, Inc. Cvista pdfcompressor 4.0., <http://www.cvisiontech.com/>
8. Imai, S., Kanamori, Y., Shuto, N.: Retrieving tsunami digital library by use of mobile phones. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 525–528. Springer, Heidelberg (2007)
9. Janssen, W.C.: Collaborative extensions for the uplib system. In: JCDL, pp. 239–240 (2004)
10. Janssen, W.C., Popat, K.: UpLib: A universal personal digital library system. In: ACM symposium on Document Engineering, pp. 234–242 (2003)
11. Karczmarek, J.: The arXiv on Your Harddrive, <http://www.theory.physics.ubc.ca/arxiv/>
12. Kikuchi, H., Mishina, Y., Ashizawa, M., Yamazaki, N., Fujisawa, H.: User interface for a digital library to support construction of a virtual personal library. In: International Conference on Multimedia Computing and Systems (ICMCS), pp. 429–432 (1996)
13. Krafft, D.B., Birkland, A., Cramer, E.J.: Ncore: Architecture and implementation of a flexible, collaborative digital library. In: JCDL, pp. 313–322 (2008)
14. Lagoze, C., Payette, S., Wilper, C.: Fedora: An architecture for complex objects and their relationships. International Journal of Digital Libraries 6(2), 124–138 (2006)

15. Ley, M.: DBLP: Digital Bibliography & Library Project, <http://dblp.uni-trier.de/>
16. Software Ltd, M.: WinRK., [http://www.msoftware.co.nz/WinRK\\_about.php](http://www.msoftware.co.nz/WinRK_about.php)
17. Magicteck Software. Magic pdf compressor 2.2, <http://www.magicteck.com/>
18. MaximumCompression. Adobe Acrobat document PDF file compression test, <http://www.maximumcompression.com/data/pdf.php>
19. Pavlov, I.: 7-Zip, <http://www.7-zip.org/>
20. Potthast, M.: Wikipedia in the pocket - indexing technology for near-duplicate detection and high similarity search. In: ACM SIGIR, p. 909 (2007)
21. Theng, Y., Tan, K., Lim, E., Zhang, J., Goh, D., Chatterjea, K., Chang, C., Sun, A., Yu, H., Dang, N., Li, Y., Vo, M.: Mobile g-portal supporting collaborative sharing and learning in geography fieldwork: an empirical study. In: JCDL, pp. 462–471 (2007)
22. VeryPDF.com, Inc. Advanced pdf tools v2.0, <http://www.verypdf.com/>
23. W3C. Web Services, <http://www.w3.org/2002/ws/>
24. Wilensky, R.: Personal libraries: Collection management as a tool for lightweight personal and group document management. Technical Report SDSC TR-2001-9, San Diego Supercomputer Center (2001)
25. WinZip International, LLC: WinZip, <http://www.winzip.com/>
26. Witten, I.H., Boddie, S.J., Bainbridge, D., McNab, R.J.: Greenstone: A comprehensive open-source digital library software system. In: ACM International Conference on Digital Libraries, pp. 113–121 (2000)

# Hierarchical Classification of Web Pages Using Support Vector Machine

Yi Wang and Zhiguo Gong

Faculty of Science and Technology  
University of Macau  
Macao, PRC  
{ma66609, fstzgg}@umac.mo

**Abstract.** In this paper, a novel method for web page hierarchical classification is addressed. In our approach, SVM is used as the basic algorithm to separate any two sub-categories under the same parent node. In order to alleviate the ill shift of SVM classifier caused by imbalanced training data, we try to combine the original SVM classifier with BEV algorithm to create classifier called VOTEM. Then, a web document is assigned to a sub-category based on voting from all category-to-category classifiers. This hierarchical classification algorithm starts its work from the top of the hierarchical tree downward recursively until it triggers a stop condition or reaches the leaf nodes. And our experiment reveals that proposed algorithm obtains better results.

**Keywords:** hierarchical classification, imbalanced data, web page, SVM.

## 1 Introduction

Nowadays, the web is growing at an exponential rate and can cover almost any information needed. However, the huge amount of web makes it more and more difficult to effectively find the target information for a user. Generally two solutions exist, hierarchical browsing (such as *Yahoo*) and keyword searching (such as *Google*). Most of the famous hierarchical browsing systems organize web documents with intensive human involvements. To reduce this huge amount of labor work, an automatic classification algorithm is definitely expected and it draws much attention from the researchers. This paper is going to address techniques for automatic classification of web pages.

Many achievements have been obtained in document classification field. But more previous researches focused on flat classification and did not pay much attention to hierarchical classification. Sometimes flat method may become inferior with large amount of web pages. For example, *Alibaba.com*, a hierarchical electronic commerce system, in which top categories include ‘manufactures’, ‘companies’ and ‘suppliers’. Take ‘manufactures’ as an example, there are almost two thousands leaf categories under this category. How to effectively and efficiently organize such massive information automatically is a challenge work. With such a motivation, hierarchical organization is often preferred.

Some web directory services provide convenient accesses for the user to narrow down their browsing scope of the huge web. However, assigning web-pages into corresponding categories is time-consuming and needs plenty of manual work. To solve the problem, we use SVM (Support Vector Machine) as the fundamental technique for a binary classifier to separate any two classes. To overcome the drawback caused by imbalanced size of two classes, we combine the original SVM algorithm with BEV (Bagging Ensemble Variation) to create a different binary classifier. For all the sub-categories under a given parent category, assigning decision is made by voting from all those binary classifiers. A web page is thus assigned to corresponding category from top to down recursively along the hierarchy until triggering a give stop condition or reaching the leaf nodes.

The paper is organized as following. Section 2 outlines related work in this field. Section 3 discusses the main methodology of our system, including main algorithm of our proposed binary classifier, and voting-based multiple category classification algorithm. Section 4 discusses the results of our experiment compared with the other base algorithms. Finally in section 5, we conclude the paper.

## 2 Related Work

In past years, there have been extensive investigations and rapid progresses in automatic hierarchical classification. Basically, the models depend on the hierarchical structure of the dataset and the assignment of documents to nodes in the structure. Nevertheless, most of researches adopt a *top-down* model for classification.

Sun and Lim [7] have summarized four structures used in text classification: *virtual category tree*, *category tree*, *virtual directed acyclic category graph* and *directed acyclic category graph*. The second and forth structures are popularly used in many web directory services since the documents could be assigned to both internal and leaf categories. These two models used category-similarity measures and distance-based measures to describe the degree of falsely classification in judging the classification performance.

Dumais and Chen [3] found a boost of the accuracy for hierarchical models over flat models using a sequential Boolean decision rule and a multiplicative decision rules. They claimed only few of the comparisons are required in their approaches because of the efficiency of sequential approach. They also have shown that SVM take a good performance for virtual category tree, but category tree is not considered in their work.

Rousu et al. [6] proposed an efficient optimization algorithm, which is based on incremental conditional gradient ascent in single-example sub-spaces spanned by the marginal dual variables. The classification model is a variant of the Maximum Margin Markov Network framework, which is equipped with an exponential family defined on the edges. This method solved the scaling problem to medium-sized datasets but whether it is fit for huge amount data set is unclear.

Nicolo et al. [1] used a refined evaluation, which turns the hierarchical SVM classifier into an approximate value of the Bayes optimal classifier with respect to a simple stochastic model for the labels. They announced an improvement on hierarchical SVM algorithm by replacing its top-down evaluation with a recursive bottom-up scheme. Loss function is used to measure the performance in classification filed in this paper.

The aim of our work is to discuss the problem of imbalanced data and measurement of multi-label classification and find strategies to solve them. In the experiments, we compare the classification performance using both the standard measurements *precision/recall* and extended measurement *loss value*.

### 3 Methodology of Hierarchical Classification

In hierarchical classification, category labels in a tree or a directed acyclic graph structure are regarded as nodes of a given taxonomy. It is assumed that whenever a label contains a certain node  $i$  in a tree or graph structure of the taxonomy, it must also contain all the nodes along the path connecting the tree root to node  $i$ . Figure 1 shows an example of hierarchy structure. The gray nodes are leaf nodes which have no sibling node. Our goal is to define an effective algorithm which can assign web documents to corresponding categories starting from the root downwards recursively until a given stop condition is triggered or the leaf node is reached. In this section, we address our solution in detail.

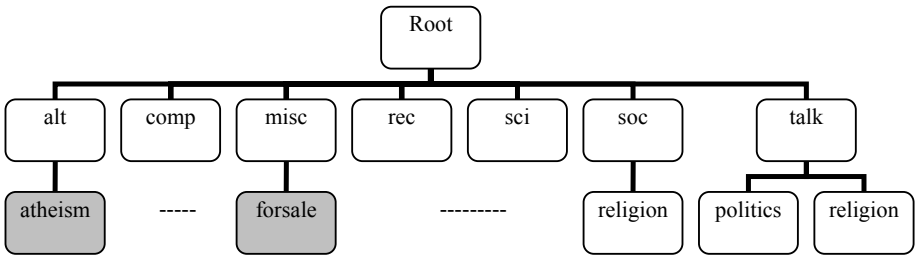


Fig. 1. A segment of the categories hierarchy (from 20newsgroup)

#### 3.1 Web Page Preprocessing

Different to traditional text documents, web pages are compiled using HTML tags. Though such tags are originally used for the purpose of presentation, they can also bring some significance information of their associated texts. We still use Term Frequency & Inverse Document Frequency (TF/IDF) as the basic model to represent web pages.

TF/IDF model could be modified with tags existed in the web pages. Two factors influence the weight of a term -- its appearing frequency and its associated tags. Let  $n$  be the total number of terms of the collection, then  $n$  dimension vector space is constructed. Each document is represented as a normalized vector  $(v_1, v_2, \dots, v_n)$ . Each dimension value is corresponding to the weight of designated term. The weight value is given as following:

$$v_i = \frac{\sum_{\alpha \in T} W_{\alpha} * tf_i^{\alpha} * \log(N / n_i)}{\sqrt{\sum_j (\sum_{\alpha} W_{\alpha} * tf_j^{\alpha} * \log(N / n_j))^2}}$$

where  $N$  is the total number of documents and  $n_i$  is the number of documents containing the  $i_{th}$  term.  $tf_i$  is the original frequency of the  $i_{th}$  term in the document.  $T$  is sets of segment content divided by different tags in web-page such as  $\langle title \rangle$ ,  $\langle a href=... \rangle \dots \langle meta name="description | keywords | classification" content="..." \rangle$  and so on. Each set  $\alpha$  is corresponding to an influenced factor  $W_\alpha$  which could be slightly adjusted in realization just necessary to satisfy the condition  $\sum_{\alpha \in T} W_\alpha = 1$ .

Generally the sequence of influenced set could be sorted by importance such as  $W_{title} > W_{meta} \dots > W_{context} > W_{image}$ . This improved TF/IDF model has been proved to be more effective in experiment.

### 3.2 SVM Binary Classifier

The Support Vector Machine, which is actually an algorithm to determine a maximal margin separating hyper-plane between two classes of data, was originally proposed by Vapnik[8]. As a binary classification algorithm, SVM gains increasingly popularity because of outstanding performance. Especially, it can tolerate the problems of high dimensions and sparse instance spaces.

The algorithm creates a maximum margin hyper-plane extremely to discriminate two classes aside separately in a high dimension. To solve the problem that the data are not exactly separable, the SVM brings in slack variables  $\xi$  and parameter  $C$ . Here,  $\xi$  provides an error estimate of the decision boundary and  $C$  is used to control balance between machine complexity and the amount of error data. The SVM computes the hyper-plane that maximizes the distances between support vectors for a given parameter setting. In fact, the problem is transformed to find the solution for:

$$\begin{aligned} \arg \min_{w, b, \xi} \quad & \frac{1}{2} \|W^{ij}\|^2 + C \sum_{i=1}^N \xi_i^{ij} \\ \text{s. t.} \quad & (W^{ij})^T \Phi(X_i) + b^{ij} \geq 1 - \xi_i^{ij}, \text{ if } x_i \text{ in class } C_i \\ & (W^{ij})^T \Phi(X_i) + b^{ij} \leq -1 + \xi_i^{ij}, \text{ if } x_i \text{ in class } C_j \\ & \xi_i^{ij} \geq 0. \end{aligned}$$

In this work, the *LIBSVM-2.84* system [2] is used to train the parameters in experiment.

### 3.3 Imbalanced Data Problem

Actually, the *imbalanced data* is a crucial problem in classification. Some categories might heavily outnumber to the other categories. Therefore sample data for training can not correctly reflect entire distribution of data spaces since ‘‘rare category’’ could be easily covered by ‘‘dense category’’. That means a test instance of ‘‘rare category’’ close to the boundary is more likely to be falsely dominated by ‘‘dense category’’. In text classification especially web-page classification this problem exists extensively.

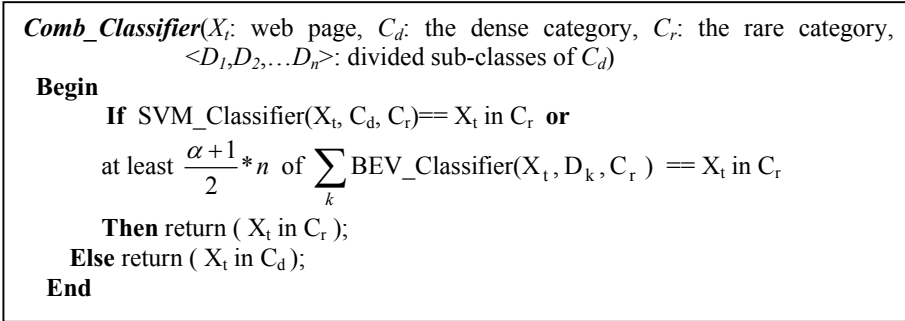
Some solutions were proposed in literature to handle this problem. Traditionally it could be summarized into three groups: under-sampling or over-sampling [4] and adjustment of penalty parameter  $C$  [9]. Li devised another method BEV (Bagging Ensemble Variation) [5]. In his work, the original ‘‘dense category’’ was divided into



several (suppose  $n$ ) sub-categories whose sizes are comparable to the “rare category” so that the positive and negative categories are in balance. Let  $C_i, C_j$  be a dense category and a rare category respectively, with BEV  $C_i$  now is randomly divided into  $n$  classes  $C_{i_1}, C_{i_2}, \dots, C_{i_n}$ , where  $C_i = \bigcup C_{i_k}$  and  $|C_j| \approx |C_{i_k}|$ . They coupled  $C_j$  with each of  $C_{i_k}$  to construct  $n$  classifiers. Then BEV classifies a web page  $X_t$  in  $C_j$  if at least half of those  $n$  classifiers support it, otherwise  $X_t$  in  $C_i$ .

With this method we perform experiment on *20newsgroups* corpus and found it can definitely enhance the performance on “rare category” but at the same time deteriorate on “dense category”. Compared with base classifier, the performance of BEV on categories ‘*alt*’, ‘*misc*’ and ‘*soc*’ which have only one sub-category are better, but the case on other four “dense categories” which take major influence in classification are even worse. The phenomenon could be explained as the divided “dense category” instances give no more perfect cues about the orientation of the hyper-plane, and there is a great range of varying the orientation of the hyper-plane which makes it move to the side of “dense category”.

Based on the above analysis, we adopted a solution called VOTEM to remedy this problem as in figure 2.



**Fig. 2.** VOTEM binary classifier

In training phase of normal method, the “rare category” is in disadvantage since the domination of “dense category”. But in classifying phase, once the classifier announces a document belong to “rare category”, the accuracy of this announcement is relatively high so we can trust it frankly. The same reason suits for BEV method while the classifier predict a document to “dense category” while this prediction could be highly correct. In another word, the classification of data items which are far from hyper-plane is relatively easy so we could guarantee the precision of this part firstly. Then the data items which are close to hyper-plane are ambiguous to distinguish are further analyzed by using both methods. VOTEM combines the advantage of “dense category” in normally classification and benefit of “rare category” in BEV classification. The cases not in these two conditions would be analyzed further, where the outcomes of normal and BEV method are combined to do the voting.

To the second condition in **If** statement,  $\alpha * n + n$  represents  $\alpha$  same votes from normal and  $n$  different votes from BEV get together to decide results. Since SVM\_Classifier already predicts  $X_i$  in  $C_d$ , so half of totally vote is a boundary condition. It suggests that  $\alpha$  should be in range  $[1/2, 1)$  and in our experiment  $\alpha=1/2$  is proved to work well. Principally  $\alpha$  value should be lower than 1 or it will be larger than the vote from BEV alone.

## 4 Experiments Analysis

Two datasets are used in our experiments. *20newsgroups\_18828*, which is announced to be duplicates-removed, is chosen as the first dataset. But we found some documents in the original set are “nearly empty”, with containing only file headers or several terms in their file bodies. We removed this part and leave 17295 documents for experiments.

The second dataset includes some web pages automatically crawled from *www.alibaba.com* [10], a popular B2B electronic commerce portal. We use about twenty thousand non-duplicated documents from 27,688 web pages of products information.

In each leaf category we randomly choose 75% documents as training samples and the left 25% testing.

### 4.1 Measurement for Multi-label Classification

For multi-label classification, the most common and clear indication used to measure the performance is defined as following. This measurement is used in the first dataset.

$$precision = \text{Correct Positive Predictions} / \text{Positive Predictions}$$

$$recall = \text{Correct Positive Predictions} / \text{Positive Data}$$

$$F\sim 1 = \text{precision} * \text{recall} * 2 / (\text{precision} + \text{recall})$$

But the measurement *precision* and *recall* may not indicate the performance of classification results. Since sometimes mis-classification occurs in children category but correct in the parent category and sometimes a document belongs to more than one categories but only one prediction is given. We prefer to making a difference between ‘partial incorrect’ and ‘total incorrect’ in multi-label predictions. So we adopt H-loss value proposed by Cesa-Bianchi [1] as our measurement in the second dataset. The main idea is: *if a parent category has not been predicted correctly, then errors in the children should not be taken into account*. That means the accumulation of loss value is terminated in this situation. A loss function  $L_H$  is defined as:

$$L_H(P, V) = \sum_{i=1}^N c_i \{ p_i \neq v_i \wedge p_j = v_j, j \in \text{ancestor}(i) \}$$

where  $c_1, \dots, c_N > 0$  are cost coefficients in each category node, predicted multi-label set  $P = (p_1, \dots, p_N)$  and the actual multi-label set  $V = (v_1, \dots, v_N)$ . In our experiment costs coefficients  $c_i$  is defined as:  $c_{root} = 1, c_i = c_{pa(i)} / |sibl(i)|$ , where  $pa(i)$  is the parent of node  $i$  and  $|sibl(i)|$  the number of sibling node of  $i$  (including  $i$  itself).

### 4.2 Experiment Results on 20newsgroups

56,841 terms are extracted as the features for 20newsgroups dataset after pre-processing. Firstly we separate documents on first level categories and measure the precision and recall of this level classification. The results are critical since it influence performance of sub-categories processing directly. Figure 4~6 shows the recall, precision and F~1 value sof first level classification with BEV, SVM and VOTEM algorithms. From figure 4&5 we could conclude BEV boost the recall on “rare category” and sacrifice the precision while SVM leads to lower performance on “rare category”. Figure 6 illustrate the comparison of performances of the three methods in this level.

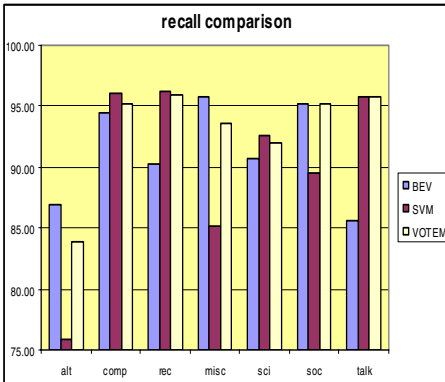


Fig. 4. The recall value of three methods

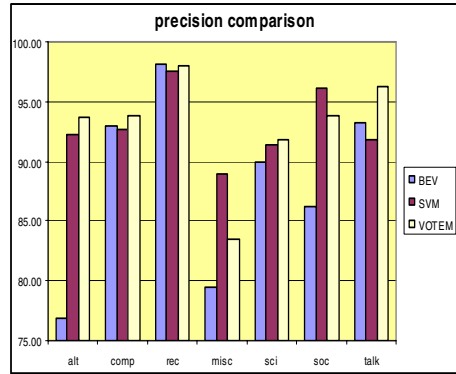


Fig. 5. The precision value of three methods

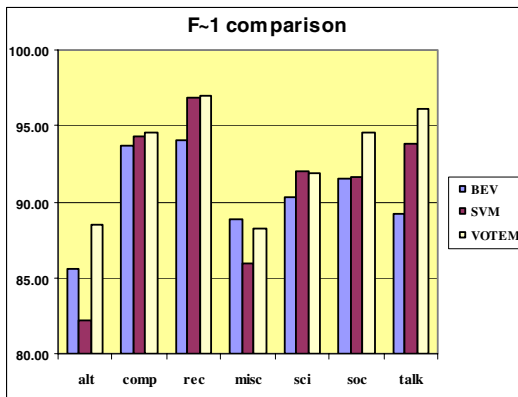


Fig. 6. The F~1 value of three methods in first level

This dataset is classified within a three-level hierarchical tree. We illustrate the classification results and compare the performance with other two individual algorithms in Table 1. VOTEM boosts the performance on “rare category” (such as “alt”,

“misc” and “soc”) while preserving the high accuracy on “dense category” (such as “comp”, “rec”, “sci” and “talk”). For the whole classification, the average-F1 value 88.11% of our method is better than the performance using the other strategies on the same dataset.

**Table 1.** F~1 comparison of three methods in leaf node upon *20newsgroups*

category F~1 measure	alt			comp				rec			
	atheism	graphics	os	sys		windows	auto	motor	sport		
			win.misc	ibm.hw	mac.hw	x			baseball	hockey	
BEV(%)	85.63	68.26	66.30	81.14	84.56	76.84	88.72	90.45	91.73	92.47	
SVM(%)	82.21	70.65	68.03	82.05	86.39	81.28	90.41	91.53	94.63	92.66	
VOTEM(%)	<b>88.52</b>	<b>72.00</b>	<b>75.24</b>	<b>83.92</b>	<b>89.03</b>	<b>83.95</b>	<b>93.75</b>	<b>94.77</b>	<b>97.62</b>	<b>95.63</b>	

misc	sci				soc	talk				micro- average
	forsale	crypt	electro	med		space	politics			
						christian	guns	mideast	misc	
<b>88.81</b>	89.41	74.16	80.42	91.71	91.52	80.37	94.20	75.11	78.30	83.51
85.96	91.00	77.45	84.07	92.37	91.64	83.98	95.27	78.03	80.11	84.98
88.29	<b>93.04</b>	<b>80.33</b>	<b>89.62</b>	<b>93.65</b>	<b>94.52</b>	<b>87.56</b>	<b>97.16</b>	<b>81.32</b>	<b>82.18</b>	<b>88.11</b>

### 4.3 Experiment Results on *alibaba.com* Web- Page

The *alibaba.com* dataset is about product information. The whole dataset is distributed as 17 categories in first level and 46 categories in the second level. The size of the categories in the first level ranges from 172 to 11732 and second from 21 to 5655 which shows the seriously imbalanced data problem. The page in category “Packaging & Paper” occupies nearly half of total number and this category is much bigger than others.

By using VOTEM algorithm, we obtained nearly zero loss value for most first level categories. The average value is 0.47, compared to 1.82 by original SVM model and 2.41 by BEV model. Table 2 illustrates the categories’ loss values within three independent strategies. The bold font represents best performance among all the methods. It is clear that the new model achieves better performance than the other two methods in most case except one. Note that the loss value in each category is the accumulation of cost when document is incorrectly classified to this category.

**Table 2.** Three methods comparison of loss value in first level

$L_H$	Chemical	Business	Home	Apparel	Beverage	Industrial	Textiles	Sports
BEV	0.294	0.118	0.294	<b>0</b>	<b>0</b>	0.352	<b>0</b>	0.118
SVM	0.235	0.352	0.118	0.059	<b>0</b>	0.352	<b>0</b>	<b>0</b>
VOTEM	<b>0</b>	<b>0</b>	<b>0.059</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>

	Health	Luggage	Office	Elec.	Security	Gifts	Package	Auto	Construct
	<b>0</b>	0.235	0.235	<b>0</b>	0.118	0.059	0.412	<b>0</b>	0.176
	<b>0</b>	0.176	0.176	0.059	<b>0</b>	0.059	0.235	<b>0</b>	<b>0</b>
	<b>0</b>	<b>0.118</b>	<b>0</b>	0.059	<b>0</b>	<b>0</b>	<b>0.176</b>	<b>0</b>	0.059

After finishing the first level categories classification, the documents contained in categories would be separated to next level nodes similarly. We take the sub-hierarchy rooted at biggest category “Packaging & Paper” as an example. Note that the incorrect items in previous level are preserved in next level classification. Table 3 shows the comparison of loss value of this level. ‘total#’ in the second line represents the amount of testing data of each category.

**Table 3.** Three methods comparison of loss value of category “Packaging & Paper”

$L_H$	Label	Stocks	Machine	Paper	Industrial	Designing	Printing	Utensils	Materials
total#	3089	278	436	363	3314	560	1555	392	5655
BEV	0.052	0.007	0.013	<b>0.007</b>	0.085	0.020	0.065	0.013	0.059
SVM	<b>0.039</b>	0.020	0.033	0.026	0.039	0.039	0.026	0.020	<b>0.033</b>
VOTEM	0.046	<b>0</b>	<b>0</b>	0.020	<b>0.026</b>	<b>0.013</b>	<b>0.020</b>	<b>0.007</b>	0.046

The results in table 3 show the performance in term of loss function measure of the three algorithms-- BEV, SVM and VOTEM. It is clear that BEV benefits rare categories such as ‘paper’, and original SVM algorithm benefit dense categories such as ‘Materials’. At the same time, our proposed approach remains the best among the three on average in term of the loss value measurement.

## 5 Conclusion

The existence of imbalanced data often decreases the performance of original SVM classifier. For this sake, we use VOTEM algorithm to provide an improved binary classifier to solve the problem brought by BEV. For a multiple category classification, we create binary classifier for any two categories. Then multiple category classification is performed using voting. The experiment results from different measurements and different level in hierarchical structure show that our algorithm outperforms other algorithms. The method can improve the performance of imbalanced data and could be treated as one strategy to solve skewed data problem.

**Acknowledgement.** This Work was supported in part by the University Research Committee under Grant No. RG066/07-08S/GZG/FST and by the Science and Technology Development Found of Macao Government under Grant No. 044/2006/A.

## References

1. Cesa-Bianchi, N., Gentile, C., Zaniboni, L.: Hierarchical Classification: Combining Bayes with SVM. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 177–184 (2006)
2. Chang, C.-C., Lin, C.-J.: LIBSVM: A Library for Support Vector Machines (2007), <http://www.csie.ntu.edu.tw/~cjlin>
3. Dumais, S., Chen, H.: Hierarchical Classification of Web Content. In: Proceeding of the 23rd ACM International Conference on Reach and Development in Information Retrieval, Athens, Greece, pp. 256–263 (2000)

4. Japkowicz, N.: The Class Imbalance Problem: Significance and Strategies. In: Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning, Las Vegas, Nevada, pp. 111–117 (2000)
5. Li, C.: Classifying Imbalanced Data Using A Bagging Ensemble Variation (BEV). In: the ACM Southeast Conference, pp. 203–208 (2007)
6. Rousu, J., Saunders, C., Szedmak, S., Shawe, T.-J.: Learning Hierarchical Multi-category Text Classification Models. In: Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, pp. 745–752 (2005)
7. Sun, A., Lim, E.-P.: Hierarchical Text Classification and Evaluation. In: Proceedings of the International Conference of Data Mining, pp. 521–528 (2001)
8. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
9. Veropoulos, K., Campbell, C., Cristianini, N.: Controlling the Sensitivity of Support Vector Machines. In: Proceedings of the International Joint Conference on AI, pp. 55–60 (1999)
10. Alibaba.com: On-line guide for the internet (2008), <http://www.alibaba.com>

# The Prevalence and Use of Web 2.0 in Libraries

Alton Yeow Kuan Chua, Dion Hoe-Lian Goh, and Chei Sian Lee

Division of Information Studies, Wee Kim Wee School of Communication and Information  
Nanyang Technological University, Singapore 637718  
{altonchua, ashlgoh, leecs}@ntu.edu.sg

**Abstract.** Libraries all over the world are undergoing fundamental paradigm shifts in the way they see their users and in how they offer their services. The thrust is on exploiting the Internet, and in particular Web 2.0 applications, to engage users not only in developing new library services but also building a community. This paper investigates the prevalence and use of Web 2.0 applications of 90 websites of libraries from North America, Europe and Asia. The findings reveal that all three categories of Web 2.0 applications, namely, those that support information push/pull, retrieval, and exchange, have been adopted in libraries across the three regions, with libraries in North America leading their European and Asian counterparts. The ways in which individual Web 2.0 applications have been used are also detailed.

**Keywords:** Web 2.0, libraries, websites.

## 1 Introduction

As hubs for information creation and flow, libraries have traditionally played an integral role in promoting literacy and supporting education in societies. However, the advent of the digital age has called into question the effectiveness of libraries [6]. For one, the ubiquitous nature of the Internet offers unparalleled convenience for users vis-à-vis location-bound libraries. Furthermore, as library resources such as books, magazines, journals and monographs are increasingly delivered in electronic formats, the distinctiveness among libraries is eroding. In response to these challenges, libraries all over the world are undergoing fundamental paradigm shifts in the way they see their users and in how they offer their services. The thrust is on exploiting the Internet, and in particular Web 2.0 applications, to engage users not only in developing library services but also building a community [10].

Hitherto, little work has been done to examine the extent to which Web 2.0 has been implemented in libraries outside the United States [7]. Furthermore, how Web 2.0 have been used to support library services have yet to attract any research attention. For these reasons, the purpose of this paper is to study the adoption of Web 2.0 in libraries. Two research questions to be addressed are as follows: (1) To what extent is Web 2.0 prevalent in libraries? (2) In what ways have Web 2.0 been used in libraries?

The scope of this paper is confined to 90 libraries selected from three geographical regions, namely, North America, Europe and Asia. The sample covers an equal number of public and academic libraries.

## 2 Literature Review

Libraries have traditionally been the hubs of information creation and flow in societies. However, with the advent of the digital age and a new breed of users who belong to the 'Net Generation', libraries are breaking from their traditional *modus operandi* by embracing Web technology to augment their services [10]. The University of Virginia Library, for example, adds thousands of eBook titles to their physical collection and is able to reach much more users than it could with only print titles (<http://etext.virginia.edu/ebooks/>). Apart from delivering digital content, libraries are also beginning to focus on enhancing users' experiences through Web 2.0 applications.

Web 2.0 represents an emerging suite of applications that are interactive, context-rich and easy-to-use [11]. When implemented in libraries, Web 2.0 applications have the potential to promote participatory networking where both librarians and users communicate, collaborate and co-create content of their interests [9]. In this paper, the following Web 2.0 applications namely, RSS, blogs, wikis, social tagging systems, instant messaging and social networking services are reviewed. While the technological components of these applications have existed even before the Web 2.0 era, they have been chosen because they are pertinent to libraries and they represent new ways in which librarians and users use the Web to harness the advantages of Web 2.0.

'Rich Site Summary' (RSS), also known as 'really simple syndication', is designed to feed users with regularly changing web-content of news-like sites, news-oriented community sites and even personal weblogs without requiring the users to visit multiple sites to receive updates [13]. Librarians can create RSS feeds to update users on new items in a collection, services provided and content in subscription databases [9]. RSS may also be used as a form of advertisement to push library information to users who would not otherwise utilize resources provided by libraries.

A blog is defined as a hierarchy of text, images and media objects arranged chronologically. Since blogs can be easily created, updated and maintained without any technical expertise, their numbers have exploded in the recent years globally. The use of blogs is appealing not only because of its low-cost implementation but it gives libraries a human voice and facilitates dialogue between readers and writers [5]. Thus, libraries that do not exploit such a simple medium to connect with users and enhance their user experience are at a losing end [2].

A wiki is a collection of web pages which allows users to add and edit content collectively. Given its ease of use, wikis not only eliminate cycles of electronic mail exchanges but also foster idea-sharing among a community of users who are interested in a given topic. Unlike blogs, wikis are thematically-organized and can be used in libraries as subject guides, policy manuals, resource listings and training resources [4, 8].

Tagging is the process of assigning keywords as a means to annotate Web sites so that they can be easily accessed in the future [6]. These tags may further be shared by others in a social tagging system which in turn creates socialization dynamics among a group of like-minded users [12]. Thus, an important use for social tagging in libraries is to facilitate information search as well as build a sense of community around the libraries' collections.

Instant messaging is a synchronous communication technology that allows users to send real-time messages to other users. Instant messaging differs from e-mail exchanges



in that communication between users takes place in real-time. Libraries can use instant messaging to provide chat-reference services so that users could ask questions and receive responses directly from librarians during specified contact timings [6].

Social networking services such as Facebook, MySpace and Frapper leverage on the Web to build online social networks amongst users who share personal interests and activities or who are keen to explore personal interests and activities of others [1]. With aggregated features found in other Web 2.0 applications such as messaging, blogging, video streaming and social tagging, librarians are able to connect with the users, raise awareness about library services and broaden their contact base.

From an information-processing standpoint, these applications can be categorized into those that support (1) information push/pull (2) information retrieval and (3) information exchange, albeit with some measure of overlap. Applications that support information push/pull allow content to be either distributed to or drawn from the users. RSS, blogs and wikis enable information to be pushed to users. They may also elicit responses from users and therefore serve to pull information from them. Applications that support information retrieval enable content to be found based on a set of user-specified criteria. Social tagging allows users to search for related materials on the Web based on tags created by others, and thus facilitates information retrieval. Applications that support information exchange offer a conducive environment to forging social relationships through content sharing. Instant messaging and social networking services help connect users and allow them to participate in information exchange. We therefore propose a model as shown in Figure 1 to depict the functional underpinnings of Web 2.0 applications. The model visualizes the role of each application and helps guide the line of inquiry in this study.

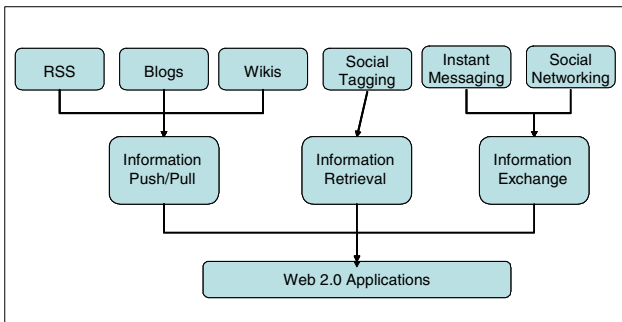


Fig. 1. Functional underpinnings of Web 2.0 applications

### 3 Methodology

#### 3.1 Data Collection

Divided equally between public and academic, 90 libraries' websites from North America, Europe and Asia were sampled, as shown in Table 1.

**Table 1.** Public and Academic Libraries Selected

<b>North America</b>	<b>Europe</b>	<b>Asia</b>
<ul style="list-style-type: none"> <li>• Baltimore County Public Library</li> <li>• Cincinnati And Hamilton County Public Library</li> <li>• Cuyahoga County Public Library</li> <li>• Denver Public Library</li> <li>• Hennepin County Library</li> <li>• Howard County Library</li> <li>• Johnson County Library</li> <li>• Kenton County Public Library</li> <li>• Madison Public Library</li> <li>• Montgomery County Public Libraries</li> <li>• Multnomah County Library</li> <li>• Pikes Peak Library District</li> <li>• Saint Joseph County Public Library</li> <li>• Salt Lake County Library System</li> <li>• Santa Clara County Library</li> </ul>	<ul style="list-style-type: none"> <li>• Bibliotheque nationale de France</li> <li>• German National Library</li> <li>• National Library of Estonia</li> <li>• National Library of Greece</li> <li>• National Library of Ireland</li> <li>• National Library of Latvia</li> <li>• National Library of Lithuania</li> <li>• National Library of Serbia</li> <li>• National Library of Sweden</li> <li>• National Library of the Czech Republic</li> <li>• National Library of the Netherlands</li> <li>• National Szechenyi Library (Hungary)</li> <li>• Russian State Library</li> <li>• The British Library</li> <li>• The National Library of Finland</li> </ul>	<ul style="list-style-type: none"> <li>• Hong Kong Public Libraries</li> <li>• National Central Library (Taiwan)</li> <li>• National Library &amp; Documentation Centre (Sri Lanka)</li> <li>• National Library Board (Singapore)</li> <li>• National Library of Australia</li> <li>• National Library of Bhutan</li> <li>• National Library of Cambodia</li> <li>• National Library of China</li> <li>• National Library of India</li> <li>• National Library of New Zealand</li> <li>• National Library of Pakistan</li> <li>• National Library of Philippines</li> <li>• National Library of Korea</li> <li>• Sarawak State Library</li> <li>• Tokyo Metropolitan Library</li> </ul>
<ul style="list-style-type: none"> <li>• California Institute of Tech.</li> <li>• Columbia University</li> <li>• Cornell University</li> <li>• Duke University</li> <li>• Harvard University</li> <li>• Johns Hopkins University</li> <li>• Massachusetts Institute of Tech.</li> <li>• Princeton University</li> <li>• Stanford University</li> <li>• University of California, Berkeley</li> <li>• University of California, LA</li> <li>• University of Chicago</li> <li>• University of Michigan</li> <li>• University of Pennsylvania</li> <li>• Yale University</li> </ul>	<ul style="list-style-type: none"> <li>• Central European University</li> <li>• Imperial College London</li> <li>• INSEAD, Paris</li> <li>• King's College London</li> <li>• London School of Economics</li> <li>• Trinity College Dublin</li> <li>• University College London</li> <li>• University of Bristol</li> <li>• University of Cambridge</li> <li>• University of Copenhagen</li> <li>• University of Edinburgh</li> <li>• University of Geneva</li> <li>• University of Manchester</li> <li>• University of Oslo</li> <li>• University of Oxford</li> </ul>	<ul style="list-style-type: none"> <li>• Australian National University</li> <li>• Chinese University of Hong Kong</li> <li>• HK University of Science &amp; Tech</li> <li>• Indian Institute of Tech, Madras</li> <li>• Kyoto University</li> <li>• Nanyang Technological University</li> <li>• National University of Singapore</li> <li>• Tsinghua University</li> <li>• University of Auckland</li> <li>• University of Hong Kong</li> <li>• University of Malaya</li> <li>• University of Melbourne</li> <li>• University of New South Wales</li> <li>• University of Sydney</li> <li>• University of Tokyo</li> </ul>

We focused on these three regions because of the availability of a large number of libraries' websites for selection. Websites that are not available in English were excluded. Fifteen public libraries from each region were randomly selected from Hennen's 2006 American Public Library Ratings (HAPLR) report ([http://www.haplr-index.com/ALProofHAPLR\\_2006.pdf](http://www.haplr-index.com/ALProofHAPLR_2006.pdf)), the European Library ([www.theeuropeanlibrary.org](http://www.theeuropeanlibrary.org)) and the Libraries of Asia Pacific Directory (<http://www.nla.gov.au/lap/>). Academic libraries were drawn from QS World University Rankings 2007 ([www.topuniversities.com](http://www.topuniversities.com)). We selected the 15 well-ranked universities from each region and further cross-referenced our list with 1) The Times Higher World University Ranking 2007 ([www.thes.co.uk](http://www.thes.co.uk)), 2) America's Best Colleges 2008 ([www.colleges.usnews.rankingsandreviews.com](http://www.colleges.usnews.rankingsandreviews.com)), 3) Top 100 Europe Universities 2007 ([www.webometrics.info](http://www.webometrics.info)), and 4) Top 100 Asia Pacific Universities ([www.arwu.org](http://www.arwu.org)).

### 3.2 Method of Analysis

All 90 websites selected were analyzed using a two-step content analysis by three graduate research assistants who were familiar with Web 2.0 applications. First, the graduate assistants independently trawled each website for the presence of the various Web 2.0 applications. Informed by the research model presented earlier, the coding scheme covered three dimensions, which are further divided into six variables: RSS (Rich Site Summary), blogs (B), wikis (W), social tagging (ST), instant messaging (IM) and social networking services (SN). All six variables were binary-coded either as 'yes' or 'no' to denote the presence or absence of the respective Web 2.0 applications. Given the participation of multiple coders, inter-coder reliability was established using Cohen's kappa for all the variables. The average pair-wise inter-coder reliability was found to fall between 0.74 and 0.95, indicating a strong non-chance agreement among the coders. The results were statistically analyzed according to the libraries' geographical regions, namely, North America, Europe and Asia, as well as their types, namely, public and academic.

Where a website had been coded 'yes' for any of the variables, a second step of analysis was conducted to understand how the particular Web 2.0 application found was used. Findings on the use of each Web 2.0 application were then aggregated for all adopting libraries.

## 4 Analysis and Findings

### 4.1 Prevalence of Web 2.0 Applications

Shown in Table 2, the findings reveal that all three categories of Web 2.0 applications, namely, those that support information push/pull, retrieval and exchange, have been adopted in libraries across the three regions, albeit in varying degree of prevalence. In terms of specific applications, blogs are most popular (58%), followed by RSS (47%) and instant messaging (43%). Applications less-widely used are wikis (19%), social networking services (16%) and social tagging (9%).

**Table 2.** Number of websites featuring Web 2.0 Applications by regions

Region	Information Push/Pull			Information Retrieval	Information Exchange	
	RSS	B	W	ST	IM	SN
North America	19	23	9	6	28	11
Europe	13	15	3	1	7	2
Asia	10	14	5	1	4	1
Total (N = 90)	42 (47%)	52 (58%)	17 (19%)	8 (9%)	39 (43%)	14 (16%)

Region-wise, libraries in North America were found to embrace all three categories of Web 2.0 applications most unanimously. In particular, applications that support information exchange have been widely adopted in North America vis-à-vis the other

regions. Ranked at a distant second are European libraries which marginally lead their Asian counterparts in the adoption of all applications except for social tagging and wikis. Chi-square analyses show the differences in prevalence of Web 2.0 applications among the three regions to be statistically significant for blogs [ $\chi^2(2, N = 90) = 6.65, p < .05$ ], social tagging [ $\chi^2(2, N = 90) = 6.86, p < .05$ ], instant messaging [ $\chi^2(2, N = 90) = 46.43, p < .001$ ] and social networking services [ $\chi^2(2, N = 90) = 15.39, p < .001$ ]. The rest of the applications were not found to be statistically significant.

**Table 3.** Number of websites featuring Web 2.0 Applications by library types

Region	Information Push/Pull			Information Retrieval	Information Exchange	
	RSS	B	W	ST	IM	SNS
Public	17	23	9	1	20	8
Academic	25	29	8	7	19	6
Total (N = 90)	42 (47%)	52 (58%)	17 (19%)	8 (9%)	39 (43%)	14 (16%)

Table 3 presents the results tabulated by library type. On the whole, the differences in the adoption rates of Web 2.0 applications between public and academic libraries are less stark than those among the regions. While public libraries were found to have a marginally higher adoption rates for wikis, instant messaging and social networking services, they trail behind academic libraries in adopting RSS, blogs and social tagging. Chi-square analyses show the difference to be statistically significant only for social tagging [ $\chi^2(1, N = 90) = 4.94, p < .05$ ].

## 4.2 Use of Web 2.0 Applications

Libraries that adopt RSS mostly use it to communicate news, updates on the collections or new postings appearing in blogs. Denver Public Library, for example, offers RSS feeds to the latest U.S. and world news. Australian National University Library's RSS feeds are linked to the library's electronic resources that serve to notify users whenever papers or journals of interest emerge. Cornell University Library uses RSS as a conduit to its "LibCast" which features audio-visual recordings and updates on exhibitions and events.

Blogs are commonly used to generate interest in subject-specific topics as well as to engage users. For example, Copenhagen University Library uses blogs to introduce new books organized by disciplines such as health science, humanities and theology. Hennepin County Library encourages participation amongst users by offering dedicated blogs for demographically-similar users. For instance, it has a Teen Speak Section that attracts teenage bloggers. The blogs hosted by National Library Board of Singapore are intended for users to share photographs on local themes such as Singapore landmarks and festivals.

Libraries that feature wikis in their websites develop their own subject guides either by using open source wiki software, as seen in Saint Joseph County Public Library, or simply providing links to external subject-based wikis as seen in Salt Lake

County Library and Hennepin County Library. In University of Hong Kong Libraries, wikis are used to archive past questions on a range of topics such as library services and book renewal procedures posted by users.

With the exception of University of Pennsylvania Library which develops its own social tagging tool known as Penn Tags, all libraries that promote social tagging provide a link to websites such as Connotea, del.icio.us and Digg. Following the link, users may register with these social tagging sites which enable them to save, organize and share any references addressable by a URL. Libraries at Duke University and Stanford University offer instructions on how to get started with social tagging.

Libraries adopt instant messaging to handle users' enquiries synchronously during predefined timeslots. Common tools including Yahoo Instant Messenger, MSN Messenger, ICQ and Skype are used in Princeton University Library and National Szechenyi Library (Hungary) while custom-built applications such as ChatRef and AskNow are used in Harvard University Libraries and National Library of Australia respectively. An interesting use of instant messaging was found among many public libraries in North America such as Multnomah County Library, Cincinnati and Hamilton County Public Library, and Santa Clara County Library where after-class online tutoring services are provided free to any user holding a library card. Usually available from the afternoon to night, these services are meant to help users with their homework.

Libraries rely on social networking services such as Facebook and MySpace to forge personalized connections with their users. Hennepin County Library, for example, has a user account in Facebook which features searchable catalog, and displays messages posted by both librarians and users while the British Library uses its Facebook account to share pictures and video clips. Denver Public Library's account on MySpace, called Denver eVolver, has been designed for teenage users by delving into topics of interest to them such as music/movie reviews and homework help.

## 5 Discussion and Conclusion

The data collected from this study has yielded two main findings. First, the order of popularity among different Web 2.0 applications adopted in all libraries is as follows: blogs, RSS, instant messaging, wikis, social networking services and finally, social tagging applications. Libraries in North America lead significantly in the adoption of blogs, social tagging, instant messaging and social networking services vis-à-vis their European and Asian counterparts. One possible reason could be disparate internet penetration rates among the regions. As of 2007, internet penetration rate in North American stands at 69.7% while those in Europe and Asia lag at 38.9% and 10.7% respectively (<http://www.internetworldstats.com/stats.htm>). The results may well reflect the responsiveness among North American libraries to meet the needs and expectations of the Web savvy populace. Conversely, Asian libraries, which remain at the poorer end of the global digital divide at this stage, may find embracing the Web a luxury rather than a necessity. Another possible reason could be that it is usually expensive for libraries in non-English speaking regions to provide the Web services in English. Thus, the English version of Web 2.0 applications in Europe and Asia may well be limited compared to those in their respective native languages.

Analyzing between library types, public and academic libraries share comparable adoption rates of all Web 2.0 applications except for social tagging. This difference could be attributed to the disparate profiles of their users. The presence of sizeable cohorts of students sharing similar interests and disciplinary areas in academic libraries may have necessitated a more prevalent adoption of social tagging applications. Users of public libraries, on the other hand, are conceivably more heterogeneous even though fragmentary pockets of like-minded users may exist. Going forward, as usage levels in public libraries increase, the emergence of a greater number of sustained user communities could then heighten the need for social tagging applications.

Second, while the underlying function of all Web 2.0 applications points towards piquing users' interests and enhancing their experience with library services, the manner in which individual applications can be used is limited by the imaginations of librarians and decision makers. A case in point is the interesting extension of the use of instant messaging. In addition to providing users synchronous access to librarians on library-related matters, a number of public libraries in the North America offer online tutoring services via instant messaging to help students with their homework.

Moreover, rather than treating each Web 2.0 application as being distinct, libraries have correctly recognized how the applications can complement each other to increase the level of user engagement. In particular, it is common to find blogs used in conjunction with RSS feeds so that users can be notified whenever new posts related to a particular topic of interest emerge. An example of concerted provision of multiple Web 2.0 applications including blogs, instant messaging and links to social networking services can be found in the teen's section of Kenton County Public Library's website. Such aggregated deployment of Web 2.0 applications, which conceivably serves the needs of users better, is likely to become more common in libraries' websites.

Our research is significant because it represents one of the earliest works to shed light on the current level of adoption and use of Web 2.0 applications in libraries globally. Not only is the prevalence of Web 2.0 applications compared across regions and between library types, the ways in which individual application has been used are examined. Decision makers and Web designers of libraries may benchmark their own efforts in deploying Web 2.0 applications against this study.

Three main limitations in our study must be acknowledged. One, the selection of libraries from North America, Europe and Asia was limited to those whose websites were available in English. Future work can look into examining libraries' websites in other languages, and from other regions. Two, only a sample of six Web 2.0 applications have been chosen for analysis. Our study can be expanded to include applications such as mobile services and a slew of other emerging Web 2.0 applications such as Ning, Twitter, Pageflakes and Diigo. Three, the scope of data collection was limited to what was publicly available on the Internet. The influence of Web 2.0 on users' behavior, for example, could not be studied. Thus, future work can delve into how the implementation of Web 2.0 applications has changed human dimensions such as perceptions, needs and behaviors of users and librarians. Additionally, what Web 2.0 applications contribute to OPAC usability and how they can improve the exploitation of library holdings are also research-worthy efforts to be undertaken.

## References

- [1] Barsky, E., Purdon, M.: Introducing Web 2.0: Social networking and social bookmarking for health librarians. *Journal of the Canadian Health Libraries Association* 27, 65–67 (2006)
- [2] Clyde, L.: Library Weblogs. *Library Management* 25(4/5), 183–189 (2004)
- [3] Curran, K., Murray, M., Christian, M.: Taking information to the public through Library 2.0. *Library Hi Tech* 25(2), 288–297 (2007)
- [4] Frumkin, J.: The wiki and the digital library. *OCLC Systems & Services* 21(1), 18–22 (2005)
- [5] Goodfellow, T., Graham, S.: The blog as a high-impact institutional communication tool. *The Electronic Library* 24(4), 395–400 (2007)
- [6] Gibbons, S.: *The Academic Library and the Net Gen Student: Making the Connections*. American Library Association, Chicago (2007)
- [7] Liu, S.: Engaging Users: The Future of Academic Library Web Sites. *College & Research Libraries* 69(1), 6–27 (2008)
- [8] Macgregor, G., McCulloch, E.: Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review* 55(5), 291–300 (2005)
- [9] Maness, J.M.: Library 2.0: The next generation of Web-based library services. *LOGOS: Journal of the World Book Community* 17(3), 139–145 (2006)
- [10] Miller, P.: Web 2.0: Building the new library. *Ariadne* 45 (2005), <http://www.ariadne.ac.uk/issue45/miller/>
- [11] O'Reilly, T.: What is Web 2.0: Design patterns and business models for the next generation of software. *Communications and Strategies* 65, 17–37 (2007)
- [12] Razikin, K., Goh, D.H.-L., Chua, A.Y.K., Lee, C.S.: Can social tags help you find what you want? In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) *ECDL 2008. LNCS*, vol. 5173, pp. 50–61. Springer, Heidelberg (2008)
- [13] Stephens, M.: *Web 2.0 & Libraries: Best Practices for Social Software*. American Library Association, Chicago (2006)

# Usability of Digital Repository Software: A Study of DSpace Installation and Configuration

Nils Körber and Hussein Suleman

Department of Computer Science, University of Cape Town  
Private Bag, Rondebosch, 7701, South Africa  
nils.koerber@gmail.com, hussein@cs.uct.ac.za

**Abstract.** Usability of the installation and configuration of digital repository software is a key factor for the implementation of digital repositories. Many universities, laboratories and companies want to place their collections online but the installation and configuration processes of digital repositories are sometimes time-consuming and unnecessarily complicated. This paper describes efforts to highlight usability issues while setting up and configuring DSpace. The focus of three studies that were performed was not end-user usability but usability of the administrative functionality. User evaluations performed on a recent version of DSpace were followed by participatory design of a tool to increase usability by abstracting away the lower-level details. Users agreed that such a tool would be suitably usable. Thus it was found that significant usability problems exist, but these problems may in fact be easily addressed.

## 1 Introduction

The emerging Open Access movement has in recent years played a defining role in the evolution of digital repository software, which is used primarily to archive and disseminate research-related documents. In the 90s such tools were usually custom-developed to meet the objectives of specific organisations or projects. However, for Open Access to be widely adopted, it was necessary for repository tools to easily be redeployed in different scenarios. This led to or supported the creation and ongoing development of digital repository tools such as DSpace [2], EPrints [1] and Fedora [3]. The Budapest Open Access Initiative has as one of its core activities the monitoring and evaluation of such tools, as Open Access hinges on this to a large degree [4].

While the link between Open Access and reusable repository tools is acknowledged, less effort has gone into the degree of reusability. Tools that are easier to reuse should lead to greater use, hence more Open Access activities. This has not been a focus. Digital repository tools have not yet joined the growing Open Source software repositories - neither EPrints, DSpace nor Fedora are found in the standard Ubuntu software repository (with 25031 tools) or FreeBSD ports collection (with 18700 tools) as of early 2008.



End-user usability has improved to keep pace with developments in Web applications, for example by the incorporation of AJAX techniques in newer systems. Administrative interfaces (such as installation and configuration), on the other hand, have not improved substantially. Historically, a systems administrator was required to install and configure digital library tools. Recent releases of both EPrints and DSpace, however, can be installed on MS-Windows. In addition, Linux-based OSes such as Ubuntu are being used by more end-users on the desktop. Sophisticated package management tools (e.g., Synaptic, emerge, MSI, FreeBSD ports) are available on all OSes to enable painless end-user installation of software - many of these tools handle the installation of dependencies automatically. Thus the underlying facilities are in place for end-users to install and configure their own repositories - all that remains is for the repository tools to present suitable interfaces to users for the tool-specific installation and configuration tasks.

The ultimate goal of this project is to abstract away the complexity of installation and configuration to lower the bar for adoption of digital repository tools. This paper discusses the initial stages of this process: confirming and highlighting problem areas in administrative usability that are previously only discussed informally; and formulating and testing a design for improved administrative usability based on the problems identified, using a participatory design approach with end-users. The paper has focused on DSpace as a candidate system, but the approach and results can be generalised to varying degrees to other systems.

The paper first presents some background on usability of digital repositories, then proceeds to describe the 3 user-centred studies that were performed, finally concluding and presenting avenues for further work.

## 2 Installation and Configuration in Digital Libraries

Like other open source projects, usability is often far behind other development-related issues, even if usability is a continuing topic of interest in the digital library community. Theng [5] stated in 2000 that little work has been done to understand the purpose and usability of digital libraries. Nichols [6] pointed out that there is possibly a general Open Source usability problem. He comments, though, that Open Source Software development has not completely ignored the importance of good usability.

Usability of digital libraries depends on three key components: content, functionality and the user interface [7]. In keeping with this theme, Dillion [8] defined digital libraries usability as how easily and effectively users can find information from a digital library, with an increasing emphasis being placed on the user. The JDCL 2002 “Usability of digital libraries” workshop emphasized usability, but with a focus on just end-user usability rather than usability of administrative mechanisms for the set-up and maintenance of DL systems.

Greenstone, produced by the New Zealand Digital Library Project at the University of Waikato, has paid more attention to usability than most other digital library systems. The distribution includes ready-to-use binaries for the most

common operating systems. Many previous studies [9] [10] [11] evaluated Greenstone's user interface and customization - though most of the issues reported on are traditional end-user usability issues. For example, the *Send feedback* button in Greenstone is an attempt to improve the usability of the Greenstone interface - every time someone makes use of the button, information is collected about which action was performed, which browser was used and the screen settings [12].

DSpace was created as a digital repository tool to capture the intellectual output of multidisciplinary research organizations. As of 2008, over 250 institutions are currently the DSpace software within their organizations in a production or project environment. The most common use is by research libraries as an institutional repository, however there are many organizations using the software to manage digital data for a project, subject repository, Web archive or dataset repository. During user studies Ottaviani [13] discovered that the DSpace administrator interface is difficult to navigate and stressed the importance of having a clear interface. In an analysis of activity of the DSpace-tech mailing list archive it was found that topics related to the installation and configuration issues were common. Messages about *installation* could be found in 1451 listings and *configuration* in 1168 from a total of 11815 messages [14]. This can be interpreted as a need for enhancement in the process of installation and configuration.

Most efforts have focused on the end-user. Usability of the installation and configuration of digital library systems has not often been a point of interest.

### 3 Methodology

The goals of these studies were to assess the impact of usability problems during the DSpace installation and configuration processes and develop and test a set of guidelines to improve on the usability of the administrative interfaces. There were 3 stages: a user study of the DSpace installation; a user study of the DSpace configuration; and a paper prototyping workshop to develop an interface design for a DSpace system tool to simplify the installation process.

#### 3.1 User Study - Installation

This first study examined how average users proceed through the installation process of DSpace. The invited 10 participants were a homogeneous mix of digital library beginners, users and professionals, with various occupational and educational background. 5 participants had never worked with a digital library before but all had worked with computers on a daily basis and had installed software on a Linux operating system before. The 3 digital library users worked almost daily with a digital library system but had never installed or configured one before, while the 2 professionals had done that at least once. The 10 participants were asked to perform an installation of DSpace on a clean Ubuntu Linux operating system. They were allowed to use the DSpace online help and additional information to guide them in the installation. Users were observed and

given a questionnaire which had to be answered after the installation. All participants successfully installed DSpace but most of them had significant problems. Comments from participants include:

- “You have to do everything by hand, why is there no wizard [that] guides [you] through the installation”
- “I installed software on Linux before, but this is challenging”

The fastest user, a librarian who had installed DSpace before, needed nearly 30 minutes, while the average installation time was more than 45 minutes. Observation showed that most of the participants struggled while working with the Linux shell and editing configuration files. 8 out of 10 participants strongly disagreed that “The installation process is simple” and “Overall usability is satisfying”. There was agreement from all users with “I would need documentation for a second installation” and “A system tool to ease the configuration and to reduce the single steps is necessary”. Figure 1 shows the results in detail.

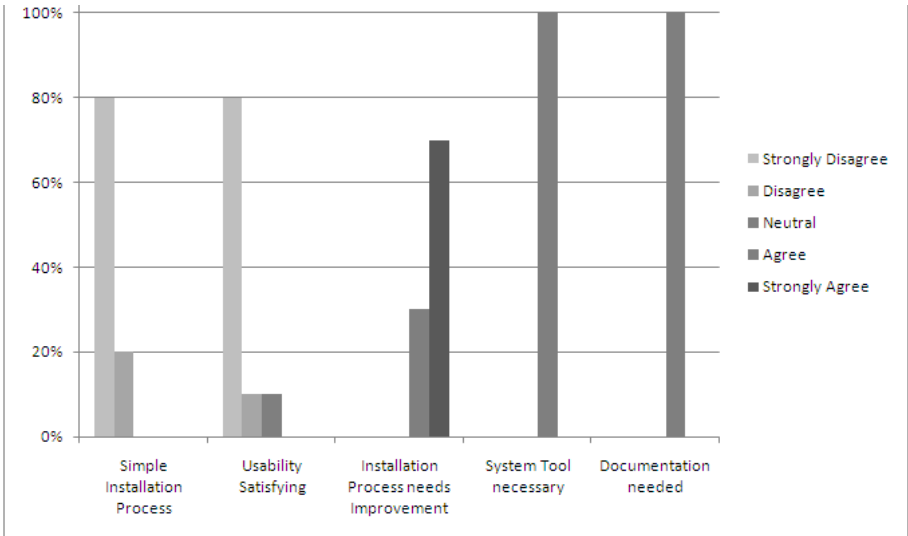


Fig. 1. Questionnaire User Study Installation

One of the users commented: “DSpace is such a great tool but the installation is just frustrating - I wouldn’t be able to set it up on my personal computer without help”. According to users, the most complicated step during the installation process was the task where users had to copy a database driver and edit the DSpace CFG file. The analysis of the results suggest that a wizard-style system tool is desirable to make the installation process easier.

### 3.2 User Study - Configuration

The second study investigated usability problems related to DSpace configuration. To complete the setup of DSpace various configuration steps are needed.

These are, for example, database and file storage configuration and general settings like the log file directory or changing of the logo. There were 10 participants in this study - this time not beginners but all with some background in digital libraries. 5 of the 10 participants had worked with digital libraries on a daily basis and the others had worked at least once on a digital library system. Observation and a user satisfaction questionnaire were used again as the usability testing methods. Most of the participants succeeded with the basic configuration, for instance customization of the base URL and editing the name of the site. More problems were caused when customizing the overall layout, where 7 of the 10 participants could not edit the HTML code from different JSP files. There was broad agreement with the question “I would need documentation for a second configuration” and disagreement with “The configuration process is simple”. Furthermore, 9 people strongly agreed that “A system tool to ease the configuration and to reduce the single steps is necessary”. Figure 2 shows an analysis, rating each participant’s reaction to the configuration process.

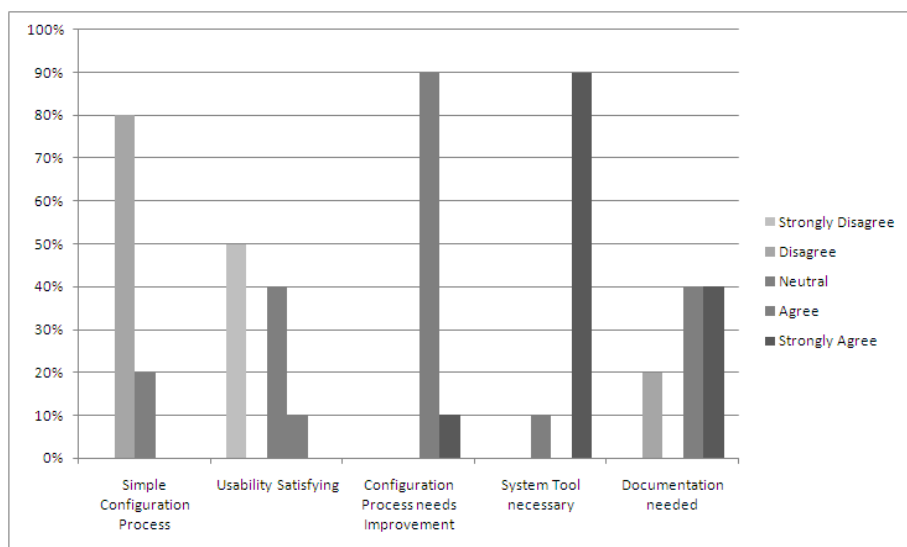


Fig. 2. Questionnaire User Study Configuration

A user who had previously configured DSpace several times commented: “I always need documentation for the configuration, the system is not self explaining at all”. Another participant stated: “It’s good that we’ve got our system administrators configuring DSpace for us, we users would need training”.

Together, these studies suggest that the installation and configuration process of DSpace is too complex for users who are not administrators. The lower level details of the process confuses users and makes it more difficult for them to install this open source digital repository tool.

### 3.3 User Development and User Evaluation of Prototype

The first 2 studies confirmed that usability of installation and configuration is a problem and that these processes can be improved. During these studies, users had suggested that abstraction of DSpace's installation and configuration processes could possibly be achieved using a high level system tool. Thus, the impact of such a tool was investigated as a natural next step, to determine if such an intervention would indeed improve the administrative usability of DSpace.

An initial design was developed based on the feedback from the first 2 studies. Users were asked for feedback on the interface design and workflow through a paper prototyping workshop. There were 6 participants in this process - three of them were postgraduate computer science students with HCI experience while the other three had a library background. This design was developed in a 3-stage process, with 2 iterations of the process.

Paper prototyping is a widely used method for designing, testing, and refining user interfaces. Snyder [16] defined paper prototyping as follows: Paper prototyping is variation of usability testing where representative users perform realistic tasks by interacting with a paper version of the interface that is manipulated by a person playing computer, who doesn't explain how the interface is intended to work.

The first stage of each iteration involved discussion of the overall design of the wizard, with paper sketches and Post-it notes to represent the system tool screens and interface elements. Participants could rearrange single elements freely (see Figure 3). The next stage of the study was to determine if the elements in each window were logically grouped or not. Participants were asked to *click* through each window and comment on positive and negative aspects. All users were satisfied with the outcome.

The final stage in each iteration involved going back to the participants to determine the level of usability of the different individual elements and the whole interface. Participants had to evaluate the site using a usability checklist for clarity of communication, accessibility, consistency, navigation, design and maintenance and visual presentation [15].

The first set of interface designs created during the sessions is the installation wizard. Instead of editing files and copying drivers a 3 screen wizard will guide users through the installation. The first screen is a welcome screen guiding users through the installation (see Figure 4).

The manage or configuration screen (see Figure 5) provides an overview of existing repositories and gives users the option to create a new instance. Each repository is displayed as a DSpace icon and the name. Because of simplicity the design session participants abstained from using buttons to edit existing repositories. After double-clicking on a repository the interface shown in Figure 6 will be opened.

The configuration manager is an interface to change basic settings on single repositories. During the design iterations the decision was made to not use a wizard style guide but a tabbed interface. One tab gives users the possibility to edit repositories and the other to copy or delete an instance.

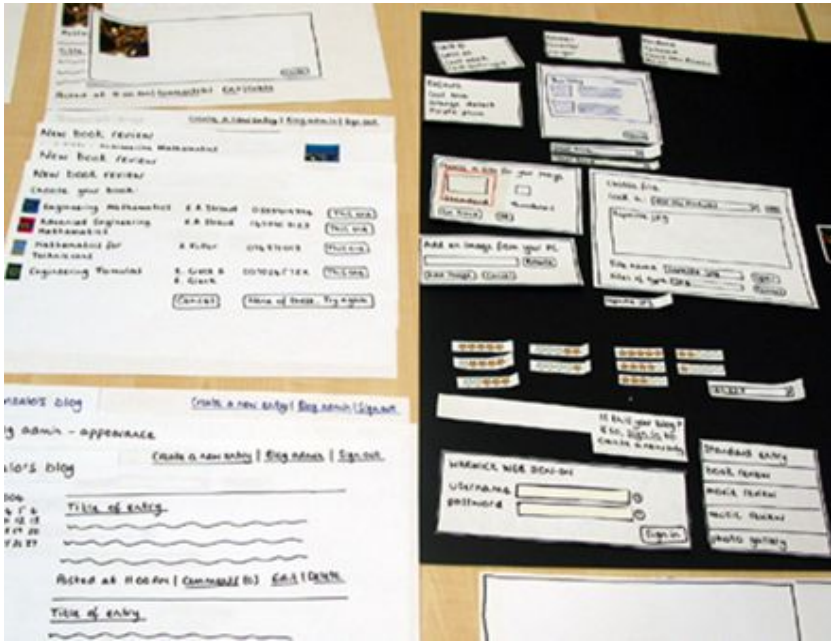


Fig. 3. Paper prototyping of system tool

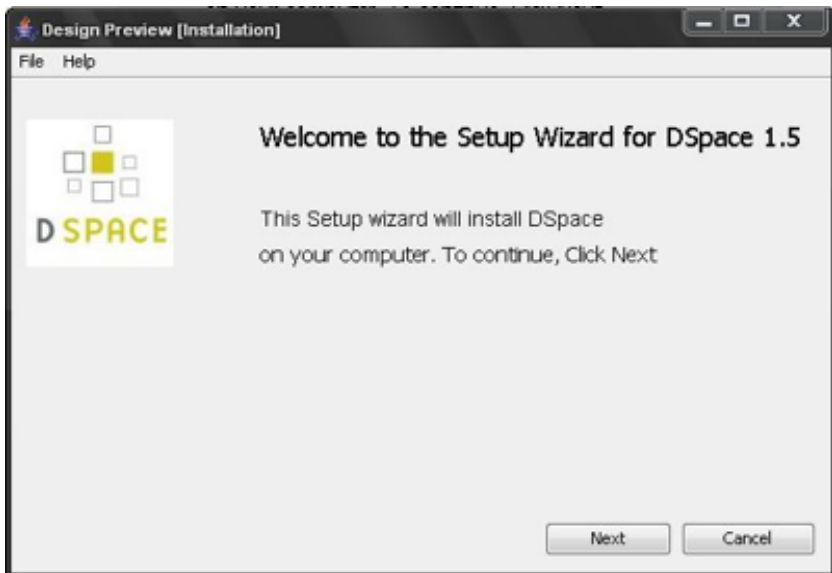


Fig. 4. Interface Design Installation



Fig. 5. Interface Design - Repository Manager

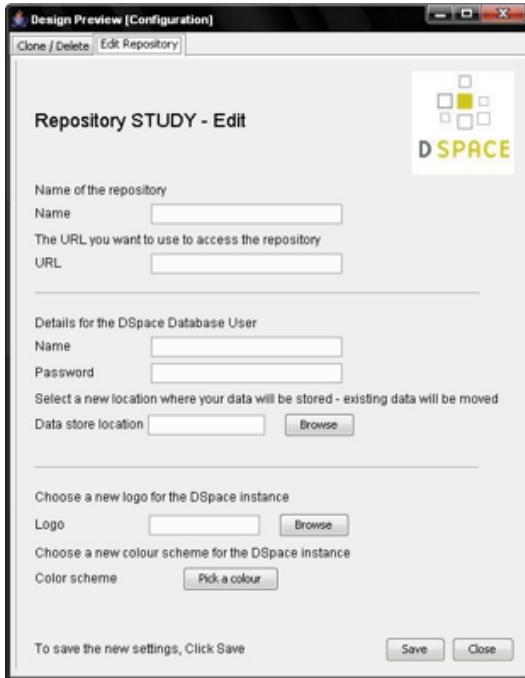


Fig. 6. Interface Design - Repository Configuration

A user who participated in both user studies and the interface design process commented after completion of the design process: “The usability will be improved, the interface design is promising”.

The final set of core features of the system tool, as designed by users, is as follows:

- Users should be guided through installation and configuration, with a wizard where appropriate.
- User should be able to manipulate existing repositories.
- Before setup, dependencies should be checked for automatically.
- Current repositories should be managed through a configuration manager, which reads and writes configuration information from and to various sources e.g., the `dspace.cfg` file
- Cloning and deleting repositories should also be supported.

## 4 Conclusions and Future Work

This paper set out to formally identify problems with the administrative usability of digital repository tools, with a focus on DSpace. While this study was conducted with only DSpace, the results are arguably applicable to other similar systems as well.

The user studies confirmed the existence of usability problems in installation and configuration. Specific issues were highlighted and serve as a reference for any solutions to this problem. One solution that is proposed is the design of a high-level system tool to guide users through the process of installation and configuration, abstracting away the lower-level details. Participatory design with end-users rather than systems administrators should lead to greater usability of all aspects of the repository, including administrative interfaces. This study confirmed that, at a design level, usability of installation and configuration can indeed be improved.

This work thus suggests that it should be possible to develop digital repository tools as *any-user* applications rather than specialist administrator-only tools. Some of this is already possible with Greenstone (especially installation) but most other tools have poor administrative usability.

Future work will involve development of prototype tools for abstraction of installation, configuration and management of repositories, and further usability evaluation of such tools.

In future, it is hoped that digital repository software will enter the mainstream and be packaged and distributed through the emerging software repositories available for most modern operating systems, as this will contribute further to adoption of such tools.

Digital repositories should also integrate with software configuration management tools where they are available, thus making it easier to manage multiple instances of repository software and instances of different repository tools within a single environment, without sacrificing usability along the way.



## References

1. DSpace (2008), <http://www.dspace.org>
2. EPrints (2008), <http://www.eprints.org>
3. Fedora (2008), <http://www.fedora.org>
4. Crow, R.: A Guide to Institutional Repository Software, 3rd edn, June 2008. Open Society Institute (retrieved June 2004), <http://www.soros.org/openaccess/software/>
5. Theng, Y.L., Mohd-Nasir, N., Thimbleby, H.: Purpose and usability of digital libraries. In: Proceedings of the fifth ACM Conference on Digital Libraries, pp. 238–239. ACM Press, New York (2000)
6. Nichols, D.M., Twidale, M.B.: The usability of open source software, 8(1) (2003) (retrieved June 2008), <http://www.firstmonday.org/issues/issue8.1/nichols>
7. Van House, N.A., Butler, M.H., Ogle, V., Schiff, L.: User-centered iterative design for digital libraries, D-Lib Magazine (1996) (retrieved June 2008), [www.dlib.org/dlib/february96/02vanhouse.html](http://www.dlib.org/dlib/february96/02vanhouse.html)
8. Dillon, A.: Designing Usable Electronic Text: Ergonomic Aspects of Human Information Usage. Taylor and Francis, Bristol (1994)
9. Kemp, E., Phillips, C., Kinsluk, Haynes, J.: Usability and open-source software development, Palmerston North, New Zealand. ACM SIGCHI New Zealand, pp. 49–54 (2001) ISBN: 0-473-07559-8
10. Chilana, P., Groetsch, K.: Greenstone Librarian Interface: Usability Analysis / Proposed Improvements, LIS590 II, Interfaces to Information Systems (retrieved June 2008), [http://www.oldtasty.com/classes/LIS590II/Assignment\\_2c/](http://www.oldtasty.com/classes/LIS590II/Assignment_2c/)
11. Witten Ian, H.: Customizing digital library interfaces with Greenstone, tcld Bulletin (2003) (retrieved June 2008), <http://www.ieee-tcdl.org/Bulletin/v1n1/witten/witten.html>
12. Greenstone Send Feedback, User feedback about the Greenstone user interface (retrieved June 2008), <http://nzdl.sadl.uleth.ca/gSDL/usability/about.html>
13. Ottaviani, J.: University of Michigan DSpace (AKA Deep Blue) Usability Studies: Summary Findings, University of Michigan (2006) (retrieved June 2008), [http://sambuca.umdl.umich.edu/bitstream/2027.42/40249/1/Deep\\_Blue\(DSpace\)\\_usability\\_summary.pdf](http://sambuca.umdl.umich.edu/bitstream/2027.42/40249/1/Deep_Blue(DSpace)_usability_summary.pdf)
14. DSpace advanced search, mailing lists, dspace-tech (retrieved June 2008), [http://sourceforge.net/search/?group\\_id=19984&type\\_of\\_search=mlists](http://sourceforge.net/search/?group_id=19984&type_of_search=mlists)
15. Sullivan, T.: User Testing Techniques - A Reader-Friendliness Checklist (1996) (retrieved June 2008), <http://www.pantos.org/atw/35317.html>
16. Snyder, C.: Paper Prototyping: The Fast and Easy Way to Design and Refine User Interfaces. Morgan Kaufmann, San Francisco (2003)

# Developing a Traditional Mongolian Script Digital Library

Garmaabazar Khaltarkhuu and Akira Maeda

College of Information Science and Engineering, Ritsumeikan University  
1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan  
garmaabazar@gmail.com, amaeda@is.ritsumei.ac.jp

**Abstract.** This paper discusses our approaches to create a digital library on traditional Mongolian script. We introduce system architecture of a digital library that stores books and materials of historical importance written in traditional Mongolian which contain history of 1,000 years and are important part of Mongolian culture. Specifically, we discuss the issues of traditional Mongolian script encoding. There are several non-Unicode encoding developed, some of them are still commonly used these days. Furthermore, a diverse character encoding, code pages and keyboard drivers are precluding proper enrichment of Unicode based content. We analyzed existing non-Unicode encoding of traditional Mongolian script. Developing text conversion technique from diverse encoding to Unicode is becoming essential demand in order to adopt already made digital content in traditional Mongolian script easily. Without such a conversion technique, the traditional Mongolian script digital library encounters difficulties such as content enrichment by retyping and unable to utilize already developed dictionaries in non-Unicode encoding.

**Keywords:** Traditional Mongolian Script, Digital library, Unicode, Encoding.

## 1 Introduction

The main purpose of this research is to develop a method to keep over 1,000 years old historical records written in traditional Mongolian script including the history of Chinggis Khaan for future use, to digitize all existing records, and to make those valuable data available for public viewing and screening.

There are over 50,000 registered manuscripts and historical records written in traditional Mongolian script stored in the National Library of Mongolia. About 21,100 of them are handwritten documents. There are many more manuscripts stored in libraries of other countries. Despite the importance of keeping 1,000 years old historical materials in good conditions, the Mongolian environments for material storage is not satisfactory to keep historical records for a long period of time. Most efficient and effective way to keep old historical materials while making it publicly available is to digitalize historical records and create a digital library. Thus this paper introduces some techniques to build a digital library on traditional Mongolian script.

We introduced the system architecture of the traditional Mongolian script digital library (TMSDL)<sup>1</sup> [6][7] to preserve historical records over 1,000 years old and to make them available for the public. Developing robust retrieval system with rich digital content and high level of usability is our key aspiration. In the TMSDL we utilized Unicode for digital text collection. While improving TMSDL, we have realized that Unicode usage for the information processing on traditional Mongolian script is not widely spread among researchers and developers. Researchers and developers were introduced several encoding and code pages. Furthermore, a diverse character encoding and code pages are precluding enrichment for the content of TMSDL. This paper discusses complexities of traditional Mongolian information processing, disparities between Unicode and other character encoding, and its implementations.

## 2 Traditional Mongolian Information Processing

Because of the unique characteristics of traditional Mongolian script, procedures to process the information such as inputting, displaying, encoding, typing, typesetting, editing, printing, and recognizing the traditional script have become more complicated.

Difficulties of the traditional Mongolian script include: the existence of at least three different variations of one letter, based on its position in a word or its proximity to a preceding letter which forms a ligature, written vertically from top to bottom in columns advancing from left to right and have more than one phoneme for some shapes, such as letter “t” and “d”. In this section we explore codepage and encoding complexity of traditional Mongolian script, since they are directly associated with the development of TMSDL.

### 2.1 Encoding Standard for Traditional Mongolian Script

One of the biggest problems is that the traditional Mongolian script differs from the modern Mongolian language. Mongolia introduced a new writing system (Cyrillic) in 1946. This has been a radical change and alienated the traditional Mongolian language.

Although the traditional Mongolian character code set has been placed in Unicode at the range of 1800-18AF [18], it is not enough to solve problems of information processing in Mongolian. Problems described below still exist. The traditional Mongolian writing system is known to be quite different from the western systems as well as the CJK (Chinese, Japanese and Korean) writing systems.

- Traditional Mongolian script is written vertically from top to bottom in columns advancing from left to right. This directional pattern is unique.
- Traditional Mongolian characters are written in succession, meaning that depending on where a letter is placed in a word, it may have different forms. There are at least three different forms for each letter and some letters have a dozen different forms. Those are called: isolate, initial, middle, and final form. All of these forms are decided by their position in a word (Fig. 1).

---

<sup>1</sup> <http://www.dl.is.ritsumei.ac.jp/tmsdl/>

Mongolian letter (transliteration)	Isolate	Initial	Medial	Final
ᠠ (A)	ᠠ ᠡ	ᠠ	ᠠ ᠡ ᠢ	ᠠ ᠡ
ᠢ (OE)	ᠢ	ᠢ	ᠢ ᠣ ᠤ	ᠢ ᠣ
ᠤ (L)		ᠤ	ᠤ	ᠤ

Fig. 1. Initial, medial and final forms of Mongolian script letters [5]

The form of a letter may also depend on the preceding letter with which it forms a ligature (Fig. 2). Each letter has a basic form as well as some possible variations of forms, while certain combinations of letters combined form ligatures.

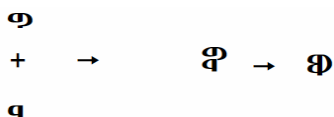


Fig. 2. Mongolian script ligatures [5]

The Unicode standard includes the basic character sets, special punctuation symbols and numerals, but does not explicitly encode the variant forms or the ligatures, although the correct variant form or ligature can, in most cases, be determined from the context.

Besides the issues associated with the unique characteristics of the traditional Mongolian script, there are several encoding and modified code pages exist. Recently Microsoft Windows Vista has included IME which supports traditional Mongolian script. Before that there was no standardized IME available for inputting text in traditional Mongolian script.

Prior to the Unicode standard, Mongolia did not have national character encoding like American ASCII or the Japanese JIS for the encoding of traditional Mongolian script. In China (PRC) GuoBiao established 7 and 8-bit encoding standard GB 8045-87 [15] for traditional Mongolian character set. In Mongolia GB 8045-87 was not used.

In general, the 8-bit encoding of commonly used systems cannot define more than 256 characters of which 128 have been used already. If traditional Mongolian script is in need to be included in one common codepage with modern Mongolian Cyrillic and Latin, it is only possible to overwrite letter shapes of Latin or Cyrillic characters range. There was no other solution if a user wants to use one file without switching the code pages.

Due to unavailability of encoding standard and IME, all available implementations either defined their own codepage or relied on ASCII representations of Mongolian which are then converted into traditional Mongolian. There are several encoding and modified code pages; implementations are discussed below.

### 2.2 Codepage Implementation

In recent years there have been many efforts to solve the problems of information processing on traditional Mongolian script. Those include the "Sudar" package of the National University of Mongolia (1991-92), MBE (Mongol Bichig Editor) for DOS by Peter Cheung (aka dEgi, Taiwan, 1992) [8], MLS (Mongolian Language Support) for UNIX/MSDOS by Oliver Corff [13], QAGUCIN traditional Mongolian script editor for Microsoft Windows by Michael Warmuth [14], Manchu Script Creator of Anaku [11], BabelPad Unicode text editor for Windows [1], Manchu IME, ManjuKey by Vincent Magiya [10], the Mongolian-English Dictionary by Lingua Mongolia [16], A Method for Electronizing the Traditional Mongolian Script by Man et al [9], Mongolia Language system by Chigen [4], UNESCO project "Development of E-tools for the Traditional Mongolian Script Text Processing" by National University of Mongolia [3] and Waseda research [17].


As a result, several encoding and code pages were introduced. In fact, the designer of an encoding scheme had decided where all shapes should be located. Several code pages were surveyed and example character sets are shown in Fig.3. Most approaches used in ASCII representations are by overwriting Basic Latin (U+0020~U+007E) and Latin-1 Supplement (U+00A0~U+00FF) code range. They developed traditional Mongolian fonts based on their codepage. Complexities are: each shape of different codepage (character set) does not match with each other thus complicated conversion techniques were developed based on that character set.

Some systems use transcriptions and transliteration technique in addition to the codepage encoding. Other systems included all canonical letters in codepage while canonical letters may be composed out by several glyphs in some systems. Some sample encoding of a Mongolian word **ᠠᠨᠠᠭᠠ**(abu: father) is shown in Table 1. However, several systems such as BabelPad, Manchu IME and TMSDL are based on Unicode. Some Unicode fonts including Code2000, Manchu2005, Simsun-18030, Mongolian Baiti and MongolScript were made as instructed in the Microsoft guidelines [2] and Unicode standard [5].

Code	Manchu Script Creator	QAGUCIN	CMs	MLS, Dula, Chigen	Code	Manchu Script Creator	QAGUCIN	CMs	MLS, Dula, Chigen	Code	Manchu Script Creator	QAGUCIN	CMs	MLS, Dula, Chigen
U+0021	ᠠ	ᠠ	ᠠ	ᠠ	U+0058	ᠠ	ᠠ	ᠠ	ᠠ	U+0071	ᠠ	ᠠ	ᠠ	ᠠ
U+0041	ᠠ	ᠠ	ᠠ	ᠠ	U+0059	ᠠ	ᠠ	ᠠ	ᠠ	U+0076	ᠠ	ᠠ	ᠠ	ᠠ
U+0042	ᠠ	ᠠ	ᠠ	ᠠ	U+005D	ᠠ	ᠠ	ᠠ	ᠠ	U+0061	ᠠ	ᠠ	ᠠ	ᠠ
U+0043	ᠠ	ᠠ	ᠠ	ᠠ	U+005E	ᠠ	ᠠ	ᠠ	ᠠ	U+0062	ᠠ	ᠠ	ᠠ	ᠠ
U+004D	ᠠ	ᠠ	ᠠ	ᠠ	U+006F	ᠠ	ᠠ	ᠠ	ᠠ	U+0063	ᠠ	ᠠ	ᠠ	ᠠ
U+0054	ᠠ	ᠠ	ᠠ	ᠠ	U+0070	ᠠ	ᠠ	ᠠ	ᠠ	U+0064	ᠠ	ᠠ	ᠠ	ᠠ

Fig. 3. Example character set of various encoding schemes

Within many code pages, CMs (Classical Mongolian script) by Peter Cheung (aka dEgi, Taiwan) became de facto for DTP applications, digital typesetting and word processing in Mongolia until Microsoft Vista's built-in IME support was released. A set of TrueType font CMs Ulaanbaatar, CMs Huree and CMs Urga (1998) is widely used because of its simplicity: no need of extra keyboard drivers and programs. Thousands of contents are stored in this codepage and in CMs fonts. Orthography dictionary of UNESCO project [3], Lingua Mongolia Mongolian-English Dictionary [16] and many more systems are utilizing CMS codepage. Thus, developing CMs support scheme for TMSDL is becoming crucial matter in terms of content enrichment of TMSDL and disambiguation of words which have same pronunciation by using orthography dictionary [3].

**Table 1.** Mongolian word  in different encoding schemes

Encoding Scheme (Codepage)	Encoded Text	Note
Manchu Script Creator	yaB1	Stores transcriptions, converts to Mongolian at the time to display
CMs (Classical Mongolian script)	FeB	Widely used in Mongolia
GB8045-87	@M	Replaced by GB 18030-2000
QAGUCIN traditional Mongolian script editor	Abü	Stores transcriptions, converts to Mongolian at the time to display

### 3 Traditional Mongolian Script Digital Library (TMSDL)

#### 3.1 Overview

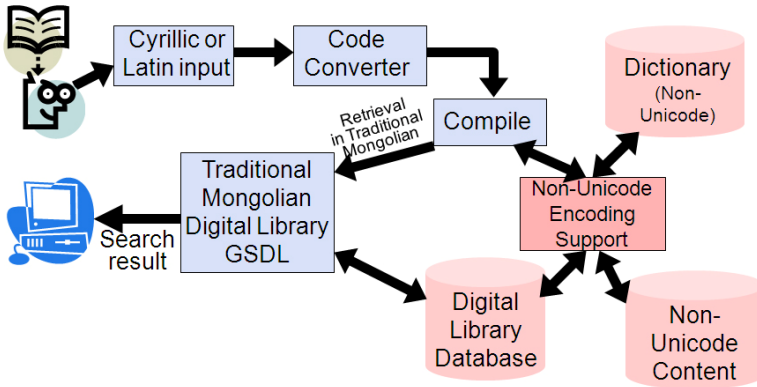
In this section we will introduce the traditional Mongolian script digital library (TMSDL). We utilized Greenstone Digital Library (GSDL), developed by New Zealand Digital Library Consortium at the University of Waikato [7]. GSDL is an open source software for building and distributing digital library collections.

We have enhanced the TMSDL, because of substantial needs to develop supporting system for diverse non-Unicode encoding in the TMSDL. The adapted structure of the TMSDL is shown in Fig. 4. Non-Unicode (include CMs) encoding support has following two main functions:

- a) Supports disambiguation of words by using non-Unicode orthography dictionary during retrieval and;
- b) Converts non-Unicode content into Unicode traditional Mongolian script.

#### 3.2 Cyrillic Code Converter and Traditional Mongolian Retrieval in GSDL

One of the main functions of a digital library is the search engine. Currently, Input Method Editor (IME) is not available for traditional Mongolian script text input. If we take into account that it is relatively easy to find Cyrillic IME, code converter should be used in our digital library's search engine and user will input keyword(s) in Cyrillic.



**Fig. 4.** The structure of Traditional Mongolian Script Digital Library. General architecture consist of user search input interface, converter, compiler, GSDL core and display interface.

Content is stored in un-coded traditional Mongolian basic forms and not variations, since Unicode standard includes only the basic character set. Thus our system converts modern Mongolian query to traditional Mongolian and displays correct variant form, when user input Cyrillic search text. Then traditional Mongolian query is retrieved from the GSDL. The collection which we have built is shown in the Fig. 5.



**Fig. 5.** Cyrillic code converter and Traditional Mongolian retrieval in Cyrillic input

Our approach is not to touch GSDL source code and not to modify standard macro files. Instead we created collection specific macro file extra.dm to add or override our functions.

### 3.3 Code Converter in Display Interface

We use code converter to display already stored basic characters correctly since content is stored in basic forms. Example of conversion is shown in Fig. 6.

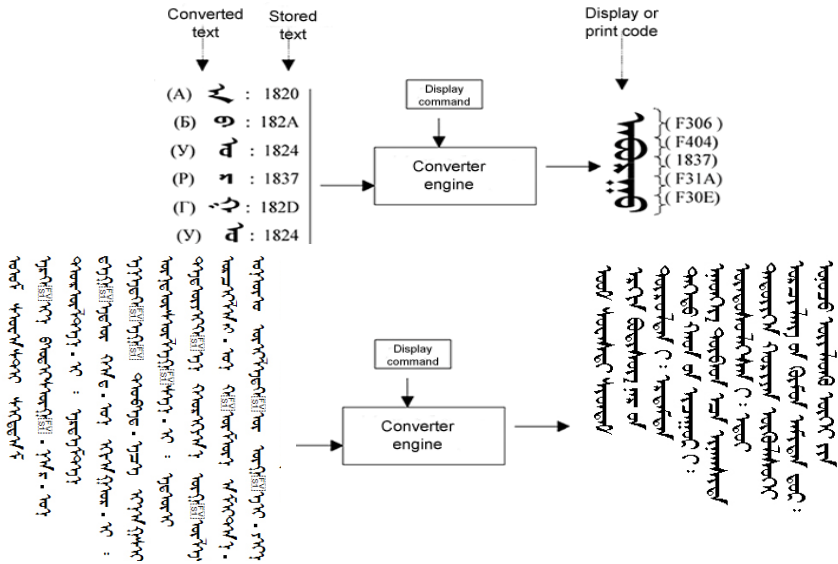


Fig. 6. Converter engine for traditional Mongolian script [5]

The control symbols are used to resolve ambiguities in some cases where the context rules are inadequate or to override the default forms. Those are the Mongolian free variant selectors (180B-<sup>[FV]</sup><sub>[S1]</sub>, 180C-<sup>[FV]</sup><sub>[S2]</sub>, 180D-<sup>[FV]</sup><sub>[S3]</sub>) for selecting alternative variants of a given positional form, and the Mongolian vowel separator 180E-<sup>[M]</sup><sub>[VS]</sub>. The Mongolian vowel separator serves as a distinguisher of the vowels “A” and “E”. It is because, these two vowels are written exactly the same when they are placed at the end of a word. Examples shown in Fig. 7 illustrate the use of the Mongolian vowel separator <sup>[M]</sup><sub>[VS]</sub>.

Character sequence	Display	Character sequence	Display
... ᠠ <sup>[M]</sup> <sub>[VS]</sub> ᠠᠨᠢ	ᠠᠨᠢ	ᠠ ᠠᠨᠢ	ᠠᠨᠢ
... ᠡ <sup>[M]</sup> <sub>[VS]</sub> ᠠᠨᠢ	ᠡᠨᠢ	ᠡ ᠠᠨᠢ	ᠡᠨᠢ
... ᠢ <sup>[M]</sup> <sub>[VS]</sub> ᠠᠨᠢ	ᠢᠨᠢ	ᠢ ᠠᠨᠢ	ᠢᠨᠢ
... ᠣ <sup>[M]</sup> <sub>[VS]</sub> ᠠᠨᠢ	ᠣᠨᠢ	ᠣ ᠠᠨᠢ	ᠣᠨᠢ
... ᠤ <sup>[M]</sup> <sub>[VS]</sub> ᠠᠨᠢ	ᠤᠨᠢ	ᠤ ᠠᠨᠢ	ᠤᠨᠢ

Fig. 7. Use of the Mongolian vowel separator [5]

The Mongolian free variant selectors are used to distinguish different variants of the same positional form of a character. They modify only the character immediately preceding them and will have no effect on the character following. Examples shown in Fig. 8 illustrate the use of the Mongolian free variant selectors.







Character sequence	Example of use	Character sequence	Example of use
ᠠ 	ᠠ	ᠠ	ᠠ
... ᠠ 	ᠠ	... ᠠ	ᠠ
ᠠ  ...	ᠠᠠᠠ (traditional form)	ᠠ ...	ᠠᠠᠠ
... ᠠ 	ᠠᠠ	... ᠠ	ᠠᠠ

Fig. 8. Some uses of the free variant selectors [5]

We developed an algorithm to display the traditional Mongolian characters correctly using control characters and/or basic characters. Then we integrated our algorithm into GSDL.

## 4 Experiments

### 4.1 Experiment on Retrieval

After creating our collection, we tested our system by inputting Cyrillic query. We compared our retrieval result with statistics of the textological study of source document.

We tested with major grammars of traditional Mongolian. Also we tested several most repeatedly used noun and numeral.

Retrieval was successful for given traditional Mongolian words and retrieved word counts were matched with source document. Thus our algorithm and modern Mongolian to a traditional Mongolian query converter works perfectly for commonly-used traditional Mongolian words. Some examples of search keywords are shown in Table 2.

Table 2. Example of query result

Cyrillic Input	Traditional Mongolian Query (meaning)	Retrieved	All numbers
Эзэн	ᠡᠵᠡᠨ (Lord)	146	146
Жил	ᠵᠢᠯ (year)	86	86
Энэ	ᠡᠨᠢ (this)	86	86
Зарлиг	ᠵᠠᠷᠯᠢᠭ (order)	65	65
Хан	ᠬᠠᠨ (prince)	61	61

### 4.2 User Evaluation

We conducted a user evaluation in order to assess the usefulness of our system. The evaluation was carried out with 9 native Mongolian participants. Of these, 5 participants (A-E in Table 3) are students living in Japan, and 4 (F-I) are researchers of

Mongolian language or script living in Mongolia. The results of the evaluation are shown in Table 3. In the table, evaluation scores are from 1 (worst) to 5 (best).

The results of the user evaluation clearly show that our system obtained high ratings in all inquiry items. On the other hand, the ratings for the inquiry of “Display of page texts” were relatively low, which is because the characters cannot be displayed vertically due to the limitation of current web browsers.

**Table 3.** The results of user experiments

Inquiries	Participants									Average
	A	B	C	D	E	F	G	H	I	
Retrieval using Cyrillic keywords	5	4	5	4	4	5	5	5	5	<b>4.6</b>
Usefulness of the proposed method	4	4	5	4	4	5	5	5	5	<b>4.5</b>
Display of page images	5	5	5	5	5	5	5	5	5	<b>5</b>
Display of page texts	4	4	4	4	3	5	5	5	5	<b>4.3</b>
Usefulness compared with other methods	5	4	5	4	4	5	5	5	5	<b>4.6</b>
Overall evaluation of the system	5	5	4	5	4	5	5	5	5	<b>4.7</b>

## 5 Conclusion and Future Work

In this paper, we have described the development of Traditional Mongolian Script Digital Library (TMSDL). Besides, we have surveyed existing approaches on development of traditional Mongolian information processing. We found that there still exists disparities between Unicode based approaches and other developer’s application. Non-Unicode encoding is still used and thousand of digital contents are stored in various encoding. We have analyzed common encoding and code pages. Developing non-Unicode support scheme for TMSDL is becoming crucial matter for future content enrichment of TMSDL. Support for CMs encoding need to be developed at the earliest step to utilize traditional Mongolian dictionaries [3][15].

## References

1. BabelPad: Unicode Text Editor for Windows, <http://www.babelstone.co.uk/>
2. Creating and Supporting OpenType Fonts for the Mongolian Script, <http://www.microsoft.com/typography/otfntdev/mongolot/>
3. Development of E-tools for the Traditional Mongolian Script Text Processing, UNESCO project, [http://portal.unesco.org/ci/en/ev.php-URL\\_ID=20774&URL\\_DO=DO\\_TOPIC&URL\\_SECTION=201.html](http://portal.unesco.org/ci/en/ev.php-URL_ID=20774&URL_DO=DO_TOPIC&URL_SECTION=201.html)
4. Electronizing Project of Mongolian (in Japanese), [http://texa.human.is.tohoku.ac.jp/~chigen/md\\_cnt\\_j.htm](http://texa.human.is.tohoku.ac.jp/~chigen/md_cnt_j.htm)
5. Erdenechimeg, M., Moore, R. M., Namsrai, Y.: UNU/IIST Technical Report No. 170 – Traditional Mongolian Script in the ISO/Unicode Standards (1999)

6. Garmaabazar, K., Maeda, A.: Retrieval Technique with the Modern Mongolian Query on Traditional Mongolian Text. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 478–481. Springer, Heidelberg (2006)
7. Garmaabazar, K., Maeda, A.: Building a Digital Library of Traditional Mongolian Historical Documents. In: 7th ACM/IEEE-CS Joint Conference on Digital Libraries, USA, p. 483. ACM Press, New York (2007)
8. GESER.STUDIO.COM, <http://www.geocities.com/geseree/>
9. Man, D., Fujii, A., Ishikawa, T.: A Method for Electronizing the Traditional Mongolian Script and Its Application to Text Retrieval. The IEICE Transactions D-II 88(10), 2102–2111 (2005)
10. Manchu IME.: ManjuKey and ManchuFont, <http://groups.msn.com/Manchu>
11. Manchu Script Creator, <http://www.anaku.cn/eng/download.php>
12. Mongolian Fonts, McGill University, <http://cg.scs.carleton.ca/~luc/mongolian.html>
13. Mongolian Language Support (MLS) for UNIX/MSDOS, <http://userpage.fu-berlin.de/~corff/im/MLS/overview.MLS.html>
14. QAGUCIN; Traditional Mongolian Script Editor, <http://members.aol.com/ayuu/download.html>
15. Standardization Administration of the Peoples Republic of China, <http://www.sac.gov.cn/>
16. The Mongolian-English Dictionary, <http://www.linguamongolia.co.uk/>
17. Kataoka, T.I., Kataoka, Y., Uezono, K., Ohara, H.: Internationalized text manipulation covering perso-arabic enhanced for mongolian scripts. In: Hersch, R.D., André, J., Brown, H. (eds.) RIDT 1998 and EPub 1998. LNCS, vol. 1375, p. 305. Springer, Heidelberg (1998)
18. The Unicode Consortium: The Unicode Standard 4.0. Addison-Wesley, Boston (2003)

# Weighing the Usefulness of Social Tags for Content Discovery\*

Khasfariyati Razikin, Dion Hoe-Lian Goh, Chei Sian Lee,  
and Alton Yeow Kuan Chua

Wee Kim Wee School of Communication & Information, Nanyang Technological  
University, 31 Nanyang Link, Singapore 637718, Singapore  
{khasfariyati, ashlgoh, leecs, altonchua}@ntu.edu.sg

**Abstract.** A new wave of social computing applications has empowered users to create and share a variety of content. This upsurge of user-generated data involves a paradigm shift in terms of the management, searching and accessing of information. Social tagging is one of these ways. This paper serves as an extension to the existing work done on investigating the effectiveness of tags for content discovery using text categorization techniques. In particular, we explored how different tag weighting schemes affect classifier performance. Six text categorization experiments were conducted using a dataset drawn from del.icio.us. The results suggest that not all tags are useful for content discovery even with different weights associated with them. Content analysis was done to understand the relationships between the use of a tag on a document and the document's terms. Implications of this research are discussed.

**Keywords:** Social tagging, Social computing, Web 2.0, Content discovery and organization, Effectiveness, Text categorization.

## 1 Introduction

Social computing/Web 2.0 applications empower users to create and share a variety of multimedia content including text, images and video. As this new avenue for content creation becomes increasingly popular through services such as blogs, wikis and media sharing, the resulting upsurge of user-generated content requires new approaches for their management, search and access. Beyond established tools such as search engines and directories, social tagging is a promising approach for organizing and discovering content on the Web. It allows users to annotate useful content with uncontrolled keywords (tags), facilitating their future access by the tag creator [12]. In addition, tags may also be shared by other users of the social tagging system, providing an alternative way to discovering and accessing content on the Web.

Social tagging differs from conventional methods of content organization based on taxonomies, controlled vocabularies, faceted classification and ontologies. These methods require experts with domain knowledge, and are rule-bound to ensure that the classification schemes created are consistent [14]. In contrast, tags are “flat”, lacking a

---

\* This work is partly funded by A\*STAR grant 062 130 0057.

predefined taxonomic structure, and their use relies on shared, emergent social structures and behaviors, as well as a common conceptual and linguistic understanding within the community [13]. Hence, tags are also known as “folksonomies”, short for “folk taxonomies”, suggesting that they are created by lay users, as opposed to domain experts or information professionals such as librarians.

This unstructured approach to organizing content has drawn criticism by some researchers. For example, the flexibility of free keyword assignment may result in ambiguity of meaning due to a lack of controlled vocabulary [12], resulting in the problem of vocabulary mismatch [5] between tag creators and users [6, 15]. Further, content may be tagged with subjective or ego-centric terms (e.g. “cool”, “todo”, “me”, “toread”) that have meaning only for the tag creator or a select few within a group of users. Tags may also sometimes be driven by the tag creator’s self-serving agenda [3], leading to problems such as tag spamming where popular but irrelevant tags are deliberately used to attract traffic to certain Web sites [9]. Taken together, these issues may hinder the effective use of tags for organizing and sharing content.

Despite these shortcomings, the use of social tagging continues to grow in popularity. Concurrently, there is an emerging body of research that explores their effectiveness for content organization and sharing. For example, from a user’s perspective, work has also been conducted on motivations on behind tagging [1], comparing the use of tags against author assigned index terms in academic papers [8], and on tagging dynamics and usage [4]. Automated, machine learning approaches have also been used to study the ability of tags to classify blogs using text categorization methods [19], and on investigating the effectiveness of tags to classify Web resources in *del.icio.us* [16].

In this paper, we extend existing work in investigating the effectiveness of tags for content discovery using a machine learning approach (e.g. [17, 19]). In particular, we explore different tag weighting schemes to determine if these affect performance, which is relatively under-researched. Further, to better understand how tags are created, we conduct a content analysis to study the relationships between the use of a tag on a document, and the document’s terms. To the best of our knowledge, there are a limited number of studies that have been done on examining the effectiveness of tags for content discovery using both a machine learning and content analytic approach. The results of our work can be used to tune automated techniques that help users in both seeking content via tags, as well as suggest tags for organizing content. The remainder of this paper is organized as follows. In the next section, we review research related to the present study. We then describe our experimental methodology and present the results. Next, we provide a discussion of the implications of our findings and conclude with opportunities for further work in this area.

## 2 Related Work

We focus our review on related literature that investigates effectiveness of tags as a means for organizing and sharing content.

From a tag creator’s perspective, work has been done to compare tags with controlled vocabularies. Tags from Connotea (<http://www.connotea.org/>) and Medical Subject Heading (MeSH) terms were evaluated by [11] who found that there was only

11% similarity between MeSH terms and tags. This was because MeSH terms served as descriptors while tags primarily focused on areas that were of interest to users. Likewise, [8] compared tags with author supplied tags from Cite-U-Like and indexing terms from INSPEC and Library Literature to determine the usage overlap. Results showed that approximately 21% of the tags were the same as the indexing terms. The reason for the divergence was attributed to the different emphases placed on an article by these two groups. For example, tag creators may consider time management information (e.g. “todo”, “toread”, “maybe”) to be important as a tag to indicate a desire to read them in the future, while such information will be disregarded by expert indexers. Together, these findings suggest that tag creators and indexing experts employ vocabularies that have little overlap, potentially causing access problems in social tagging systems.

Research has also been conducted on tag effectiveness using different machine learning approaches. For example, [2] used 350 popular tags and 250 of the most recent blog articles from Technorati. Clustering was done on these articles, and the results suggested that tags were able to organize articles in a broad sense, but not as effective in indicating the specific content for an article. Besides blogs, [17] studied the effectiveness of tags to classify Web content from 100 tags and 20210 documents in del.icio.us. Using Support Vector Machines, experiments were run on two feature sets: document terms only, and document terms plus tags. Surprisingly, results indicated that using document terms only produced better classification results in terms of F-measure than using terms plus tags. Nevertheless, both F-measures from the experiments were relatively low at 0.59 and 0.52, suggesting that not all tags were effective at content discovery, and that the classifier’s performance was likely to be influenced by the tag creator’s motivations, and his/her interpretation of the document content. Finally, [10] investigated tags as a source for metadata to describe music. Using 236974 tags collected for 5722 tracks from last.fm and MyStrands, a Correspondence Analysis was performed to visualize a two-dimensional semantic space defined by the tags. Findings suggest that tags were effective in capturing music similarity, and could be used to describe mood and emotion in music.

While our present study shares the goal of investigating tag effectiveness with the above work, our research is differentiated in several important ways. Firstly, the studies by [11] and [8] were limited to scholarly articles while [2] used blogs. Here, the context of the medium of communication differs in our study. Specifically, in our dataset, the pages that are tagged in del.icio.us are more diverse, and not limited to ordinary Web pages, but also include blogs and academic articles. Put differently, we capture a wider spectrum of content found on the Web. While [17] employed del.icio.us data, it did not examine the influence of different tag weighting schemes, and this may have impacted the results, since the number of tags was significantly fewer than the terms within a document. Taken together, the present study is thus timely as we extend existing work, and perform a more comprehensive analysis of the effectiveness of tags for content discovery.

### 3 Experimental Setup

The dataset used in these experiments is based on the one used in [17]. The tags and Web documents were harvested from del.icio.us, a popular social tagging service,

from August 2007 and October 2007. During this period, 100 tags and 20210 English language documents were collected. Following [2], the tags were mined starting from the popular tags page. As such our tags will be biased towards the more commonly used ones. Nevertheless, by using popular tags, we were assured that there are a significant number of documents related to each tag that will provide a sufficient dataset size for our experiments.

The tags that were mined consisted of single token terms. The collected documents were processed by removing the HTML elements, JavaScript codes and Cascading Style Sheets elements. This was followed by stop word removal and stemming of the remaining words. For each tag, we selected all the documents that were tagged with the keyword and these were grouped as the positive samples for that particular tag. An equal number of documents, which were associated with a different tag, were randomly selected as negative samples. From this set of positive and negative samples, two-thirds of the documents were used as the training sample while the rest were part of the test set. We made use of the SVM<sup>light</sup> [7] package for the classification experiments. Binary classifiers were utilized for the study and the default parameters of the SVM package were used during the experiments.

In all, six text categorization experiments were conducted. The first experiment made use of only terms from the documents as features. This experiment served as a baseline for the other experiments due to its use of fundamental features. Here, the TF-IDF values of the terms were used as the feature set. The five subsequent experiments included tags, in addition to terms, as part of its feature vector. The TF-IDF values of the tags were increased by one, three, five, seven and ten times to reflect the differing importance of the tags relatively to the rest of the feature set. The motivation for increasing the TF-IDF values would help determine if this would help in the performance of the classifier.

The output from the classifier, which are precision, recall and F-measure were used to determine the effectiveness of the tags and the different weighting schemes. The performance of the tags was evaluated based on the macro-averaged and micro-averaged precision, recall and F-measure. Macro-averaged values give an indication of the overall performance of the classifier over all tag categories. Micro-averaged values, on the other hand, measure the performance over each document. It emphasizes on the performance of tags that have a larger number of documents.

## 4 Results

The results obtained for all the experiments are shown in Table 1. The values in bold indicate the highest value obtained among all the six experiments. Surprisingly, as observed from the table, the experiment whose feature vector consisted only of terms obtained the highest scores for the majority of the performance measures. There is a general downward trend for the majority of the performance values as the weighting of the tags were increased. An exception would be the precision values. Apart from this, it appears that the addition of tags as feature vectors and increasing their weights do not help in the improvement for identifying documents associated with the tags. Figures 1 and 2 illustrate the trends for the macro-averaged and micro-averaged values among all experiments respectively.

In terms of macro-averaged values, the precision values obtained were surprisingly almost constant despite increasing the weights of the tags. The increase in precision for the tag weight of three and ten are relatively insignificant with respect to the rest of the precision values obtained. Put differently, increasing the tags' weights does not help in improving the precision of the classification. The recall values show that increasing the weight of the tags does not improve classification performance. Instead, misclassified errors increased as the weights increased. A similar trend is observed for the F-measure values.

**Table 1.** The macro- and micro-averaged, and standard deviation values for precision, recall and F-measure. The values in bold are the highest values obtained for all experiments.

Experiment	Macro-averaged			Micro-averaged		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Terms only	<b>52.66</b> (sd = 4.21)	<b>54.86</b> (sd = 19.05)	<b>52.05</b> (sd = 10.99)	<b>64.76</b>	54.40	<b>59.14</b>
Tag Weight 1	52.56 (sd = 6.06)	51.60 (sd = 20.75)	50.10 (sd = 13.21)	56.47	<b>54.93</b>	55.69
Tag Weight 3	52.67 (sd = 7.38)	44.04 (sd = 22.94)	44.96 (sd = 15.47)	54.78	48.76	51.60
Tag Weight 5	52.01 (sd = 10.15)	37.93 (sd = 24.90)	39.53 (sd = 18.83)	54.80	42.84	48.09
Tag Weight 7	51.13 (sd = 13.72)	34.76 (sd = 26.26)	36.23 (sd = 21.14)	55.00	39.81	46.19
Tag Weight 10	51.23 (sd = 16.27)	32.38 (sd = 27.08)	33.60 (sd = 22.69)	55.30	37.78	44.89

Micro-averaged performance values, in general, exhibit similar trends as the macro-averaged values. Here, the precision value obtained was at 64.74% in the terms only experiment. The precision values obtained for the subsequent experiments were lower than the first and were almost constant, decreasing when the weight of the tag was set to three, but increasing slightly thereafter. For recall, there was a slight improvement when tags were part of the feature vector. However, increasing the tag weights only made the recall values to deteriorate thereafter. Likewise, the results for micro-averaged F-measure values were similar to recall.

On the whole, these experiments provide a clearer picture on the effect of different tag weighting schemes and their influence on the performance of the classification process. The results show that the addition of tags does not help in improving precision, recall and F-measure values, suggesting further that not all tags could be effectively used for content discovery. There are three possible reasons for this. First, tag creators may have goals other than information discovery by other users [1], and therefore some tags could not be used for this purpose. For instance the documents that are associated with a user-centric, or subjective, tag such as “funny” would vary between tag creators. It was found that some of the documents that had been tagged with “funny” might be humorous to the tag creator, but tag consumers might find it otherwise. Next, social tagging does not have well-defined rules, and the vocabularies employed by different users could vary significantly, leading to inconsistencies in associating documents with tags, impeding effective retrieval of documents [12]. A blog that focuses its contents on cooking and recipes was found to be associated with a diversity of tags such as “blogs”, “blog”, “foodblog” and “foodblogs”. Lastly, another related reason to account for the



poor performance of the classifiers is due to the polysemic nature of tags [6]. As a result, users may impute different interpretations to the same tag so that documents that are associated with a particular tag might not be semantically related at all. For instance, the tag “ruby” might refer to the gem for the uninitiated. However, to a Web applications developer, the tag may refer to a programming language instead.

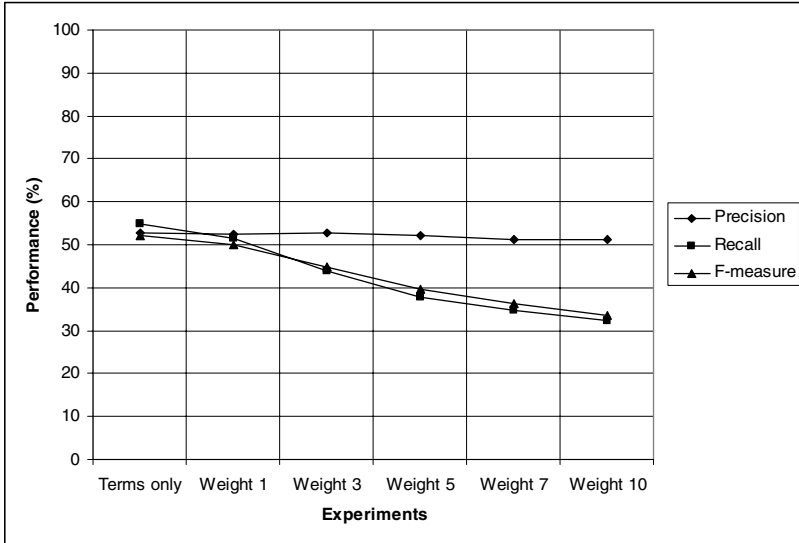


Fig. 1. The macro-averaged performance for all the experiments

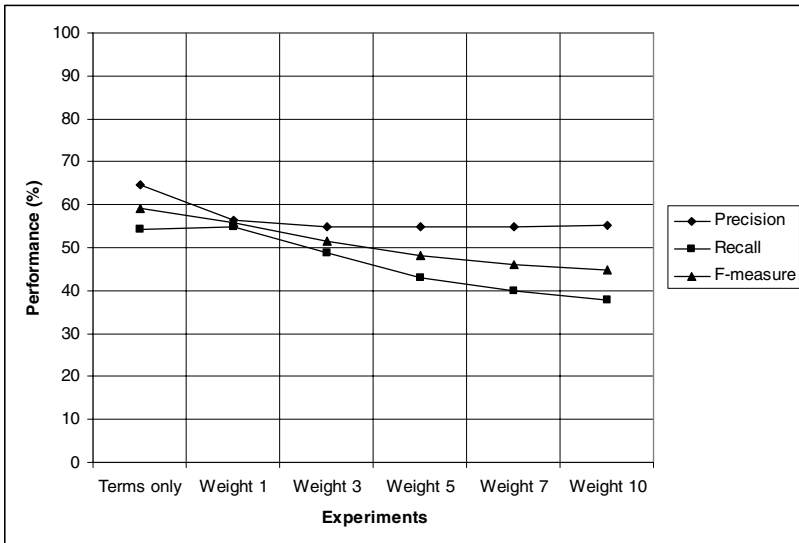


Fig. 2. The micro-averaged performance for all the experiments

## 5 Tag and Document Analysis

To better understand the performance of the classifiers, we manually analyzed the content of 20 tags to uncover distinguishing patterns. This set comprised the top ten best and poorest performing tags by averaging the F-measure values obtained for all the six experiments (see Table 2).

**Table 2.** The ten top ten and bottom ten tags and their average F-measure values

<b>Top Ten Tags</b>	<b>Average F-measure</b>	<b>Bottom Ten Tags</b>	<b>Average F-measure</b>
news	69.10	adobe	20.89
politics	66.78	server	19.45
blogs	65.94	portfolio	19.45
software	64.89	windows	18.62
web2.0	64.45	journal	18.35
interesting	64.42	actionsript	17.68
site	64.17	comic	16.98
resources	63.95	xml	16.42
california	63.62	download	14.83
foodblog	62.20	ajax	13.79

At a glance, the ten best performing tags seemed to have rather broad meanings. In contrast, the bottom ten tags appear to have a narrower definition. It is also interesting to note that at least seven of these tags are computer related. The selected documents were analyzed based on generic/distinctive, subjective/objective characteristics, and the type of medium the content was found in. Our findings are elaborated in the following paragraphs.

Among the top 10 tags, those such as “news”, “site” and “resources” seemed to be more generic in comparison with the rest of the tags. Here, it appears that there is a common understanding among the community of tag creators on the meanings behind these generic tags [11], while the technical tags seem to suffer from more uncertainty in their meanings, given the results in Table 2. Further, the meanings of the top 10 tags appear to be more easily understood by a majority of users as opposed to the poorer performing technical tags whose meanings are presumably known only by those familiar with the field (e.g. “actionsript” and “ajax”).

In addition, a generic tag implies a wider variety of topics associated with it, which also suggests that the terms found in these documents would be varied. Yet, the classifiers performed well on average. A reason for this might be that the tags did not occur frequently in the entire corpus as terms, causing them to have significant TF-IDF values. For example, the subject content associated with the “news” tag was diverse, including the New York Times. While the terms were diverse owing to variety of topics the newspaper accommodates, the term “news” itself did not occur frequently. In contrast, the bottom ten tags had much narrower definitions and were reflected in the associated documents as well. For example, some documents that were tagged with “server” included a tutorial on Microsoft SQL Server and an introduction to Mono. The tag appeared quite frequently in the documents for these two cases. In fact, the tag is the most common word in the dataset, and could be one of the reasons why it performed poorly.

Next, subjective tags are those which could be adjectives, verbs or have other intrinsic qualities, while objective tags are nouns and exhibit extrinsic characteristics [6]. “Interesting” is the only subjective tag that performed well in our experiments. The other nine tags were objective. As found by a prior analysis on subjective tags [17], documents associated with this tag were not topic-specific. Additionally, the documents rarely contained the tag “interesting”, thus making the tag a useful term, leading to a good classifier performance. The bottom ten tags also only had one subjective tag that is “download”. An analysis of the documents associated with the tag revealed that the term appears frequently in the dataset. This finding is in contrast to the best performing subjective tag. Perhaps it is due to the commonness of the tag that “download” performed rather poorly. Similar patterns of findings were observed for the objective tags ranked in the top and bottom ten. However, the documents associated with objective tags were found to be more subject-specific.

Finally, tag creators often associate a document with a tag that defines the type of media the document can be found in [6]. Some examples in the best performing tags were “blogs” and “foodblog”. This observation is true in a social tagging system like del.icio.us where the documents are not restricted to just HTML pages. While the tag “blogs” suggests a diverse range of topics due to its broad meaning, “foodblog” tends to concentrate more on the notion of food. Some examples of documents which are associated with “foodblog” include those which provide recipes, reviews and recommendations of eating establishments, and recipe books and cooking tips. A similar trend was observed in the bottom ten tags, namely, “comic”, “journal” and “portfolio”. A reason for the difference in performance could be due to the amount of text found. In “blogs” and “foodblog”, a significant amount of the content was made up of text. For the others (“comic” and “portfolio”), there was less textual content and more multimedia content, thus affecting classifier performance. Here, other content-based techniques would be needed to address this shortcoming.

In sum, four observations may be drawn on using tags for content discovery: (1) generic tags might surprisingly be better for retrieving relevant documents than specific ones; but more possibly, (2) tags with more accessible meanings are likely to be understood and adopted by the community of tag creators and users; (3) tags which do not appear frequently in the corpus should be recommended for use over those that appear commonly; and (4) automated techniques that suggest tags to users (e.g. [4]) should consider the need for analyzing multimedia content as well.

## 6 Discussion and Conclusion

To reiterate, the aim of this paper is to investigate the effectiveness of social tags for discovery of relevant content. Six text categorization experiments were conducted in the present study using different tag weighting schemes. Our results indicate that the terms only experiment performed the best in terms of precision, recall and F-measure, further reinforcing existing research that suggest that tags may serve purposes other than content discovery, and that caution must be exercised by tag consumers in their search for relevant information [1, 17].

Our work has yielded the following conclusions. First, introducing tags as feature vectors and increasing their weights does not help in content discovery as the values for precision, recall and F-measure deteriorate, implying that not all tags are useful for

content discovery. This appears to be in contradiction with the Web 2.0 notion of the benefits of harnessing the collective intelligence of users. A reason for this could be related to the level of familiarity of the tag creators with the different aspects of social tagging and/or the nature of the topic itself. A novice tag creator might use ineffective tags or those that are not drawn from the vocabulary of the tagging community, impeding future retrieval by others [18]. Another factor could be due to the polysemic, synonymous and homonymous nature of tags [6]. Here, there would be difficulty in disambiguating the tags without the need for looking at the context of the associated document. As such, measures should be put in place when using tags to search relevant documents to ensure that ambiguity in meanings are addressed. For example, the system could prompt the user to determine the exact context of the ambiguous tag, thus helping them obtain relevant documents that meet their information needs. Next, the best performing tags are seemingly generic tags with broad meanings. Our content analysis suggests that their meanings appear to be widely understood. Put differently, despite their generality, there is some stabilization in terms of tag usage in which tag creators tend to select tags which are commonly understood by other members of the del.icio.us community [18]. For example, a tag creator may apply similar tags to a document previously given by other tag creators. Thus, simply recommending popular tags to users might not be an appropriate method. Instead, recommendations based on the patterns of knowledgeable members of the tagging community would be preferable. This would ensure that the tags have more relevancy and semantic relation to the document in question [4].

There are limitations in our study that could be addressed in future work. First, all terms appearing in the documents, with the exception of stopwords, were deemed useful. There could be a possibility that the classifier could incorrectly generalize its model based on the existence of a rare term that does not have any meaning to the tag. Perhaps it might be helpful to identify which features are useful for a tag to eliminate noisy terms. Examples might include  $\chi^2$  or information gain [20]. Secondly, this study made use of only popular tags. However, the number of such tags is proportionately smaller than the entire collection of tags in del.icio.us. Future work could utilize a wider variety of tags to determine if performance may be affected, as less popular tags may be associated with more esoteric, but more specific concepts, and therefore could result in better classifier performance. In addition, research could be done on using tags to filter or rank documents associated with them. Lastly, future work could look into replicating the same experiments for different datasets such as Cite-U-Like or Amazon, where the intent of the tag creators could be different from that of del.icio.us, as well as extending the analyses to multimedia content such as those found in YouTube and Flickr.

## References

1. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: SIGCHI Conference on Human Factors in Computing Systems, pp. 971–980. ACM, New York (2007)
2. Brooks, C.H., Montanez, N.: Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proceedings of the 15th international conference on World Wide Web, pp. 625–632. ACM, New York (2006)

3. Chua, A.: Knowledge sharing: A game people play. *Aslib Proceedings* 55(3), 117–129 (2003)
4. Farooq, U., Kannampallil, T.G., Song, Y., Ganoë, C.H., Carroll, J.M., Giles, L.: Evaluating tagging behavior in social bookmarking systems: metrics and design heuristics. In: *Proceedings of the 2007 international ACM Conference on Supporting Group Work*, pp. 351–360. ACM, New York (2007)
5. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T.: The vocabulary problem in human-system communication. *Commun. ACM.* 30(11), 964–971 (1987)
6. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. *Journal of Information Science* 32(2), 198–208 (2006)
7. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Proceedings of the 10th European Conference on Machine Learning*, pp. 137–142. Springer, Berlin (1998)
8. Kipp, M.E.: Exploring the context of user, creator and intermediate tagging. In: *ASIS&T 2006 Information Architecture Summit* (2006)
9. Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P., Garcia-Molina, H.: Combating spam in tagging systems. In: *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pp. 57–64. ACM, New York (2007)
10. Levy, M., Sandler, M.: A semantic space for music derived from social tags. In: *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007* (2007)
11. Lin, X., Beaudoin, J.E., Bui, Y., Desai, K.: Exploring characteristics of social classification. In: *Proceedings 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research* (2006)
12. Macgregor, G., McCulloch, E.: Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review* 55(5), 291–300 (2006)
13. Marlow, C., Naaman, M., Boyd, d., Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: *Proceedings of the seventeenth conference on Hypertext and hypermedia*, pp. 31–40. ACM, New York (2006)
14. Morville, P.: *Ambient findability*. O'Reilly, Beijing (2005)
15. Olsen, K.A., Sochats, K.M., Williams, J.G.: Full text searching and information overload. *International Information and Library Review* 30(2), 105–122 (1998)
16. Razikin, K., Goh, D.H.-L., Cheong, E.K.C., Ow, Y.F.: The efficacy of tags in social tagging systems. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) *ICADL 2007*. LNCS, vol. 4822, pp. 506–507. Springer, Heidelberg (2007)
17. Razikin, K., Goh, D.H.-L., Chua, A.Y.K., Lee, C.S.: Can social tags help you find what you want? In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) *ECDL 2008*. LNCS, vol. 5173, pp. 50–61. Springer, Heidelberg (2008)
18. Sen, S., Lam, S., Rashid, A.M., Cosley, D., Frankowski, D., Osterhous, J., Harper, F.M., Riedl, J.: Tagging, communities, vocabulary, evolution. In: *Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work*, pp. 181–190. ACM, New York (2006)
19. Sun, A., Suryanto, M.A., Liu, Y.: Blog classification using tags: An empirical study. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) *ICADL 2007*. LNCS, vol. 4822, pp. 307–316. Springer, Heidelberg (2007)
20. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: *Proceedings of ICML 1997, 14th International Conference on Machine Learning*, pp. 412–420. Morgan Kaufmann Publishers, San Francisco (1997)

# A User Reputation Model for DLDE Learning 2.0 Community

Fusheng Jin<sup>1,2</sup>, Zhendong Niu<sup>1</sup>, Quanxin Zhang<sup>1</sup>, Haiyang Lang<sup>2</sup>, and Kai Qin<sup>2</sup>

<sup>1</sup> School of Computer Science and Technology, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, 100081, Beijing, China

<sup>2</sup> School of Software, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, 100081, Beijing, China

{jfs21cn, zniu}@bit.edu.cn, zhangqx@china.com.cn,  
Lhy762@yahoo.com.cn, kk860@sohu.com

**Abstract.** With the development of digital library, social networks, and user-generated content, need for trust and reputation models become prime. In this paper, we propose a user reputation model. As an encouraging and sanctioning mechanism, it has been applied to the DLDE (Digital Library and Digital Education) Learning 2.0 Community that is developed by our lab based on digital repositories management etc. The model combines user's individual activity analysis approach and collaborative activity analysis approach. Individual activity analysis approach is used to analyze the activities in which users participate individually and give its evaluation method. Collaborative activity analysis approach is used to analyze users' collaborative activities; three different categories of users' collaborative activities and corresponding evaluation methods were proposed in this paper. Experiments show that the proposed reputation model can accomplish the mission of encouraging good behaviors and differentiating the ability of students. Therefore it can fit well in our Community.

**Keywords:** User Reputation model, DLDE Learning 2.0 community, Collaborative activity analysis, Individual activity analysis, evaluation.

## 1 Introduction

Over the past decade digital libraries, electronic commerce, user generated content and recommendations, and social networking applications have become common on the web. As reliance on these applications grow the need for a trust and reputation model has become essential. Users look for recommendations on products, web site, suppliers, buyers, content, and peers based upon the nature of interaction. Recommendations are provided based upon personal experience, usefulness of prior recommendations, and other behaviors on the system. The roles and behaviors of any user in an online application define the trust that other users of the system have on the user. This also defines the reputation of the user in the community. [1]

Traditional online reputation systems usually focus on online trade systems such as eBay [6]. These systems are usually based on user ratings among the users, mainly

because the relations between the participants are very simple in online trade systems, which are limited to selling and buying. [2]

Our scenario is different. In this paper, In order to promote students to participate learning actively, we propose a user reputation model and apply it to the DLDE Learning 2.0 community, which is based on Digital Education Resource Repositories. In the DLDE Learning 2.0 community, users can post topics, reply to others' topics, write blogs, upload/download learning resources, rate for others, challenge the learning units etc. If a user plays well in the community, the user will fetch more score from the community and others. This kind of community is much more similar to a Digital Education Resource Repositories based E-Learning Platform from the view-point of providing an online learning environment. The main differences lie in the web2.0 features that are added in our community, users have become the content producers besides the content consumers.

Experiment shows that our reputation model can provide an incentive for good behaviors and differentiate the ability of students in the DLDE Learning 2.0 community and therefore have a positive effect on student learning.

This rest of the paper is organized as follows. Section 2 introduces the activities in the DLDE Learning 2.0 Community. Section 3 presents the reputation model and the computation method for user's reputation. Experiment in our community for reputation model is presented in section 4. Section 5 concludes the whole paper and discusses the future work.

## 2 The DLDE Learning 2.0 Community

The DLDE Learning 2.0 Community (See Figure 1) is developed based on the Digital Education Resource repositories and the WEB2.0 concepts[11-13]. Within the community, users are not only consumers of resources but also resource producers. Therefore traditional resource repositories is no longer the only source of resources, every user can produce resources in various format. In this way, we can make full use of every user's expertise.

As in Figure 1, the DLDE Learning 2.0 Community's main purpose is to create a user-centered live learning environment based on digital resource repositories. Contribution, Participation and Collaboration are its main characters. In this community, each user has his personal space where users can maintain their profiles; customize their preferences, for example, their personalized interface and resources. Meantime, lots of tools are provided to users based on Web 2.0, such as RSS, BLOG, and BBS etc. Users can do what they like, but their activities are regulated by our reputation system. In the community, Incentive mechanism is used to encourage Contribution, Participation and Collaboration. Generally speaking, any positive activity in which users participate will bring them some scores and the increasing of reputation according to the extent of importance. The detail will be explained in below.

Different kinds of activities that users participate in the community will increase or decrease their score. The weights for the different activities are given arbitrarily according to empirical value that is from our community running experiments. The main activities include,

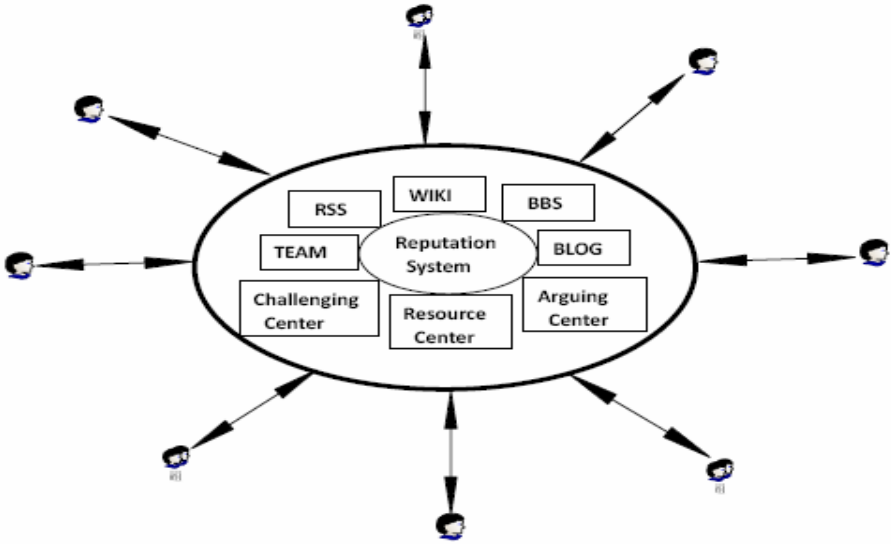


Fig. 1. Community structure

1. *Register*. New comers will obtain 10 scores after registration.
2. *Login*. If a user logs everyday, the user will achieve 2 scores. But the score will be counted only once when the user logs more than one time in a day.
3. *Punish*. If the topic or the reply is deleted the user will be fined 50 scores for his activity. Those who attempt and accomplish nothing will decrease 1 score per day until their score reaches zero.
4. *Post and Reply*. Any user can post a topic for discussion, others can reply the topic if they are interested in and give their comments by rating. Range (-3-3).the final scores of poster is the average rating scores.
5. *Write*. Original articles and good resources are warmly welcomed in the community. It is the representation of innovation spirit that such a kind of resources will achieve a higher score. The scores are computed by formula:  $10 * \text{average}(\text{rating scores}) + 0.1 * \text{viewing times} + 0.1 * \text{replying times}$ .
6. *Top/Highlight of topics*. This kind of topics is usually very popular among users. Therefore the user who posted it will be rewarded 50 scores. This is determined by administrator.
7. *Opinion leader*. Those who post, reply, rate and view top 20 in a week are called opinion leader, who will get 50 scores as a reward. We count these four kinds of actions together for each user every week. Which is computed by formula:  $10 * \text{Post times} + 5 * \text{reply times} + 0.1 * \text{rate times} + 0.1 * \text{view times}$ .
8. *View a topic and Rating*. Rating is not necessary but is encouraged. In order to encourage participation and obtain a rating data from participants as many as possible, anyone who views one time will get 0.05 score and rates one time for others will receive 1 score. In the DLDE Learning 2.0 community, rating is classified by six levels, extremely bad, worse, bad, average, good, and excellent. Range (-3-3). Users can choose one to express their idea.



9. *Use multi-learning policies.* According to the Constructivism Learning Theory, A user who uses multi-learning policies during learning will learn more than those who do not use it. So those who use a kind of learning policy will get 10 scores in the learning.
10. *Upload learning resources.* Whether the resource is good or not is measured by rating, downloading times, viewing times. Rating is a more effective way to judge the resources so it has a higher weight than downloading times and viewing times. The scores is computed by formula:  $10 * \text{average}(\text{rating scores}) + \text{downloading times} + 0.1 * \text{viewing times}$
11. *Download learning resources.* Users can download the helpful learning resources by paying 10 scores per time to the system.
12. An *arguing center* used for complaining misbehaviors is established. This will make fraud easily detectable. Users who conduct in fraudulent activities that are confirmed will be punished severely. 200 scores in the first time, double in the second time ,and so on.
13. *Accomplish learning unit* (See Fig. 2). Users can challenge the learning unit by themselves; they also can form a team freely and share the unit score by rating arithmetic that we have proposed. (See formula (4)-(5)), the score is given by the course administrator.
14. *Offer a reward.* Any participant can find answers for their questions or find learning resources they are looking for by offering a reward. In this way, the scores are transferred from one user to another.

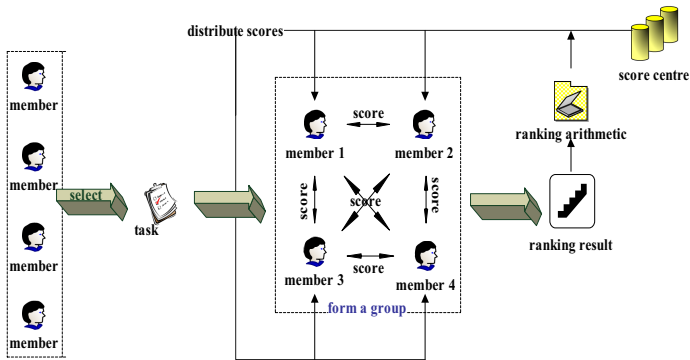


Fig. 2. Team activities

### 3 Users' Reputation Model

The reputation model is used to evaluate the reputation of users in the DLDE Learning 2.0 Community. It is equivalent to the score the user gains in the community, which mainly depends on what a user contributes. To encourage users to participate in the community, a system scores and a collaboration score are provided for each user. The system score is obtained by attending individual activities, such as post topics, write blogs, upload resources etc. The collaboration score can be obtained through collaborative activities, such as team activities and rating activities etc.

We use  $User\_score$  to represent the total score of a user,  $System\_score$  represents his/her system score and  $Collaboration\_score$  represents his/her collaboration score. The computation of a User's score is shown in the formula (1) :

$$User\_score = System\_score + Collaboration\_score. \quad (1)$$

### 3.1 System Score

As we mentioned before, in the DLDE Learning 2.0 community, users may participate in some activities in term of his personal interests. We name this kind of activities as individual activities. Each individual activity may increase or decrease their score. We call the score obtained by this way is the system score, and a user's system score is computed by formula (2).

In formula (2),  $Action\_score(i)$  represents the score that a user obtained by engaging in activity  $i$ ,  $User\_weight$  represents the user's weight.(the computation of  $User\_weight$  see section 3.3 )

Meanwhile, we use  $1 - User\_weight$  instead of  $User\_weight$  as the weight of computation, the reason is our encouraging policy that makes users achieve scores easily in the beginning stages and more difficult in the following stages.

$$System\_score = \sum_{i=1}^n Action\_score(i) * (1 - User\_weight) \quad (2)$$

### 3.2 Collaboration Score

In the DLDE Learning 2.0 community, user-centered, participation and collaboration are the main characters. Users are not only the resource producer but also the resource consumer. Every user can give comments to others and receive comments from others. The contribution of user is judged by rating, downloading and viewing activities etc. We define the activity that occurs among users as collaborative activity and the score that is attained by collaborative activity as collaboration score. We classify all the collaborative activities into three categories, team activities, rating activities and offering reward activities. According to the above three kinds of activities, the collaboration score is divided into three parts; Rank score which can be achieved by team collaboration, Encouragement score that can be gained by popularity computation and Transfer score that can be obtained through accomplishing offer reward. We name  $collaboration\_score$  as Collaboration score,  $Rank\_score$  as the score obtained by team working,  $Encouragement\_score$  as Encouragement score and  $Transfer\_score$  as Transfer score. The relation among them is shown in formula (3).

$$Collaboration\_score = Rank\_score + Encouragement\_score + Transfer\_score. \quad (3)$$

#### 3.2.1 Rank Score

In the DLDE Learning 2.0 community, there are a lot of challenging problems, for example, coursework, project etc, which need to be solved through teamwork. We assume different problems have different score according to the complexity of the problem.

The team score that each team member achieves is determined by rating each other. Each member gives a score to others besides himself/herself in the team (In our community, the score lies in scope[0-100]). We use the score to multiply his/her weight as the computing score. So the weight will have more impact on the score. We add up all scores for each member. See formula (4).

In this formula, Team\_score(i) represents the score that member i obtains, Team\_member\_score(j) represents the score that member j gives to member i, User\_weight(j) represents the weight of team member j.

$$\text{Team\_score}(i) = \sum_{j=1}^n \text{Team\_member\_score}(j) * \text{User\_weight}(j) \quad (4)$$

The Rank score that member i gains after solving the problem is in proportion to Team\_score. As formula (5) shows. Rank\_score represents the score that member i get in the end, Problem\_Score represents the problem score.

$$\text{Rank\_score} = (\text{Team\_score}(i) / \sum_{i=1}^n \text{Team\_score}(i)) * \text{Problem\_Score} \quad (5)$$

### 3.2.2 Encouragement Score

In our community, if the resource that a user contributes is very popular, for instance, the uploaded resources, the posted topic, the original article; the user will get an extra encouragement score. The popularity is mainly evaluated through influencing factors by other users, for example, rating results, viewing times, replying times and downloading times. The computation of Encouragement\_score is shown in the formula (6).

$$\text{Encouragement\_score} = \sum_{i=1}^n a(i) * \text{Element}(i) \quad (6)$$

In this formula, Element(i) represents the influencing factor i that we mentioned above, a(i) representing the weight of the influencing factor i.

### 3.2.3 Transfer Score

In our community, participants can find answer for their questions or learn resources they are looking for by offering a rewarding score. If the user finds the best answer and marks it, or when the time expires through rating the best answer, he/she must pay for it. We name this kind of score obtained by rewarding as Transfer\_score. Transfer\_score realizes the score transferred from one user to another.

## 3.3 User Weight

In our DLDE Learning 2.0 Community, every user has a weight named User\_weight. User\_weight is equivalent to the users' reputation. It is mainly determined by users' current score. If you have higher score than others, you will have higher weight. The computation of User\_weight is shown in the formula (7), (8) and (9). A user's weight

is determined by his/her current score and online time, when users' online time meets designated requirement, it will not have effect on users' weight. Therefore the users' score is the only factor that affects the weight. After a user enrolls, his/her weight is assigned an initial value. The initial value of weight is 0.3 in our community by default, which can be changed by system administrator according to needs. (the initial value lies in scope [0-1] ) .

$$\text{User\_weight} = \text{initialization\_weight} + (\text{Score\_weight} * \text{Days\_weight}). \quad (7)$$

In the formula (7), Score\_weight represents the weight of score that users own, Days\_weight represents the weight of online time, Initialization\_weight represents the initial weight.

Score\_weight is determined by users' current score and initial weight. The computation of Score\_weight is shown in the formula (8).

$$\text{Score\_weight} = \tan^{-1} \text{User\_score} * \left(\frac{2}{\pi}\right) * (1 - \text{initialization\_weight}) \quad (8)$$

Days\_weights is computed by the user's online time dividing the number of days that system administrator sets. See formula (9).

$$\text{Days\_weights} = (\text{User\_days}) / (\text{System\_days}). \quad (9)$$

In this formula, Days\_weights represents the intermediate variable to compute Days\_weight, User\_days represents the days that current user is online, System\_days represents the value that system sets.

The value of Days\_weight is computed by following arithmetic :

```

If • Days_weights < 1 •
    Days_weight = Days_weights;
Else
    Days_weight = 1;

```

## 4 Experiment

So far, the DLDE Learning 2.0 Community is used to provide an online learning environment for the students in our university, some courses have been created in the community for years, such as, programming skills training, Data Structure and Arithmetic, Software Engineering etc. Up to now, it has been running more than three years and a lot of digital resources of different formats have been accumulated. How to make full use of the resources, encourage participation and contribution, improve the self-study ability of students became the problem for us to solve urgently, so the reputation model as an incentive and differentiating mechanism was introduced to the community one year ago. Those who work hard and clever enough will get more scores than others.

In the evaluation, we selected 350 users from my course -Data Structure and Arithmetic- in the community, they are the sophomore students who are from School of Computer Science and Technology and school of software. In the course, I prepared

thirty assignments for them, twenty of them are required to be solved individually and others should be conquered by teamwork. They can use all the resources provided in the community and share their knowledge except the answer of assignments. Their performances according to the reputation model in the community will occupy twenty percent of their final examination.

In the process of evaluation, we gathered data for two times, the first time is three months later; the other time is six months later. Fig 3 shows the user distribution three months later. In order to analyze them easily, their scores are normalized to scope [0,100] when displayed. The method of normalization is to select the highest score first among 350 users, we name it as top\_score, and we use norm\_score to represent normalized score. The computation of norm\_score is shown below.

$$\text{Norm\_score} = (\text{User\_score}/\text{top\_score}) * 100. \tag{10}$$

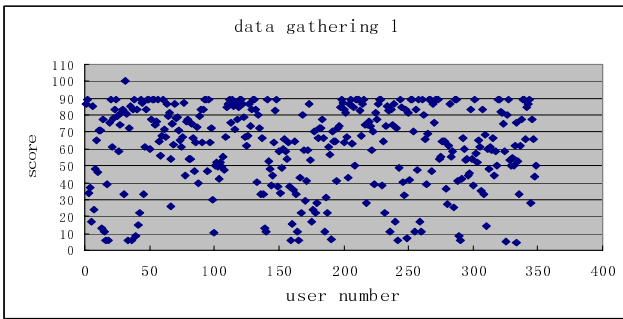


Fig. 3. Data gathering after running three months

From Fig 3 we can see that most data are scattered, only a small part of data is compacted on the top. According to our model, in the beginning stages, if users are active enough, it is very easy for them to get score by attending the activities and challenging the assignments actively. So the encouraging effect is obvious, but the discrimination effect is not too obvious. It reflects that the model begin to take effect.

Fig 4 shows the user distribution six months later. In order to compare it with Fig.3, their score is also normalized to scope [0,100] when displayed.

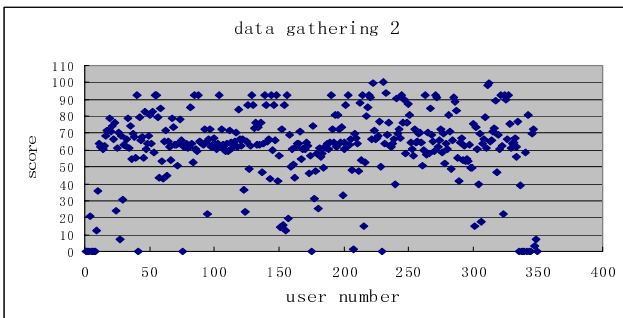


Fig. 4. Data gathering after running six months

From Fig 4 we can see that most data are compact in the middle, only a small proportion of data are scattered on the top and few data reach the bottom. According to the model, when the score of users reach a value, though most of users passionately participate in the individual activities and collaborative activities. It has been more and more difficult for users to get scores. At this time, those clever enough will get higher score, the score of those who are inactive will be reduced to zero. It shows that we have realized a reasonable encouraging and discrimination effect.

## 5 Conclusion and Future work

In this paper, we present a user reputation model for our DLDE Learning 2.0 Community. The main advantage of our model is that it combines user's individual activity analysis approach and collaborative activity analysis approach. The individual activity analysis approach focuses on encouraging participation and contribution and the collaborative activity analysis approach puts emphasis on collaboration and original contribution. As we can see from our experiments, using the combination of the two approaches leads to a reasonable result. It accomplishes our encouraging and discriminating purpose, eventually promotes our students' learning.

However, dealing with cheating activities is always a hot and difficult topic in trust and reputation system. In the DLDE Learning 2.0 Community, we set up an arguing center to handle fraud by people. When the numbers of users become more and more, it is too difficult to handle cheating activities by people. So research on detecting fraud automatically is our next step. On another hand, the weights for the different activities are given arbitrarily according to empirical value, how to give the most appropriate values of weight is our next researching work.

## Acknowledgments

The work described in this paper was fully supported by IBM Eclipse Innovation Awards 2005 and National Natural Science Foundation of China Project No. 60773053.

## References

1. Neel, S.: Online Trust and Reputation Systems. In: Proceedings of the 8th ACM conference on Electronic commerce, pp. 366–367. ACM Press, New York (2007)
2. Wei, C., Qingtian, Z., Wenyin, L.: A User Reputation Model for a User-Interactive Question Answering System. In: Proceedings of the Second International Conference on Semantics, Knowledge, and Grid, pp. 40–45. IEEE Computer Society, Washington (2006)
3. Xiaoming, X., Zhendong, N., Rongjian, L.: An Agent Based Analysis of Study Information in E – learning. *J. New Technology of Library and Information Service* 05, 33–36 (2005)
4. Feng, H., Zhendong, N., Dou, M.: The Design of a Knowledge - based Web Testing Environment. *J. New Technology of Library and Information Service* 08, 65–68 (2005)

5. Lizhe, S., Zhendong, N., Hantao, S., Zhengtao, Y., Xuelin, S.: Study on the User Profile of Personalized Service in Digital Library. *J. Journal of Beijing Institute of Technology* 01, 58–62 (2005)
6. Zaiqing, N., Yuanzhi, Z., Ji-Rong, W., Wei-Ying, M.: Object-level Ranking: Bringing Order to web Objects. In: Proceedings of the 14th international World Wide Web Conference, pp. 567–574 (2005)
7. Resnick, P., Zeckhauser, R., Friedman, R., Kuwabara, K.: Reputation Systems. *J. Communications of the ACM* 12, 45–48 (2000)
8. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The EigenTrust Algorithm for Reputation Management in P2P Networks. In: Proceedings of the Twelfth International World Wide Web Conference, Budapest (2003)
9. Resnick, P., Zeckhauser, R., Swanson, J., Lockwood, K.: The Value of Reputation on eBay. *J. Experimental Economics* 02, 79–101 (2006)
10. Google Answers, <http://answers.google.com/>
11. Jøsang, A., Ismail, R., Boyd, C.: A Survey of Trust and Reputation System. *J. Online Service Provision, Decision Support System* 02, 619–644 (2005)
12. Yahoo! Answers, <http://answers.yahoo.com/>
13. Alvarez, A.R., Stephen, H.: Supporting Trust in Virtual Communities. In: Proceedings of the 33rd Hawaii International Conference on System Sciences, IEEE Computer Society, Los Alamitos (2000)

# Query Relaxation Based on Users' Unconfidences on Query Terms and Web Knowledge Extraction

Yasufumi Kaneko, Satoshi Nakamura, Hiroaki Ohshima, and Katsumi Tanaka

Department of Social Informatics, Graduate School of Informatics, Kyoto University  
Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan  
{kaneko, nakamura, ohshima, tanaka}@dl.kuis.kyoto-u.ac.jp

**Abstract.** This paper proposes a method to allow users to search for Web pages according to their search intentions. We introduce a degree of “unconfidence” for each term in a Web search query. We first investigate the relationships among query terms by accessing a Web search engine. Next, according to a degree of users' unconfidences for each query term and the relationships among query terms, our system finds alternative terms by accessing to a Web search engine. Then, our system generates a collection of new queries that are different from the original query and merges the Web search results obtained for each new query. We implemented our system and showed the usefulness of our system based on user evaluation.

**Keywords:** Web search, Intent Detection.

## 1 Introduction

Recently, Web search engines have become the most common tools for finding information. However, it is not easy for users to retrieve Web pages that contain information they really require. The reasons for this difficulty are as follows:

- **The difficulty of creating an appropriate query**

Because of the lack of enough knowledge in the domain of users' interests, it is usually difficult for users to formulate an appropriate query for obtaining relevant Web pages. For example, if a user wish to have a Japanese food that is typical in Kyoto, he may not know those typical Japanese food names, or he may know partially about those typical Kyoto Japanese food names.

- **The difficulty for users to express clearly their search intentions**

Users cannot inform search engines of their search intentions, especially, to what extent he is confident on a query term. Suppose that a user wishes to have famous Japanese foods in Kyoto and that he partially knows such a name such as “Tofu”. In this case, he may not stick to “Tofu” so much and want to have other famous Japanese food in Kyoto such as “Yuba”. However, there is no way for him to specify how much he sticks to “Tofu”.



### – The difficulty for search engines to detect relationships

When a user wishes to have information on Kyoto’s Japanese food, it is not so easy for search engines to understand the query intent if the user’s query is a conjunction of “Kyoto”, “Japanese“ and “food”.

If users could inform the systems of their search intentions, especially users’ unconfidences on query terms, and the systems can utilize knowledge on semantics relationships among terms from Web, then the systems could return the desired results, and we believe that users could reach target Web pages more easily.

In this paper, we propose the notion of “unconfidence” of each term in a query in regard to users’ search intentions. To indicate the unconfidences, we introduce “**relaxation value**” for each query term. When there is a term with a high relaxation value, we replace it by the disjunction of it’s alternative terms to get more appropriate search results. Based on the assumption that users can input the relaxation value to each term in a query, we propose how to use the value to make search results more appropriate and evaluate our method. As an example method to input relaxation values on query terms, we propose a method to provide users’ intentions to the search engines by using the speed at which each query term is input.

In our method, our system first investigates relationships among query terms by accessing to a Web search engine. It is necessary to find relationships among query terms since finding alternative terms for a given term depends on relationships the term has. For example, “Yuba” is an alternative term of the term “Tofu” if “Tofu” is much related to “Kyoto food”. Second, our system obtains alternative terms for each query term according to the relaxation value and the relationships obtained. Thirdly, our system creates several queries based on the terms in the original query and the alternative terms obtained, and merges the Web search results retrieved by the queries. Finally, our system orders the search results and shows them to users.

The major contribution of the present paper is as follows:

- To introduce the notion of the “unconfidence” on each query term.
- To propose a way to find alternative terms for a query term with high “unconfidence” by extracting knowledge from the Web.
- To propose a way to automatically find the relationships among query terms.
- To propose a way to input users’ unconfidences on query terms.

## 2 Idea of a Query with “Relaxation Value”

Let us consider the following case. A user inputs “*tanaka web-search icadl*” as a query. Is it possible to recognize his search intentions clearly? The search engine will only return simple search results that contain the terms “tanaka”, “web-search”, and “icadl” in spite of the his search intentions. However, there is no guarantee that these results will satisfy him. He may think “I want to find Tanaka’s papers in the proceedings of ICADL. Firstly, I want to find papers which are related to the Web field. However, I want to find papers which are related not only to the Web field but also to other fields such as database

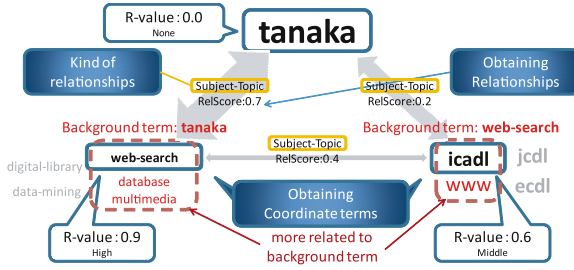


Fig. 1. Diagram of section 3

and multimedia which are related to Tanaka.” When he has such intentions, he will be unconfident for the term “web-search”. On the other hand, the desired results are different depending on which relationships “tanaka and web-search” or “web-search and icadl” he focuses on. If former relationship he focuses on, he is interested in the fields such as “database” or “multimedia” which are related to Tanaka. However, if latter relationship he focuses on, he is interested in the fields such as “digital-library” or “data-mining” which are related to ICADL. In this case, it is difficult for the search engine to return search results that he wants. He might well be dissatisfied if the search results are not desirable.

For this reason, a system in which users can input only query terms in retrieving Web pages is very inconvenient not only for users but also for the system. In order to make the Web search environment more comfortable, it is important to enable users to provide more search intentions to the system.

In this paper, we focus on the unconfidences of search intention because we believe the idea “Which query term do users are unconfident in?” is especially important. Terms which users have low confidence does not contain terms about which users feel uncertain based on spelling error or fluctuation of notation. By considering the unconfidences, we believe that the system can select more appropriate relationships of query terms and optimize search results. We introduce “relaxation value” to indicate users’ unconfidences in each term of a query. A “relaxation value” is continuously-valued from 0 to 1 and is assigned to each term. Here, we call it “R-value”. Given a high R-value, the term is replaced with “related terms” and is set to a low weight which is used in ranking search results. “Related terms” are the terms which our system obtains according to R-value and are related to the original term. In the following section, we propose a method of how to detect users’ intentions on the assumption that users can set R-value for each query term.

### 3 How to Recognize Users’ Intentions

We detect users’ intentions by detecting related terms users are interested in according to R-value and relationships of query terms. In this section, we explain two relationships we consider, and the method of obtaining the related terms

according to R-value and the relationships. A diagram of this section is shown in Figure 1 with an example using a query “tanaka web-search icadl”.

**Relationships of query terms.** We explain the two kinds of relationships we consider. These relationships involve two terms. If we want to obtain relationships among more than two terms, we treat the terms as pairs.

- **Subject-Topic:** When two terms  $A$  and  $B$  such as “japan sushi” and “ipod price” can be shown to be “ $A$  of  $B$ ” or “ $B$  of  $A$ ”, we define this as a “Subject-Topic” relationship. As in the case of “ $B$  of  $A$ ”, subject  $A$  and topic  $B$  tend to exist. Therefore, our system can obtain more appropriate related terms that have the same attribute  $B$  from  $A$  by considering  $B$  and vice versa.
- **Coordinate:** When two terms  $A$  and  $B$  such as “wii xbox360” and “quality quantity” can be shown to be “ $A$  and  $B$ ” or “ $A$  or  $B$ ”, we define this as a “Coordinate” relationship. With this relationship,  $A$  and  $B$  tend to have the same hypernyms or hyponyms. These terms often have co-occurring words and the terms tend to be related to the users’ intentions. The system can obtain appropriate related terms for such a term by considering those terms.

**Obtaining the relationships among query terms.** To obtain the relationship among two query terms, we use the following expression for two terms  $T_1$ ,  $T_2$  and conjunction  $W_c$ .

$$RelScore(T_1, W_c, T_2) = \frac{DF(“T_1 W_c T_2”)}{DF(“T_1 W_c”) * \frac{DF(“T_1 W_c T_2”)}{DF(“W_c T_2”)}} \quad (1)$$

The expression  $RelScore(T_1, W_c, T_2)$  indicates how reasonable the phrase “ $T_1 W_c T_2$ ” is. In the expression,  $DF(“X”)$  means the number of pages found by a phrase search using the phrase  $X$ , and  $T_1 W_c T_2$ ,  $T_1 W_c$ , and  $W_c T_2$  mean phrases that are combined in that order. For example, when we check whether “tanaka” and “web-search” can be reasonably combined using the conjunction “of”, the first item on the right side of the expression means the proportion of the number of pages found using the phrase search “tanaka of web-search” to that found using the phrase search “tanaka of”, and the second item means the proportion of the phrase search “of web-search”. Therefore, the expression gives a score indicating how common the phrase “tanaka of web-search” is.

For two terms  $A$  and  $B$ , our system gets six scores by using three conjunctions “of”, “and”, and “or” in two different orders “ $A B$ ” and “ $B A$ ”. The system regards the relationship that receives the highest of the six scores as the relationship among the terms. We set a threshold  $L_R$  and consider no relationship if there is none of the score above the threshold. By doing this for each pair of query terms, our system get related query term for each query term. If a query term has several related query terms, our system uses only the related query term with the highest score. For example in Figure 1, “tanaka” is more related to “web-search” than “icadl” because of the difference in those scores.

**Obtaining related terms for each query term.** To obtain related terms for each query term, we use “coordinate terms” that have the same hypernyms.

“Windows” and “Mac” are examples of such terms. Regarding methods of obtaining the terms, Ohshima et al. proposed using a web search [1], and many other methods have been proposed such as using HTML structure [2], a data set [3]. We use the former method because terms can be obtained relatively quickly without large corpora. By using the method, our system can obtain terms in the order of the score which each obtained term has. In this method, we can use a “background term” to get more appropriate terms than without the term. By using background terms, our system gets terms that are related not only to the original term but also to the background term above terms that are related only to the original term. In our method, for each query term, our system gets a background term according to the R-value and related query term, and obtains related terms that reflect the background term. First, our system gets background terms for each query term. For convenience, we set thresholds that assign one of four grades to the level of the R-value: “none”, “low”, “middle”, and “high”. If  $B$  is related to  $A$  and  $B$ 's R-value level is “low” or “none”, our system adds  $B$  to  $A$ 's background term. For example in Figure 1, “web-search”'s background term is “tanaka” according to the relationship and R-value. Then, our system sets the number of terms that it should get according to the R-value for each query term, and obtains coordinate terms according to this number and the background terms. For example in Figure 1, our system obtains two terms for “web-search” and one term for “icadl” based on their R-value.

## 4 Creating Queries and Optimizing Search Result

**Creating new queries and obtaining search results.** Our system creates new queries by replacing each original query term with the related terms. When there are  $n$  terms in the original query  $q$ , we describe the group as  $\{t_1, \dots, t_j, \dots, t_n\}$  and define  $RT(t_j)$  as the related terms for  $t_j$ .  $Cand(t_j)$  means the group of candidates for terms that our system replaces  $t_j$  with.  $Cand(t_j)$  contains  $t_j$  and  $RT(t_j)$ . In addition, if  $t_j$ 's R-value is high,  $Cand(t_j)$  contains “” (blank) to create queries that do not contain the term. By selecting a term in  $Cand(t_j)$  for each  $t_j$  in order, our system creates new queries using all combinations. We describe the group of new  $m$  queries as  $Q' = \{q'_1, \dots, q'_i, \dots, q'_m\}$ .

Then, our system obtains a new result by merging the results of each  $q'_i$ .  $P_{total}$  and  $Count(Cand(t_j))$  means the number of pages that users want to retrieve and the number of elements in  $Cand(t_j)$ . We define  $PageCount(q'_i)$  which means the number of pages that our system retrieves by using  $q'_i$  as in the following expression.

$$PageCount(q'_i) = P_{total} * \prod_{j=1}^n \frac{1}{Count(Cand(t_j))} \quad (2)$$

Our system merges the results of each  $q'_i$  depending on the  $PageCount(q'_i)$ . If some pages of a query have already been retrieved by other queries, our system skips these pages and retrieves lower ranked pages.

**Ordering pages in the search result.** First, our system sets weights for each query term based on the R-value. We define  $w(t_j)$ , the weight of  $t_j$ , using the following based on  $RV(t_j)$ , the R-value of  $t_j$ .

$$w(t_j) = 1 - RV(t_j) \quad (3)$$

Secondly, our system sets scores for each page based on the weights. Each page  $p$  has at least one query  $q'_i$  that our system uses to get the result containing the page. For the queries for each page, we define  $SubScore(p, q'_i)$  as follows.

$$SubScore(p, q'_i) = \frac{S_{query}(p) + S_{summary}(p, q'_i)}{Rank(p, q'_i)} \quad (4)$$

When our system gets page  $p$  by using query  $q'_i$ ,  $S_{query}(p)$ ,  $S_{summary}(p, q'_i)$ , and  $Rank(p, q'_i)$  mean the sum of the weights of the query terms that the query contains, the sum of the weights of the query terms that the summary contains, and the rank of the page. We describe  $q'_i$  which has the highest  $SubScore$  for a page  $p$  as  $q'_{max}$ . Based on this, we define the score of a page  $p$  as  $Score(p)$ , thus

$$Score(p) = \frac{S_{title}(p)}{Rank(p, q_{max})} + SubScore(p, q_{max}) \quad (5)$$

$S_{title}(p)$  means the sum of the weights of the query terms that the title of a page  $p$  contains. Our system uses  $Rank$  in computing  $SubScore$  or  $Score$  to ensure that the merged results contain the pages that have the highest rank for each new query. This correction minimizes the risk that only the pages retrieved by a few new queries out of all the new queries are in the highest ranked pages in the final results. Our system orders pages based on  $Score(p)$ , and returns the result.

## 5 Implementation

To evaluate the usefulness of our method, we implemented a prototype using programming language C#, and we used the Yahoo! Search Engine.

In designing the input interface, we aimed to achieve the following features.

- **No need to switch among keyboard and mouse operation**
- **Use of only one input box to avoid confusing users**
- **Simplified process without special search options**<sup>1</sup>

After considering these factors, we propose the use of the speed of inputting query terms as the method of providing R-value to our system. If users have query terms for which they want to set a high R-value, all they have to do is input the terms slowly and our system will set the R-value depending on the input speed. This method enables users to input query terms into a normal input box using only a keyboard. In addition, they do not have to consider an explicit value for each query term. Our system therefore incorporates the three features described above. Converting the input speed to R-value is based on the ratio of the average input speed for each query term to the average input speed of the

<sup>1</sup> Only 8.72% of search engine users use special search options such as “-”<sup>[4]</sup>.

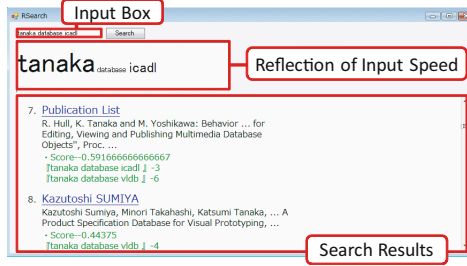


Fig. 2. Prototype of this system

Table 1. Examples of search intentions and queries used in the evaluation

Intention(Something which users want to know)	Query
1 Comparing Casio's digital camera with other brands.	casio(high) digital-camera
2 Difference between Windows and Mac about graphics and so on.	windows mac graphics(high)
3 Party stunts such as tricks.	party-stunt trick(high)

whole query. For this reason, personal differences in users' input speed do not affect our system. In addition, varying the speed of input is practical because all users are capable of inputting words more slowly. In order to check the possibility of this operation, we asked each of six test subjects to adjust the input speed based on three types of speed, "normal", "slower", and "slowest" for each query term of six queries containing three terms. The average R-value provided by the simple experiment are **0.060** for "normal", **0.505** for "slower", and **0.644** for "slowest". We think that there is a probability of this interface.

Our prototype is shown in Figure 2. In our system, users input a query into an input box, and get optimized results. The more slowly users input a query term, the smaller the term appears under the input box. By checking the size, users can see whether they have informed our system of their intentions accurately.

## 6 Evaluation

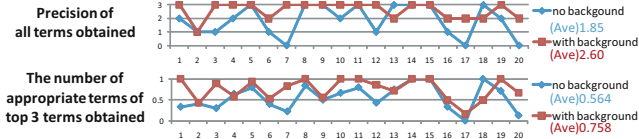
To evaluate our method, we prepared 20 intentions and queries. Some of them are shown in Table 1. In the queries in Table 1, the terms in parentheses behind the query terms mean the level of the R-value.

### 6.1 Evaluation of Related Terms

For each query term which has a high R-value in the 20 intentions, we first compared terms obtained with background term to terms obtained without background term in order to show the importance of considering relationships. We show related terms obtained for each intention in Table 2 as examples. For each intention, we show the precision of all terms obtained and the number of appropriate terms in top 3 terms obtained in Figure 3. The results show that our

**Table 2.** Related terms obtained by our method

	original term	background term	related terms obtained
1	casio	digital-camera	Sony, Nikon, Ricoh
2	graphic	windows	game, music, soft
3	stunt	party-stunt	comic-show, dance, short-performance

**Fig. 3.** Result for evaluation of related terms

method obviously obtained more appropriate terms by using background term based on relationships. Although there are some low points in precision of all terms, our system can obtain more than 2 terms in top 3 terms for most of the intentions. According to the results, we decided to use only top 3 terms when we need related terms.

## 6.2 Evaluation of Search Results

**Evaluation method.** In this evaluation, we used three search results.  $R_{pm}$  are the results for our proposed method.  $R_{oq}$  are the results for the original queries.  $R_{cq}$  are the results for the complex queries containing all related terms retrieved by our method and each related term connected by the Yahoo! Search Engine’s search option “OR”. For example, when the original query was “*tanaka web-search icadl*” and our system obtained “database” and “multimedia” as related terms for “web-search” and “www” as a related term for “icadl”,  $R_{oq}$  was the result obtained by the query “*tanaka web-search icadl*” and  $R_{cq}$  was the result obtained by the query “*tanaka web-search OR database OR multimedia icadl OR www*”. For each intention of the 20 intentions, we asked the users to rank  $R_{pm}$ ,  $R_{oq}$ , and  $R_{cq}$  according to the two following views.

- **View 1:** Which result covers the most topics reflecting the users’ intention?
- **View 2:** Which result contains the most pages reflecting the users’ intention?

We carried out the evaluation with seven users in the 20-30 year age group who often use search engines. For each intention, we showed users the intention, the query, the views, and the top 10 pages for the three results that were allocated randomly, and asked them to rank the three results.

**Evaluation results.** Each result is given score 1, 0, or -1 in the order of ascending the ranks that users assigned to each result. For each view, average scores based on the ranks are shown in Figure 4.

First, we evaluated from view 1. A comparison of  $R_{pm}$  and  $R_{oq}$  showed obvious differences. This means that the results of our method can cover more topics that reflected users’ intentions. When we compared  $R_{pm}$  and  $R_{cq}$  to check the

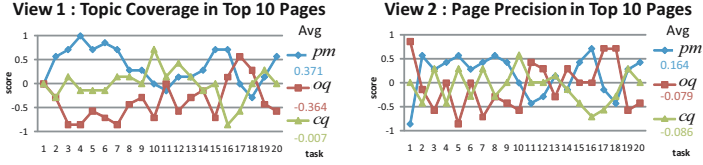


Fig. 4. Results for each view

usefulness of the methods of merging and ordering, the results of the evaluation showed that  $R_{pm}$  contained more topics that reflected users' intentions. We consider that this was because the top 10 pages tended to contain a variety of pages evenly as a result of our methods of merging and scoring.

Next, we evaluated from view 2. The results of the evaluation showed that  $R_{pm}$  had fewer noisy pages than  $R_{oq}$ . A comparison of  $R_{pm}$  and  $R_{cq}$  showed similar results. Therefore, the results obtained by our method contained more pages that reflected users' intentions than were obtained by normal queries.

Finally, we looked at the results for each intention. For some intention such as 1, 11, 12, 14, 17, and 18, the results showed  $R_{oq}$  were better. Checking the reasons, we found that our method's performance decreased when the probability of co-occurrence of the original query term which had a high R-value and related term was very high. Given these results, we need to consider how to incorporate the kind of query terms that tend to co-occur as related terms. In contrast, the results for intentions such as 16 and 20, which contain personal names, ranked highly. Similarly, the results of intentions such as 5, 8, and 15, which contain the other proper nouns, also ranked highly. We can therefore say that our method is effective for intentions that contain proper nouns.

### 6.3 Discussion

The main advantage of our proposed system is that users can make queries using query terms they are unconfident in or that are not highly specific, especially when they cannot remember or do not know more appropriate query terms. Our system is useful in this situation because it obtains appropriate related terms and a variety of topics automatically. In addition, because the pages are ordered, so that they do not appear unevenly, users can check pages covering various topics even if they check only a few top pages. Therefore, our system is capable of displaying a variety of topics for users.

Unfortunately, a considerable amount of time is required for processing. As the number of query terms increases, the number of accesses to the network increases because the number of combinations of query terms and new queries increases. The system therefore needs to be optimized to reduce accesses. Another problem is that the types of queries on which our method can operate effectively are limited. For example, when users want to add a query with the name of a member of the "Beatles" as a query term, and they add "Paul McCartney" for instance, which has a high R-value, they will also get information about other members of



the band. However, if they add the query term “Beatles” to a query and give the query term a high R-value, they will not get information on the other members. Although we do not use hypernyms or hyponyms because getting those terms consumes too much time, we need to consider including methods that obtain those terms to improve our system.

## 7 Related Work

There have been several studies to recognize the relationships among terms in a given query. Noda et al. proposed a method of discriminating between subject terms and topic terms in a query with two terms by using Web search[5]. Ohshima et al. proposed a method to obtain terms which are related to terms in a given query by considering which conjunctions were most often used around the terms of the query in the snippets[1]. We obtain the relationships among the terms in a query and derive related terms by considering these methods.

Some studies have tried to detect query intent by classifying the goals of the Web search[6][7]. There are also several systems that enable users to establish their search intentions. Regarding the researches for intent detection, relevance feedbacks based on documents or terms are also well known[8][9]. However, they need repeatable retrieves to reflect users’ intentions to the search results.

We can also point to some studies about query modification[10][11]. Our method is similar to these in the concept of replacing query terms with more appropriate terms. However, our method differs from them in that our method uses a lexico-syntactic pattern based on the method of Ohshima et al[1] to get related terms which are appropriate replacements for the original terms.

## 8 Conclusion

We proposed a method of optimizing search results based on users’ intentions and demonstrated its usefulness in an experimental evaluation. In addition, we described the development of an interface that enables users to inform a search system of the intention of their search.

In future work, we will investigate how to detect users’ intentions more accurately from the relationships among query terms. In this paper, we did not discriminate between two relationships during processing, but we intend to do this in future. In addition, we will include the use of hypernyms or hyponyms.

## Acknowledgements

This work was supported in part by “Informatics Education and Research Center for Knowledge-Circulating Society” (MEXT Global COE Program, Kyoto University), and by MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: “Content Fusion and Seamless Search for Information Explosion” (Grant#: 18049041).

## References

1. Ohshima, H., Oyama, S., Tanaka, K.: Searching Coordinate Terms with Their Context from the Web. In: Proc. of WISE 2007, pp. 40–47 (2007)
2. Shinzato, K., Torisawa, K.: A Simple WWW-based Method for Semantic Word Class Acquisition. In: Proc. of the Recent Advances in Natural Language Processing, pp. 493–500 (2005)
3. Ghahramani, Z., Heller, K.: Bayesian sets. *Advances in Neural Information Processing Systems* 18 (2005)
4. White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: Proc. of SIGIR 2007, pp. 255–262 (2007)
5. Noda, T., Ohshima, H., Tezuka, T., Tanaka, K.: Web information retrieval by extraction of topic terms from subject terms. *DBSJ Letters* 5(2), 65–68 (2006)
6. Kang, I.-H., Kim, G.: Query type classification for web document retrieval. In: Proc. of SIGIR 2003, pp. 64–71 (2003)
7. Lee, U., Liu, Z., Cho, J.: Automatic identification of user goals in web search. In: Proc. of the 14th international conference on World Wide Web, pp. 391–400 (2005)
8. Rocchio, J.: Relevance feedback in information retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, 313–323 (1971)
9. Tan, B., Velivelli, A., Fang, H., Zhai, C.: Term Feedback for Information Retrieval with Language Models. In: Proc. of SIGIR 2007, pp. 263–270 (2007)
10. Jones, R., Rey, B., Madani, O., Greiner, W.: Generating query substitutions. In: Proc. of the 15th international conference on World Wide Web, pp. 387–396 (2006)
11. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: Proc. of the 13th international conference on World Wide Web, pp. 666–674 (2004)

# A Query Language and Its Processing for Time-Series Document Clusters

Sophoin Khy<sup>1</sup>, Yoshiharu Ishikawa<sup>2</sup>, and Hiroyuki Kitagawa<sup>1,3</sup>

<sup>1</sup> Graduate School of Systems and Information Engineering, University of Tsukuba

<sup>2</sup> Information Technology Center, Nagoya University

<sup>3</sup> Center for Computation Sciences, University of Tsukuba

sophoin@kde.cs.tsukuba.ac.jp, ishikawa@itc.nagoya-u.ac.jp

kitagawa@cs.tsukuba.ac.jp

**Abstract.** Document clustering methods for time-series documents produce a sequence of snapshots of clustering results over time. Analyzing the contents (topics) and trends in a long sequence of clustering snapshots is hard and requires efforts since there are too many number of clusters; a user may need to access every cluster or read every document contained in each cluster. In this paper, we propose a framework to find clusters of user interest and change patterns called *transition patterns* involving the clusters. A cluster in a clustering result may persist in another cluster, branch into more than one cluster, merge with other clusters to form one cluster, or disappear in the adjacent clustering result. This research aims at providing users facilities to retrieve specific transition patterns in the clustering results. For this purpose, we propose a query language for time-series document clustering results and an approach to query processing. The first experimental results on TDT2 corpus clustering results are presented.

**Keywords:** cluster graph, cluster transition, clustering result, graph query, query language, query processing, transition pattern.

## 1 Introduction

The Internet has become a common tool for information dissemination, where on-line information, such as news and blogs, is being profusely published everyday. Organizing or extracting salient information from such profuse amount of information is thus a challenging task. Document clustering is a method used to find and group similar documents into the same clusters, while dissimilar ones into different clusters. It has been used as a fundamental and pre-processing method in many areas, such as information retrieval [3], topic detection and tracking [2], text classification [9] and summarization of documents [8].

Document clustering methods for time-series documents such as the one in reference [5] produce a sequence of snapshots of clustering results over time (Fig. 1). Analyzing the contents (topics) and trends in a long sequence of clustering snapshots is hard and requires efforts since there are too many number of clusters; a user may need to access every cluster or read every document contained in each cluster. Some recent works have embraced tracing and analyzing changes in clusters over time [6,7,10]. Some further developed visualizing tools for browsing clusters and exploring trends in time-series clustering results [14].

In these works, without browsing, it is still difficult to obtain specific information of user interest.

Consider, for example, a sequence of clustering results as shown in Fig. 1. A user is interested in the US Democrat presidential nomination campaign. The topic can be represented, for example, by the keywords {clinton, obama, nomination}. The user wants to find occurrences of clusters, as shown in Fig. 1, highly related to the keywords. To the best of our knowledge, none of existing methods support such user requirements. By manually exploring the data in the system, it is not easy for the user to judge whether those likely clusters respond well to his/her interest. In addition, tracking topic evolutions in a large stream of clusters is very important in the analysis of the characteristics of the data or the real world events. For example, does the topic {Clinton nomination campaign} dissolve and change to {Obama nomination campaign}? Or does {Clinton nomination campaign} merge with {Obama nomination campaign} and form {Democrat US presidential campaign}?

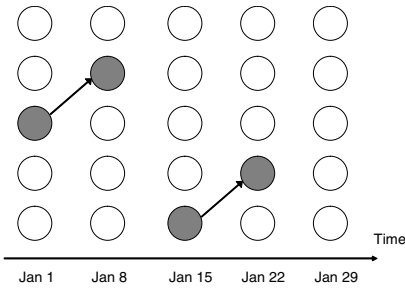


Fig. 1. A Sequence of Clustering Results

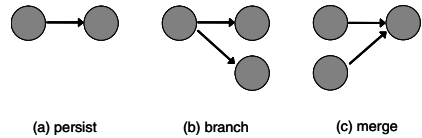


Fig. 2. Transition Patterns

In a sequence of clustering results of time-series documents, a cluster in a clustering result may persist in another cluster, branch into more than one cluster, merge with other clusters to form one cluster, or disappear in the adjacent clustering result. These phenomena is called *cluster transitions* in reference [10] and we call the occurrence patterns of one or more such phenomena together, as shown in Fig. 2, *transition patterns*. In this paper, we propose a framework to find clusters related to users’ topics of interest and transition patterns involving the clusters. Specifically, we aim at providing users facilities to retrieve specific transition patterns in the clustering results. For this purpose, we propose a query language for time-series document clustering results and an approach to query processing. Given a query and transition pattern, it finds the occurrences of cluster transitions that match the given pattern and are relevant to the query. The results are ranked and returned to the users.

The remainder of this paper is organized as follows. Section 2 reviews related work. The preparations in this research are described in Section 3. Section 4 introduces the query language and explains the query processing scheme. The evaluation of the query language is shown in Section 5. Section 6 concludes the paper and discusses future work.

## 2 Related Work

There are several researches on identifying relationship between clusters. Mei et al. proposed an approach to discovering the evolutionary patterns of themes in a text stream [6]. In the approach, themes are generated using a probabilistic mixture model and theme evolutionary relations are discovered. Nallapati et al., based on the notion of events and topics in TDT [2], identified events that make up a topic and established dependencies among them [7]. Several dependency models have been proposed in the approach. Although the underlying problem in finding transitions between clusters is relevant to our work, the two approaches do not restrict to finding transitions between two consecutive time points, but find all possible transitions between clusters from all time instances. These approaches aim at finding all relevant topics. Hence, their underlying objectives differ from our approach.

Spiliopoulou et al. proposed an approach called MONIC to model and track cluster transitions on clustering results at consecutive time points [10]. A cluster transition at a given time point is a change experienced by a cluster that has been discovered at an earlier time point. A transition may concern the content and form of the cluster, which is internal to it, or it may concern its relationship to the rest of the clustering, which is an external transition. MONIC detects both internal and external transitions of clusters and models their lifetime. The cluster transition graph (Section 3.2) in our approach is based on the idea of external transitions in MONIC.

T-Scroll [4] is an information visualization interface tool to visualize the overall trend of time-series documents. It displays links between clusters, which represents related clusters from two consecutive time points, and allows users to explore more detailed information such as keyword lists and contents of documents in a cluster. Adomavicius et al. in reference [1] proposed a data analysis and visualization technique for representing trends in multiattribute temporal data called C-TREND, a system for temporal cluster construct. The idea in T-Scroll and C-TREND is partly related to our work. However, in our work, we proposed a query language for temporal clusters aiming at providing users facilities to search for specific patterns in the clustering results.

## 3 Preparations

### 3.1 Target Data

The target data in this work is a collection of accumulated document clustering results  $D_1, \dots, D_n$  at consecutive time points  $T_1, \dots, T_n$  generated by a clustering method on time-series documents where a sliding window is adopted.

Let  $S_1, \dots, S_n$  be document sets in which  $S_i$  is the document set from which the document clustering result  $D_i$  is generated.  $S_i$  is assumed to be overlapped with  $S_{i+1}$ , i.e.,  $S_i \cap S_{i+1} \neq \emptyset$ . All clusters in each  $D_i$  ( $1 \leq i \leq n$ ) are assumed non-overlapped, i.e.,  $\forall C_p, C_q \in D_i, C_p \cap C_q = \emptyset$ , if  $p \neq q$ .

### 3.2 Constructing Cluster Transition Graph

A cluster transition, proposed in MONIC [10], at a given time point is a change experienced by a cluster that has been discovered at an earlier time point. It provides insights about the nature of cluster changes: is a cluster a newly emerged

```

Query ::= Find-Clause From-Clause With-Clause
Find-Clause ::= 'find' Transition_Pattern
From-Clause ::= 'from' G
With-Clause ::= 'with' Bindings
Transition_Pattern ::= Snapshot_Spec (-> Snapshot_Spec)*
Snapshot_Spec ::= Node | '{' Node_List '}'
Node_List ::= Node (, Node)*
G ::= Seq_Name('['TimeStamp, TimeStamp']')?
Bindings ::= Node '=' Keyword_List (, 'and' Node '=' Keyword_List)*
Keyword_List ::= '{'Keyword (, Keyword)* ((, 'not('Keyword')')*)?}'

```

**Fig. 3.** BNF of Query Language

cluster or a disappeared one or does some of its member move to different clusters. In this work, we construct the cluster transition graph based on the idea of cluster transitions in MONIC.

A cluster transition graph is constructed by connecting pairs of adjacent clustering results. A pair of two adjacent clustering results  $D_i$  at time point  $T_i$  and  $D_j$  at the successive time point  $T_j$  is connected by detecting transitions between all clusters in  $D_i$  and  $D_j$  as follows.

For each cluster  $C_i$  in  $D_i$  and  $C_j$  in  $D_j$ , the degree to which  $C_i$  overlaps  $C_j$  is measured. The following function, the transition probability of the number of documents that were in  $C_i$  and moved to  $C_j$ , is used in this work.

$$\Pr(C_j|C_i) \stackrel{\text{def}}{=} \frac{|C_i \cap C_j|}{|C_i|}, \quad (1)$$

where  $|C_i|$  and  $|C_j|$  are the number of documents in  $C_i$  and  $C_j$ , respectively.

A list of clusters in  $D_j$ , which are matched clusters of clusters in  $D_i$ , and the corresponding transition probability values are obtained. Then, links between clusters and, hence, graphs can be constructed.

## 4 The Query Language

### 4.1 The Query Language Syntax

The query language proposed in this paper is a quite simple declarative language of small number of constructs, but can represent information needed for analyzing document clustering snapshots and detect cluster transition patterns. Figure 3 shows an extended BNF notation of the query language.

### 4.2 Examples of Queries

The following examples show the query language syntax.

Query 1

```

find C -> C
from S08-7days[2008-Jan-01, 2008-Feb-28]
with C = {clinton, obama, nomination}

```

We assume that `S08-7days` is the name of a sequence of document clusters. Suppose that it contains a sequence of document clusters from January 1, 2008 to present, and the clustering was performed on seven-day basis. The notation `S08-7days[2008-Jan-01, 2008-Feb-28]` represents that we restrict the scope of the target to the period between January 1, 2008 and February 28, 2008 within `S08-7days`.

In the above query example, the *transition pattern* `C -> C` in the `find` clause specifies a transition pattern to be found. The `with` clause `with C = {clinton, obama, nomination}` represents the associate keywords for `C`; it means that we want to find the occurrences of a transition pattern in which both clusters can be represented by keywords `{clinton, obama, nomination}`.

The results of the query is given, for example, as follows:

```
1: C#Jan15-5 -> C#Jan22-4
2: C#Jan1-3 -> C#Jan8-2
...
```

It is a ranked list of cluster transitions. The user may be able to specify the preferred number of entries. The notation such as `C#Jan15-5` denotes the identity of a cluster. In this case, it represents the fifth cluster obtained at January 15.

The next example shows a query to retrieve the occurrences of a pattern with length two.

#### Query 2

```
find C -> C -> C
from S08-7days[2008-Jan-01, 2008-Feb-28]
with C = {clinton, obama, nomination}
```

The following example shows a query using different keyword sets for the two clusters.

#### Query 3

```
find C1 -> C2
from S08-7days[2008-Jan-01, 2008-Feb-28]
with C1 = {clinton, nomination, not(obama)}, and
      C2 = {obama, nomination, not(clinton)}
```

This query retrieves cluster transitions in which the first cluster corresponds to `{clinton, nomination}` but does not include `{obama}`, and the second one corresponds to `{obama, nomination}` but does not include `{clinton}`.

The following query contains a transition pattern with a branch.

#### Query 4

```
find C1 -> {C2, C3}
from S08-7days[2008-Jan-01, 2008-Feb-28]
with C1 = {clinton, obama, nomination}, and
      C2 = {clinton, not(obama)}, and
      C3 = {obama, not(clinton)}
```

The query wants to find the occurrences of two transitions  $C_1 \rightarrow C_2$  and  $C_1 \rightarrow C_3$ , where  $C_2$  and  $C_3$  branch from  $C_1$  and are highly related to the given keywords.

The next query is the opposite; it is about the merge pattern.

Query 5

```
find {C1, C2} -> C3
from S08-7days[2008-Jan-01, 2008-Feb-28]
with C1 = {clinton, not(obama)}, and
      C2 = {obama, not(clinton)}, and
      C3 = {clinton, obama, nomination}
```

### 4.3 Examples of Query Processing

In this section, the processing scheme of the query language regarding the query examples in Section 4.2 is given in the same order as follows.

**Query 1: The Most Simple Case.** We consider to answer Query 1. Let the set of keywords appearing in the `with` clause be  $Q = \{t_1, \dots, t_n\}$ . In the example,  $Q$  is {clinton, obama, nomination}.

The `find` clause specifies that the two clusters, which have a transition, correspond to the same query  $Q$ , but are from two different clustering results at consecutive time points. To avoid confusion, the two clusters are denoted as  $C_1$  and  $C_2$ , where  $C_1$  temporally precedes  $C_2$ . We define the score for the cluster transition  $C_1 \rightarrow C_2$  in terms of query  $Q$  by

$$s(C_1(Q) \rightarrow C_2(Q)) \stackrel{\text{def}}{=} \Pr(C_1|Q) \cdot \Pr(C_2|Q) \cdot \Pr(C_1 \rightarrow C_2). \quad (2)$$

Intuitively, the score gives a probability that the occurrence of a cluster transition is related to the given query  $Q$ . The score is calculated and used to rank the query results.

Next, we consider how to derive the probabilities  $\Pr(C_1 \rightarrow C_2)$  and  $\Pr(C|Q)$ .  $\Pr(C_1 \rightarrow C_2)$  can be defined as

$$\Pr(C_1 \rightarrow C_2) \stackrel{\text{def}}{=} \Pr(C_2|C_1), \quad (3)$$

where  $\Pr(C_2|C_1)$  is defined in Eq. 1 in Section 3.2.

$\Pr(C|Q)$  is a conditional probability that given the query  $Q$ , we obtain  $C$  as a relevant cluster to  $Q$ . It is defined as

$$\Pr(C|Q) \stackrel{\text{def}}{=} \prod_{i=1}^n \Pr(t_i \in C), \quad (4)$$

where  $\Pr(t_i \in C)$  is the occurrence probability of a query term  $t_i$  of  $Q$  in cluster  $C$ .

In summary, we get

$$\begin{aligned} s(C_1(Q) \rightarrow C_2(Q)) &\propto \left[ \prod_{i=1}^n \Pr(t_i \in C_1) \right] \cdot \left[ \prod_{i=1}^n \Pr(t_i \in C_2) \right] \cdot \frac{|C_1 \cap C_2|}{|C_1|} \\ &= \frac{|C_1 \cap C_2|}{|C_1|} \cdot \prod_{i=1}^n [\Pr(t_i \in C_1) \cdot \Pr(t_i \in C_2)]. \end{aligned} \quad (5)$$



**Query 2: Long Sequence.** Next, we consider Query 2. In this case, we assume that the two transitions are nearly *independent* and approximate the probability.

$$\begin{aligned} s(C_1(Q) \rightarrow C_2(Q) \rightarrow C_3(Q)) &\stackrel{\text{def}}{=} s(C_1(Q) \rightarrow C_2(Q)) \times s(C_2(Q) \rightarrow C_3(Q)) \\ &\stackrel{\text{def}}{=} \Pr(C_1|Q) \cdot [\Pr(C_2|Q)]^2 \cdot \Pr(C_3|Q) \\ &\quad \cdot \Pr(C_1 \rightarrow C_2) \cdot \Pr(C_2 \rightarrow C_3). \end{aligned} \quad (6)$$

The two probabilities can be calculated using the method of Query 1.

**Query 3: Different Keyword Sets.** We consider Query 3. In this case, we need to consider two keyword sets for  $C_1$  and  $C_2$ . Let the corresponding keyword sets for  $C_1$  be  $Q_1 = \{t_1^1, \dots, t_n^1\}$  and  $C_2$  be  $Q_2 = \{t_1^2, \dots, t_m^2\}$ . In the example, their contents are  $\{\text{clinton}, \text{nomination}, \text{not}(\text{obama})\}$  and  $\{\text{obama}, \text{nomination}, \text{not}(\text{clinton})\}$ , respectively.

In this query language, `not()` in the `with clause` is a predicate used to negate the effect of the query keyword in the parenthesis. If a query contains the predicate `is`, for example,  $C = \{\text{clinton}, \text{nomination}, \text{not}(\text{obama})\}$ ,  $\Pr(C|Q)$  is computed as follows.

$$\begin{aligned} \Pr(C|Q) &= \Pr(C = \{\text{clinton} \wedge \text{nomination} \wedge \neg \text{obama}\}) \\ &= \Pr(\text{clinton} \in C) \cdot \Pr(\text{nomination} \in C) \cdot \Pr(\text{obama} \notin C) \\ &= \Pr(\text{clinton} \in C) \cdot \Pr(\text{nomination} \in C) \cdot (1 - \Pr(\text{obama} \in C)). \end{aligned}$$

Finally, similar to Query1, we process this query as follows.

$$s(C_1(Q_1) \rightarrow C_2(Q_2)) \stackrel{\text{def}}{=} \Pr(C_1|Q_1) \cdot \Pr(C_2|Q_2) \cdot \Pr(C_1 \rightarrow C_2). \quad (7)$$

**Query 4: Query with Branch.** We consider Query 4. We assume the two cluster transitions are *independent*.

$$\begin{aligned} &s(C_1(Q_1) \rightarrow \{C_2(Q_2), C_3(Q_3)\}) \\ &\stackrel{\text{def}}{=} s(C_1(Q_1) \rightarrow C_2(Q_2)) \times s(C_1(Q_1) \rightarrow C_3(Q_3)) \end{aligned} \quad (8)$$

The results can be evaluated using the same approach as Query 1.

**Query 5: Query with Merge.** For Query 5, we use the same assumption as Query 4 that the two transitions are *independent*.

$$\begin{aligned} &s(\{C_1(Q_1), C_2(Q_2)\} \rightarrow C_3(Q_3)) \\ &\stackrel{\text{def}}{=} s(C_1(Q_1) \rightarrow C_3(Q_3)) \times s(C_2(Q_2) \rightarrow C_3(Q_3)). \end{aligned} \quad (9)$$

## 5 Implementation

In this section, we implemented the proposed query language described above.

## 5.1 Computing Term Frequency

For the implementation, the  $\Pr(t_i \in C)$ , where  $t_i$  is a term and  $C$  is a cluster, is defined as follows.

$$\Pr(t_i \in C) \stackrel{\text{def}}{=} \frac{\text{total\_freq}(t_i)}{\sum_{j=1}^l \text{total\_freq}(t_j)}, \quad (10)$$

where  $\{t_1, \dots, t_l\}$  is the set of all terms appeared in the documents in  $C$  and  $\text{total\_freq}(t_i)$  is the total frequency of  $t_i$  in  $C$  and is defined as follows.

Suppose  $C$  contains documents  $\{d_1, \dots, d_{|C|}\}$  and let the frequency of term  $t_i$  in document  $d_k$  be  $\text{freq}_k(t_i)$ . The total frequency of  $t_i$  in  $C$  is defined by:

$$\text{total\_freq}(t_i) \stackrel{\text{def}}{=} \sum_{k=1}^{|C|} \text{freq}_k(t_i). \quad (11)$$

This  $\Pr(t_i \in C)$  (Eq. 10) and the transition probability  $\Pr(C_j|C_i)$  (Eq. 1) are computed beforehand and stored persistently so that we can retrieve them when needed in the computation of query scores.

## 5.2 Experimental Evaluation

**Experimental Setup.** In this evaluation, the time-series document clustering result of an extended- $K$ -means-based incremental clustering method on TDT2 Corpus [11] in reference [5] is used as the data set. The clustering results were generated by three-day incremental basis. The TDT2 Corpus consists of chronologically ordered news stories obtained from various sources. We used 20 clustering results dated from February 2 to March 31, 1998. Each clustering result contains 17 clusters and is associated with a timestamp, the date the clustering was performed.

Due to limitations in the data, some transition patterns and query keywords may not give meaningful results. Query keywords should be chosen such that the results are not empty. For this experiment, we chose keywords which represent topics in the TDT2 Corpus for transition patterns as given below. In addition, for query patterns of branch and merge, we use the same query keyword sets for all clusters C1, C2 and C3.

- C → C pattern, C = {pope, cuba}; {monica, clinton}; {tornado, florida}; {iraq, weapon, inspection}
- C1 → {C2, C3} pattern, C1 = C2 = C3 = {monica, clinton}; {iraq, weapon, inspection}
- {C1, C2} → C3 pattern, C1 = C2 = C3 = {monica, clinton}; {iraq, weapon, inspection}

**Results and Discussion.** To evaluate the query results, topic labels for clusters in the clustering data are used to evaluate query results. A topic label of a cluster was obtained by measuring precision of the cluster against the evaluation data of the TDT2 Corpus. If the precision score of the cluster for a topic is larger than a threshold, the cluster is labeled with the topic. Table 1 shows some cluster ID's and topic labels for the clusters of the clustering data when the threshold is set

**Table 1.** Clusters and Labeled Topics

Cluster ID	Topic Name
Feb02-3, Feb02-5, Feb05-3, Feb05-6, Feb08-6, Feb11-7, Feb14-7, Feb17-10, Feb20-10, Feb23-10, Feb26-9, Feb26-10, Mar01-10, Mar04-10, Mar07-9, Mar10-9, Mar13-11, Mar16-11, Mar19-9, Mar22-5, Mar22-9, Mar25-5, Mar25-10, Mar28-5, Mar28-9, Mar31-5, Mar31-10	Monica Lewinsky Case
Feb02-6, Feb02-12, Feb05-11, Feb08-7, Feb08-12, Feb11-12, Feb14-12, Feb17-12, Feb20-12, Feb23-12, Feb26-12, Mar1-12, Mar4-12	Pope Visits Cuba
Feb2-0, Feb5-0, Feb5-5, Feb8-0, Feb8-5, Feb11-0, Feb11-6, Feb14-0, Feb14-6, Feb17-0, Feb20-0, Feb23-0, Feb26-0, Mar1-0, Mar1-8, Mar1-14, Mar4-8, Mar4-14, Mar7-7, Mar7-14, Mar10-7, Mar10-14, Mar13-9, Mar16-10, Mar19-8, Mar19-14, Mar22-13, Mar25-8, Mar25-13, Mar28-7, Mar28-12, Mar31-8, Mar31-14	Current Conflict with Iraq
Mar07-15, Mar10-15, Mar13-15, Mar16-15, Mar19-15, Mar22-15, Mar25-15, Mar28-15	Tornado in Florida

**Table 2.** Top-10 Results of  $C \rightarrow C$  Pattern of Query {pope, cuba}

Query {pope, cuba}	Score
Feb02-12 $\rightarrow$ Feb05-11	1.73E-06
Feb05-11 $\rightarrow$ Feb08-12	1.70E-06
Feb11-12 $\rightarrow$ Feb14-12	1.69E-06
Feb08-12 $\rightarrow$ Feb11-12	1.60E-06
Feb14-12 $\rightarrow$ Feb17-12	9.59E-07
Feb17-12 $\rightarrow$ Feb20-12	4.54E-07
Mar01-12 $\rightarrow$ Mar04-12	4.40E-07
Feb26-12 $\rightarrow$ Mar01-12	4.03E-07
Feb20-12 $\rightarrow$ Feb23-12	2.10E-07
Feb23-12 $\rightarrow$ Feb26-12	1.36E-07

**Table 3.** Top-10 Results of  $C \rightarrow C$  Pattern of Query {tornado, florida}

Query {tornado, florida}	Score
Mar19-15 $\rightarrow$ Mar22-15	8.06E-07
Mar25-15 $\rightarrow$ Mar28-15	6.72E-07
Mar22-15 $\rightarrow$ Mar25-15	5.52E-07
Mar16-15 $\rightarrow$ Mar19-15	1.34E-07
Mar13-15 $\rightarrow$ Mar16-15	5.89E-08
Mar10-15 $\rightarrow$ Mar13-15	4.90E-08
Mar07-15 $\rightarrow$ Mar10-15	4.88E-08
Mar04-16 $\rightarrow$ Mar07-15	8.12E-10
Mar01-16 $\rightarrow$ Mar04-16	7.68E-11
Feb26-16 $\rightarrow$ Mar01-16	2.38E-11

**Table 4.** Top-10 Results of Branch Pattern of Query {monica, clinton}

Query {monica, clinton}	Score
Mar28-9 $\rightarrow$ {Mar31-5, Mar31-10}	3.52E-16
Mar19-9 $\rightarrow$ {Mar22-9, Mar22-5}	2.87E-16
Mar22-5 $\rightarrow$ {Mar25-10, Mar25-5}	1.35E-16
Mar22-9 $\rightarrow$ {Mar25-5, Mar25-10}	7.04E-17
Mar25-10 $\rightarrow$ {Mar28-5, Mar28-9}	1.70E-17
Feb02-3 $\rightarrow$ {Feb05-6, Feb05-3}	6.28E-18
Feb02-5 $\rightarrow$ {Feb05-3, Feb05-6}	1.66E-18
Feb23-10 $\rightarrow$ {Feb26-9, Feb26-10}	5.78E-19
Mar19-9 $\rightarrow$ {Mar22-5, Mar22-13}	5.71E-20
Mar19-9 $\rightarrow$ {Mar22-13, Mar22-9}	4.24E-20

**Table 5.** Top-10 Results of Merge Pattern of Query {iraq, weapon, inspection}

Query {iraq, weapon, inspection}	Score
{Mar25-13, Mar25-8} $\rightarrow$ Mar28-7	1.10E-20
{Mar25-13, Mar25-8} $\rightarrow$ Mar28-12	1.71E-22
{Feb14-0, Feb14-6} $\rightarrow$ Feb17-0	3.89E-23
{Feb11-0, Feb11-6} $\rightarrow$ Feb14-0	1.98E-23
{Feb05-0, Feb05-5} $\rightarrow$ Feb08-0	1.68E-23
{Feb05-0, Feb05-5} $\rightarrow$ Feb08-5	1.26E-23
{Feb08-0, Feb08-5} $\rightarrow$ Feb11-0	6.13E-24
{Mar10-14, Mar10-7} $\rightarrow$ Mar13-9	4.51E-24
{Mar07-14, Mar07-7} $\rightarrow$ Mar10-7	1.54E-24
{Feb08-0, Feb08-5} $\rightarrow$ Feb11-6	1.38E-24

to 0.51. Ideally, the clusters of a topic in Table 1 should appear in top-ranked results of the query for the topic.

Due to space constraint, we show only results of some queries. For the transition pattern  $C \rightarrow C$ , query {pope, cuba} returns 21 instances as results; {monica, clinton} 53 instances; {tornado, florida} 24 instances; {iraq, weapon, inspection} 91 instances. Tables 2 and 3 show top-10 results of queries {pope, cuba} and {tornado, florida}, respectively. For the transition pattern of branch  $C1 \rightarrow \{C2, C3\}$ , query {monica, clinton} returns 20 instances as results; {iraq, weapon, inspection} 30 instances. Table 4 shows top-10 results of query {monica, clinton}. For the transition pattern of merge  $\{C1, C2\} \rightarrow C3$ , query {monica,

clinton} returns 19 instances as results; {iraq, weapon, inspection} 52 instances. Table 5 shows top-10 results of the query {iraq, weapon, inspection}.

Low-ranked results of a query generally have very small occurrence frequencies of the query keywords in the clusters and small transition probability values between the clusters. That is they are less similar to the topic of the query. The results in the middle ranks have either small occurrence frequencies of the query keywords and high transition probability values, or high occurrence frequencies of the query keywords and low transition probability values. The results in the top ranks in general have high occurrence frequencies of the query keywords and high transition probability values. Observing top-10 results of the above queries by comparing the clusters in the transitions of the top-10 results with the topic-labeled clusters in Table 1 reveals that most of the clusters in the top-10 results appear in Table 1. This shows that they are highly related to the queries since the query keywords are topic-representative keywords. In general, top-ranked results of a query mostly contain transitions highly relevant to the topic of the query.

## 6 Conclusions and Future Work

In this paper, we presented a declarative query language and its processing scheme to retrieve transition patterns in time-series document clustering results. The query language is composed of simple and small constructs but is powerful in retrieving interesting and meaningful patterns. The experimental evaluation confirmed the effectiveness of the query processing scheme.

Depending on the applications and user requirements, other transition patterns may be interesting. The query language is not limited to only clustering data set. It can be applied to any sequences of a time series of grouped data set, which can be regarded as clusters in some sense, and where cluster transitions can be detected. Applications of the query language to other data set and further extension of the query language may present more interesting retrieval performance and effectiveness of the language. In addition, a visualizing tool to highlight retrieval results in the sequence of periodical clustering results facilitates users in analyzing and exploring the data.

## Acknowledgements

This research is partly supported by the Grant-in-Aid for Scientific Research (19300027) from Japan Society for the Promotion of Science (JSPS) and the Grant-in-Aid for Scientific Research (19024006) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. In addition, this work is supported by the grants from Hosono Bunka Foundation.

## References

1. Adomavicius, G., Bockstedt, J.: C-TREND: Temporal Cluster Graphs for Identifying and Visualizing Trends in Multiattribute Transactional Data. *IEEE Trans. on Know. and Data Eng.* 20(6), 721–735 (2008)
2. Allan, J. (ed.): *Topic Detection and Tracking: Event-based Information Organization*. Kluwer, Dordrecht (2002)

3. Cutting, D., Karger, D.R., Pedersen, J.O., Tukey, J.W.: Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In: Proc. of 15th ACM SIGIR conference, pp. 318–329 (1992)
4. Ishikawa, Y., Hasegawa, M.: T-Scroll: Visualizing Trends in a Time-series of Documents for Interactive User Exploration. In: Proc. of the 11th ECDL Conference, pp. 235–246 (2007)
5. Khy, S., Ishikawa, Y., Kitagawa, H.: A Novelty-based Clustering Method for Online Documents. *World Wide Web Journal* 11(1), 1–37 (2008)
6. Mei, Q., Zhai, C.: Discovering Evolutionary Theme Patterns from Text - An Exploration of Temporal Text Mining. In: Proc. of ACM KDD Conference, pp. 198–207 (2005)
7. Nallapati, R., Feng, A., Peng, F., Allan, J.: Event Threading within News Topic. In: Proc. of CIKM Conference, pp. 446–453 (2004)
8. Radev, D., Otterbacher, J., Winkel, A., Blair-Goldensohn, S.: NewsInEssence, Summarizing Online News Topics. *Communications of the ACM*, 95–98 (2005)
9. Sebastiani, F.: Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1), 1–47 (2002)
10. Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., Schult, R.: MONIC – Modeling and Monitoring Cluster Transitions. In: Proc. of KDD Conference, pp. 706–711 (2006)
11. Linguistic Data Consortium (LDC), <http://www ldc upenn edu/>

# Ontology Construction Based on Latent Topic Extraction in a Digital Library

Jian-hua Yeh and Naomi Yang

Department of Computer Science and Information Engineering  
Aletheia University  
au4290@email.au.edu.tw  
Graduate Institute of Library and Information Studies  
National Taiwan Normal University  
shupiny@ms2.hinet.net

**Abstract.** This paper discusses the automatic ontology construction process in a digital library. Traditional automatic ontology construction uses hierarchical clustering to group similar terms, and the result hierarchy is usually not satisfactory for human's recognition. Human-provided knowledge network presents strong semantic features, but this generation process is both labor-intensive and inconsistent under large scale scenario. The method proposed in this paper combines the statistical correction and latent topic extraction of textual data in a digital library, which produces a semantic-oriented and OWL-based ontology. The experimental document collection used here is the Chinese Recorder, which served as a link between the various missions that were part of the rise and heyday of the Western effort to Christianize the Far East. The ontology construction process is described and a final ontology in OWL format is shown in our result.

## 1 Introduction

Manual ontology construction has been a labor-intensive work for human beings, since humans are capable of creating semantic hierarchy efficiently. But with the growing size of real world concepts and their relationships, it is more difficult for humans to generate and maintain large scale ontologies. Meanwhile, the quality of the knowledge structure in an ontology is hard to maintain because human is not able to keep the criteria of concept creation consistently. Therefore, human-generated knowledge networks are usually difficult to span, such as web directories, large organization category hierarchies, and so forth. Our experiences on constructing web ontology [1] and government ontology [2] also show that it is difficult to generate fully semantic-aware concept hierarchy purely relying on traditional data clustering algorithms. In this paper, we introduce an effective process to construct domain ontology automatically based on a special collection called the Chinese Recorder [5], which served as a link between the various missions that were part of the rise and heyday of the Western effort to Christianize the Far East. A special ontology construction process is designed and a final ontology described in OWL [3] format is shown in our result. This paper aims at two major issues, as listed below:

1. Create an effective process for ontology construction of historical documents  
Generally speaking, there are two ways to generate ontologies: manual and automatic construction. In recent years, a number of related discussions focus on the process of manual generation of ontologies, one of the most frequently referenced article is “Ontology 101” [4]. In [4], a complete ontology creation process is introduced, with the standard steps including determination of the domain and scope, consideration of reuse, enumeration of important terms, definition of the classes and the class hierarchy, definition of the slots, and creating instances. For the process of automatic construction of ontologies, many algorithms were proposed and developed, which will be discussed in the next section. One of the major goals in our research aims at aged historical collections, trying to develop an effective and automatic process to construct domain ontology. This process will relieve the burden of domain expert with decreasing the time consumption and increasing the collection scale.
2. Construct knowledge network for historical collections  
The knowledge contained in historical documents is both rich and wide-ranging, which leads the research focus of creating knowledge network or knowledge hierarchy. But most of the researches produce content classification only, which is called taxonomy in that scenario. The taxonomy only represents the tree structure of content classification, which lacks of variety of relationships among concepts. That is, no complex knowledge structure contained in such kind of structure. One of the major goals in our research aims at creation of complex structure to represent rich knowledge in historical collections and description of domain ontology using W3C OWL standard.

## 2 Issues of Ontology Construction Process for the Chinese Recorder

Before going into the introduction of ontology construction process, the special historical collection in our research is described first. We use the collection called “The Chinese Recorder” for our domain ontology construction experiment. The Chinese Recorder is a valuable source for studying the missionary movement in China and the effect the missions had on shaping Western perceptions of and relations with the Far East. It was published in English monthly for 72 years. It served as a link between the various missions by the Protestant missionary community in China, and was the only English mission that were part of the rise and heyday of the Western effort to Christianize the Far East. It provided information about individual missionaries and mission activities, recounting their progress on evangelical, educational, medical, and social fronts. It featured articles on China's people, history, and culture. The Robert's library of Toronto University has holdings for June 1868-December 1876 and Jan 1915- December 1932, while the Chinese Church Research Center (Taiwan) has holdings for all the collections among 72 years.

Currently, most of the Chinese Recorder collection copies are kept in microfilm form. In the early years of its publication (1860s), the printing methods is pretty primitive and we found that the scanning quality of the microfilms is poor (please refer to Fig. 1). If the optical character recognition (OCR) applied on such digital images,

many kinds of noise (mostly from the dirty spots on the pages) will affect the recognition effectiveness. Besides, the page formats of the Chinese Recorder varies from volume to volume (single-column, double-column, etc.), which creates certain barrier of whole chapter text extraction. This problem is also an important issue in related researches. Since the process proposed in this paper concerns only the relationship between document and terms contained in the document, the whole chapter text extraction is unnecessary and the noise effect can be controlled in our experiment.

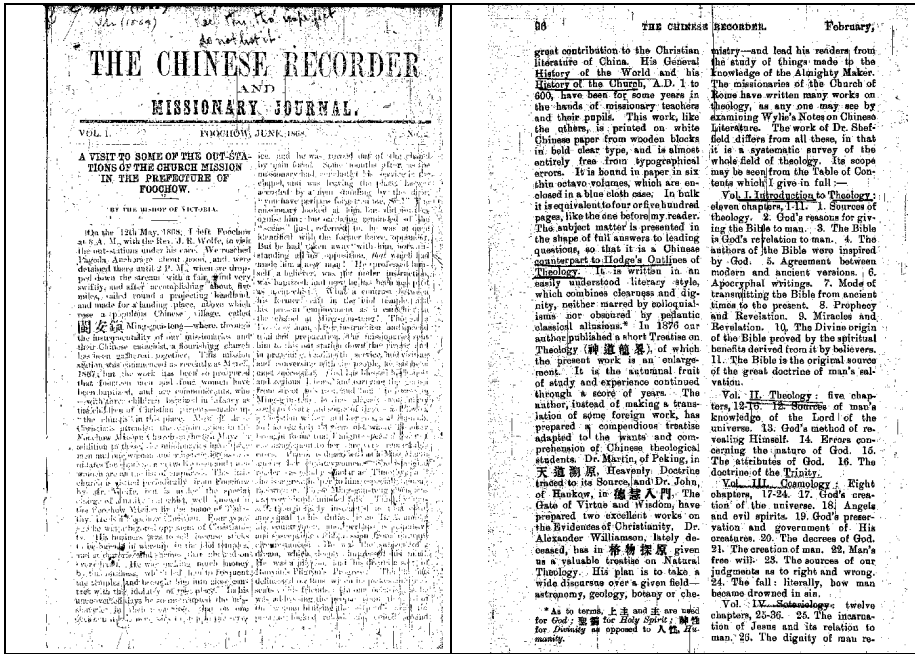


Fig. 1. The scanning quality of the Chinese Recorder microfilms is poor. (left: May 1868, right: February 1894).

When the text extraction finished, all the textual data in the Chinese Recorder are collected. The next issue will be the ontology creation process based on the extracted text. The automatic hierarchy construction researches have gained much attention in recent years [6,7]. Several classification and clustering methods have been proposed to solve the hierarchy generation problem [8,9,10,11]. Among these methods, the hierarchical agglomerative clustering (HAC) algorithms [7,12] attracts a lot of attentions since the clustering result is presented in the form of a tree structure, that is, a hierarchy is created. This kind of algorithms has great potential in automatic hierarchy generation since it keeps both scalability and consistency. While this is a promising technology, it still suffers certain shortage in the current development: the cluster generated by automatic process seldom keeps semantics, since the automatic process based on information retrieval (IR) techniques is usually term-based, not concept-based. This characteristic causes the clustering result seldom semantic-aware, which



is often not acceptable by human recognition. The automatic generation process can be improved when information retrieval and computational linguistics techniques advance, but for now, it is not applicable to generate high-quality knowledge networks through traditional clustering methods only. In recent years, the researches about topic extraction from texts are getting more and more attention, especially a promising technology called latent topic discovery. Latent topic discovery is invented to overcome the bottleneck of bag-of-words processing model in information retrieval area, trying to advance the text processing technology from pattern to semantic calculation.

For the researches in latent topic discovery, most of the research focuses aim at topic detection in text data by using term distribution calculation among the documents. Several important algorithms were developed, including Latent Semantic Analysis (LSA)[13], Probabilistic Latent Semantic Analysis (pLSA)[14], and Latent Dirichlet Allocation (LDA)[15]. LSA is one of the semantic analysis algorithms which differs from traditional term frequency-inverse document frequency (TF-IDF) model. The TF-IDF model consider the term frequency only, but the calculation of LSA combines some latent factor of textual data by adding additional vector space features such as singular value decomposition (SVD) of document-term matrix to analyze the document-term relationships. pLSA model is proposed to overcome the disadvantage found in by LSA model, trying to decrease the degree of computation by using probabilistic approach. pLSA analyzes the document-term relationships using latent topic space, just like LSA, which projects the term  $t_j$  in set  $T$  together with document  $d_i$  in set  $D$  to a set of  $k$  latent topics  $T_k$ . pLSA and LSA try to represent the original document space with a lower dimension space called latent topic space. In Hofmann [14],  $P(T_k | d)$  is treated as the lower dimension representation of document space, for any unseen document or query, trying to find the maximum similarity with fixed  $P(t | T_k)$ . Other than LSA and pLSA, the algorithm of Latent Dirichlet Allocation (LDA) is more advantageous since LDA performs even better than previous research results in latent topic detection. In fact, LDA is a general form of pLSA, the difference between LDA and pLSA model is that LDA regards the document probabilities as a term mixture model of latent topics. Girolamin and Kaban [16] shows that pLSA model is just a special case of LDA when Dirichlet distributions are of the same.

Because the latent topic discovery procedure is capable of finding semantic topics, we can further group these topics, regarding it as a semantic clustering process. All we need is to calculate the cosine similarity [17] between topics since the latent topics resolve the problems of synonym and polysemy, that is, the terms grouped under a latent topic reveal the term usage for some specific concept.

From the discussion above, it is necessary to design a latent topic based ontology construction process to ensure the successful ontology generation and overcome the difficulties created by old historical collections. In the next section, a latent topic extraction method is proposed to support automatic domain ontology construction.

### 3 The Proposed Method

This paper aims at developing an automatic domain ontology construction process for historical documents. We will also describe the final ontology with semantic web

related standards to show the knowledge structures. Before developing the construction process, all the Chinese Recorder microfilms are scanned and produce a large amount of digital images. The processing steps of this research are shown below:

### 1. Textual data generation

We adopt standard OCR procedure in this step to generate raw textual data. As mentioned earlier, a large amount of digital images were generated, and we use OmniPage® as our OCR software to produce the raw text. We found that there is about 8%-10% errors in the raw text, so we design a statistical correction methodology to correct these errors via bigram correlation model [18], as described below:

- (a) A corpus [19] is chosen to get the term bigram data, a list of 65,000 entries is fetched and form a bigram matrix  $B$ . The value of each matrix cell  $B(\text{term}_i, \text{term}_j)$  is  $\log p$  (the value  $p$  stands for the probability of  $\text{term}_j$  after  $\text{term}_i$ ).
- (b) Match the raw text with the corpus term list fetched in (a), finding the unknown term  $U$  (the possible error word) and considering with previous/next terms  $F$  and  $R$ . List possible candidate words of  $U$  based on Levenshtein distance (a kind of edit distance)[23] ( $U_1, U_2, \dots, U_k$ ) and calculate  $b = B(F, U_i) * B(U_i, R)$ , the bigger value of  $b$  shows the more possible correction term. For the consecutive term errors, we do not process them since they are below the statistical threshold in our experiment (less than 0.5%).

The second part of text generation is from the contents of The Chinese Recorder Index [20]. It contains three indexes: the Persons Index, the Missions and Organizations Index, and the Subject Index, which give call number, title, date, month and the year. Since this is a relatively young publication (published in 1986) with both good publication and preservation conditions, it is easy to fetch three kinds of term list via OCR process.

### 2. Latent topic extraction

After the raw text is generated and corrected, we are ready for latent topic extraction. In this step, the Latent Dirichlet Allocation (LDA) is used to extract latent topics from raw text generated in previous step since LDA performs smoother topic range calculation than LSA and pLSA [16]. In this research, we treat every page in the Chinese Recorder as a basic data unit called "page document". These page documents will form a document-term matrix known in information retrieval domain, generally a sparse matrix. Then the LDA estimation starts and the latent topics are generated.

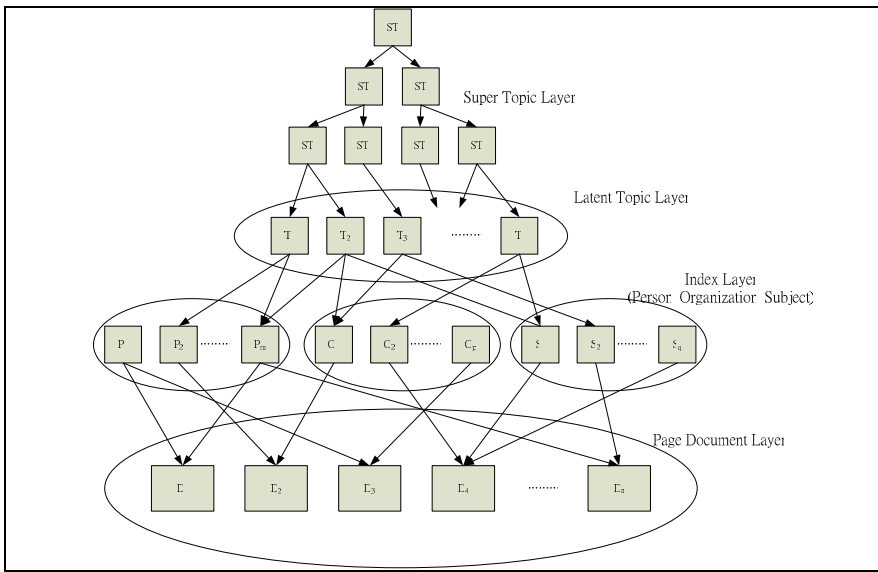
### 3. Topic clustering

In topic clustering step, we group the latent topics into higher level topics in a hierarchical manner. Because the latent topic contain semantics, so the clustering process is regarded as some kind of semantic clustering. In this research, the basic cosine similarity with hierarchical agglomerative clustering (HAC)[7,12] is adopted to generate high level topics called "super topics". The cosine similarity is calculated as follows:

Let  $t_u = \{w_{u,1}, \dots, w_{u,n}\}$  and  $t_v = \{w_{v,1}, \dots, w_{v,n}\}$  be two vectors of correlation values for the topic  $u$  and  $v$ , the topic similarity estimation function is

$$sim(u, v) = \frac{\bar{t}_u \cdot \bar{t}_v}{|\bar{t}_u| \times |\bar{t}_v|} = \frac{\sum_{i=1}^n w_{u,i} \times w_{v,i}}{\sqrt{\sum_{i=1}^n w_{u,i}^2} \times \sqrt{\sum_{i=1}^n w_{v,i}^2}}$$

The clustered super topics form a tree structure, but the whole ontology needs not to be a tree structure since the relationships among page documents, index terms, and latent topics are not simply hierarchical but graph-based, as shown in Fig. 2.



**Fig. 2.** Ontology layer and structure in this research

In Fig. 2, the nodes contained in page document layer, index layer, and latent topic layer form a graph structure. An additional feature shows that not all index terms will be referenced by latent topics because the latent topic choose terms based on obvious frequency of term occurrence.

4. OWL generation and domain expert revision

In this step, we define the OWL classes and properties for the domain ontology of the Chinese Recorder. According to the characteristics of data generated in previous steps, we propose the class and property definitions as below:

- (a) *PageDocument* class: every page text fetched from the Chinese Recorder is an instance of this class.

- (b) *Person* class: every person name shown in the Chinese Recorder is an instance of this class.
- (c) *Organization* class: every organization or mission name shown in the Chinese Recorder is an instance of this class.
- (d) *Subject* class: every subject shown in the Chinese Recorder is an instance of this class.
- (e) *Topic* class: the extracted latent topics are the instances of this class. Besides, both topic and super topics are of the instances of this class.

For OWL property, we define:

- (a) *Occurrence* property: this is the relationship between the index term (person, organization, and subject) and the page document with the index term.

The relationships between class/property/instance are shown in Fig. 3.

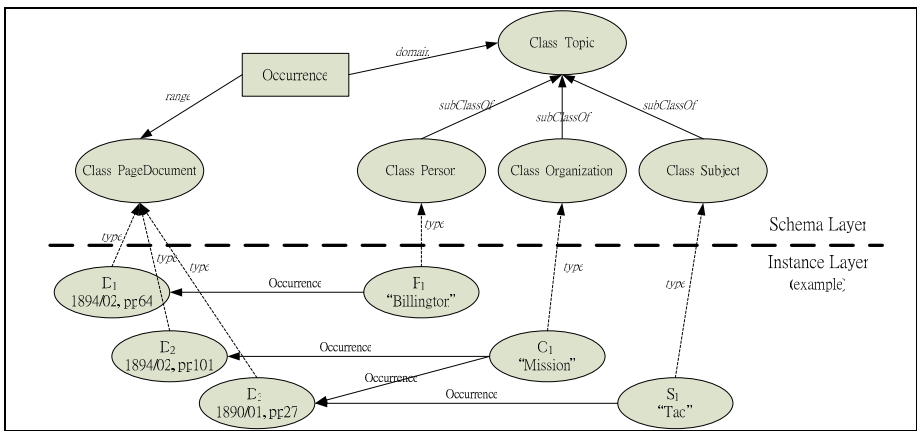


Fig. 3. Class/property/instance relationships in the Chinese Recorder ontology

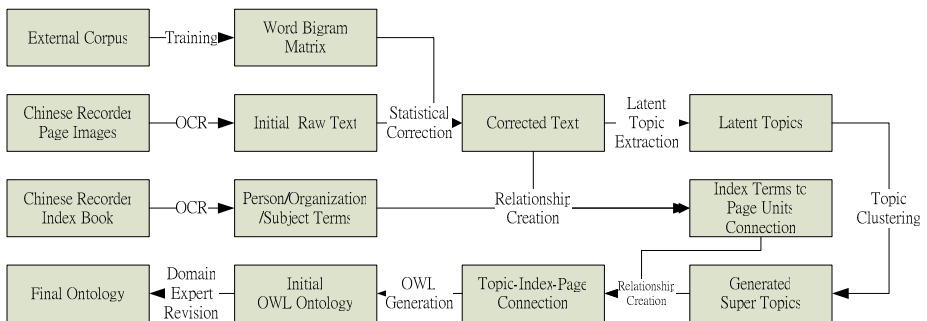


Fig. 4. The ontology processing steps in this research

With the above definitions in OWL, we can further organize the content generated in previous steps into a large graph structure, i.e. domain ontology of the Chinese Recorder. Because the data is OWL-based, it is possible to import into ontology editor such as Protégé[21] for further maintenance by domain experts.

The whole processing steps are shown in Fig. 4.

### 4 The Experiment

In this paper, we use a condensed dataset as a preliminary experiment to prove the correctness of our design. The 1890 and 1899 publications of the Chinese Recorder are chosen as our experimental data. This data set contains 204 page documents with several bad-condition pages (noises in pages). The statistical correction procedure mentioned in last section is adopted to increase the correctness of textual data. Besides, there are also non-English characters (such as Chinese contents, as shown in Fig. 1), but they are omitted in this experiment. For the index term generation, we found that the publication quality of the Chinese Recorder Index Book is pretty well, as shown in Fig. 5.

<p>Aaroe, A.K., Miss, See Olson, A.K. Aaroe [Mrs. Herman].</p> <p>Aass, D., Miss, See Bergling, D. Aass (Mrs. A. R.).</p> <p>Abbey, Louise S. Parson Whiting, [Mrs. Robert Easton], AFF:APM,6:373,9:389,10:472,14:67,23:494,25:291,464,31:324,33:320,35:272,36:270,49:210?, EAC,30:191; ARR:10:472,23:494,33:320;1880,16:428; CHI:23:494,25:464; CON:30:191;Shanghai,8:241; COR:35:198; DEP:9:389,31:324,36:270,49:210?,187B,16:428; LOC:Kiangsu,Chenchiang,14:67;Kiangsu, Nanking,6:373,9:226,389,10:472,13:395,14:67,23:484,494,25:291,464,31:324,35:272,36:270;Kiangsu, Shanghai,33:320,36:270;Shansi,T'aiyuan,16:428; Shantung,Chefoo,18:428;Japan,25:464;Turkey,16:428;United States,31:324,36:270,49:210?; OTH:25:291,30:621,33:208,256,578,629; SPO:6:373,9:226,14:67,16:428; UNS:9:226,389,16:310,428,30:191, 821,31:324,33:141,50:360.</p>	<p>1845,25:164?; DEA:25:164; DEP:25:163;1833,25:162; 1844,7:106;1845,25:164; LOC:Fukien,Amoy,7:106,15:470,18:238,19:121,25:163,26:338,31:558;Fukien, Kiangsu,31:559;Kwangtung,Canton,15:216,25:163, 65:503,504;Dutch East Indies,Batavia,25:162; Holland,25:162;Macao,19:121,25:183,65:504;Siam,7: 178,25:162;Siam,Bangkok,25:162,65:503;Singapore, 25:162,65:503; OTH:25:162; UNS:7:106,178,285,10: 148,11:338,12:154,15:216,470,18:238,19:121,25: 160-164,26:338,389,30:477,31:559,38:421,46:463, 61:772,64:728,65:503,504,67:358.</p> <p>Ackerson, Amelia C., Miss, See Conradson, Amelia C. Ackerson [Mrs. Herman J.].</p> <p>Ackzell, I.A.M., Miss, AFF:CIM,43:752,50:500,51: 812,54:630; ARR:43:752,54:630; COR:50:500-501; DEP:51:812; LOC:Shansi,Hsiao-yi,50:500;Canada,54: 830.</p> <p>Acland, ART:44:124-125; POS:Under Secretary of Foreign Affairs,44:124; UNS:44:124.</p>
---	--

Fig. 5. A partial page example of The Chinese Recorder Index

Next, the textual data generation step proposed in last section combined with statistical correction is adopted to generate 204 page documents with a total number of 95,005 words. For index term generation, a number of person, organization, and subject terms related to the document set are fetched. Related data statistics is shown in Table 1.

Table 1. Related data statistics in our experiment

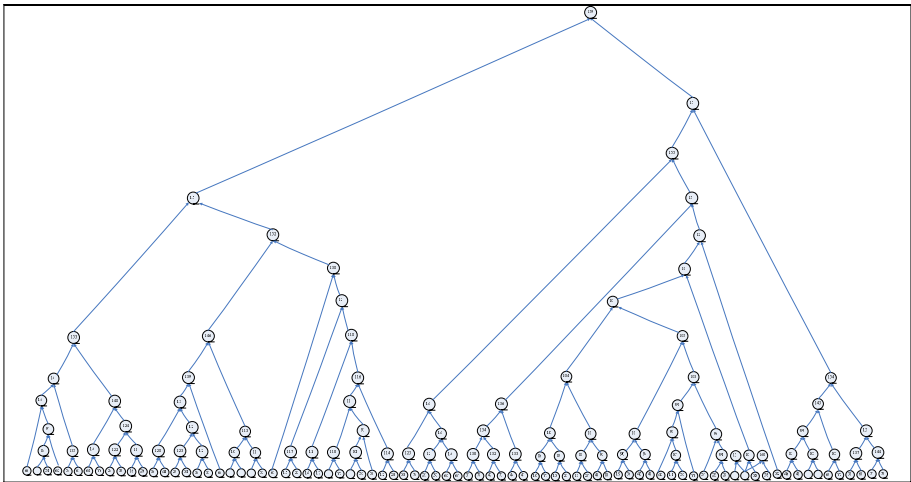
Raw text	Person index	Organization index	Subject index
95,005	544	150	180

In latent topic extraction step, 80 latent topics are extracted from 204 example page documents. These topics form the basis of topic clustering step. Partial topic list is shown in Fig. 6.

**Topic:** Love Divine Kuling hours sacred God spiritually  
**Topic:** children education school responsibility poor matter members  
**Topic:** Christian Conference Spirit movement time ground Chinese  
**Topic:** schools colleges prevent leading less American Christian  
**Topic:** Bible preaching why unable expect comment Chinese  
**Topic:** England historical bishop sufficient episcopal Church every  
**Topic:** God human Go personal everywhere everything judgment  
**Topic:** ancient Greeks contained seventh long fourth ear  
**Topic:** worship days Yellow light According sect sun  
**Topic:** method instance personal learn preaching employed first

**Fig. 6.** Partial latent topics generated by this experiment

In topic clustering step, we adopt cosine similarity as the decision function of hierarchical agglomerative clustering (HAC) algorithm to group the latent topics into “super topics”, forming the final topic hierarchy (as shown in Fig. 7). According to the index-page-topic relationships described in previous section (as shown in Fig. 4), an OWL draft of the Chinese Recorder domain ontology is generated. This ontology draft is now ready to import into Protégé software for further maintenance (the screenshot is shown in Fig. 8).



**Fig. 7.** Topic dendrogram in our experiment

After import into Protégé software, we are ready to give the ontology generation result to domain experts for further optimizations including revision and maintenance. The aspects of optimization include: 1. the original ontology classes contains Person, Organization, Subject, and Topic, the domain experts are able to refine the class hierarchy by adding more suitable classes such as deriving Event or Location subclasses from Topic class. 2. The derived classes form richer knowledge structure, which is able to add more semantics for later OWL instances.

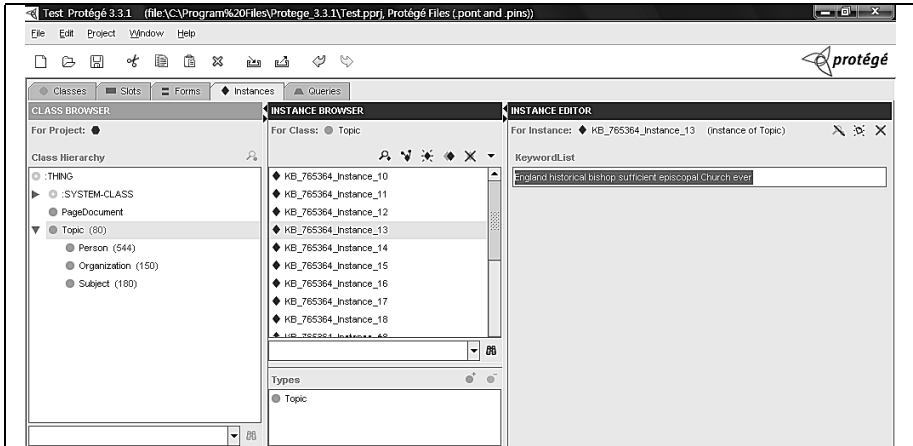


Fig. 8. OWL generated in this experiment can be imported into Protégé for further maintenance

## 5 Conclusion and Future Work

From the observation of the experimental result in this paper, we conclude that the construction of domain ontology draft is possible for historical collections through the introduction of proper information technologies. The extraction of semantics in documents shows the possibility of automatic ontology processing which minimize the human efforts in traditional ontology creation, while keeps semantics consistency with ontology generation. Meanwhile, the proposed processing steps introduce latent topic discovery as a basis of knowledge extraction, achieving both keeping important knowledge features of collection and relieving huge amount of human efforts.

The future works of this research contains: 1. expanding the data processing ranges for the whole Chinese Recorder collection, trying to create a complete domain ontology of the Chinese Recorder. 2. Describe the domain ontology with OWL-based format and published in our project web site to facilitate the content researches of the Chinese Recorder collection constantly. Besides, we will continue to improve our ontology processing algorithm to generate ontology with better semantic quality. For example, we are planning to adopt concept clustering [22] as the replacement of term vector similarity in topic clustering step through similarity propagation of term relationships.

## References

1. Yeh, J.-H., Sie, S.-h.: Towards automatic concept hierarchy generation for specific knowledge network. In: Ali, M., Dapoigny, R. (eds.) IEA/AIE 2006. LNCS (LNAI), vol. 4031, pp. 982–989. Springer, Heidelberg (2006)
2. Chen, C.-c., Yeh, J.-H., Sie, S.-h.: Government ontology and thesaurus construction: A taiwanese experience. In: Fox, E.A., Neuhold, E.J., Premssmit, P., Wuwongse, V. (eds.) ICADL 2005, vol. 3815, pp. 263–272. Springer, Heidelberg (2005)

3. Deborah, L., McGuinness, Harmelen, F.v.: OWL Web Ontology Language Overview. W3C Recommendation (February 2004), <http://www.w3.org/TR/owl-features/>
4. Noy, N.F., McGuinness, D.L.: *Ontology Development 101: A Guide to Creating Your First Ontology* (2001)
5. The Chinese Recorder, Scholarly Resources, Inc, 1867-1941
6. Jain, A.K., Dubes, R.C.: *Algorithms for clustering data*. Prentice-Hall, Englewood Cliffs (1988)
7. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. *ACM Computing Surveys* 31, 264–323 (1999)
8. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: *Proceedings of ICML 1997, 14th International Conference on Machine Learning* (1997)
9. Li, F., Yang, Y.: A loss function analysis for classification methods in text categorization. In: *The Twentieth International Conference on Machine Learning (ICML 2003)*, pp. 472–479 (2003)
10. Valdes-Perez, R.E., et al.: Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results. In: *Joint Conference on Digital Libraries (JDCL 2001)*, Roanoke, VA, June 24-28 (2001)(presented as a demonstration)
11. Yang, Y., Zhang, J., Kisiel, B.: A scalability analysis of classifiers in text categorization. In: *ACM SIGIR 2003*, pp. 96–103 (2003)
12. Widyanto, D., Ioerger, T.R., Yen, J.: An Incremental Approach to Building a Cluster Hierarchy. In: *Proceedings of the 2002 IEEE International Conference on Data Mining, ICDM 2002* (2002)
13. 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)
14. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42(1), 177–196 (2001)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(5), 993–1022 (2003)
16. Girolami, M., Kaban, A.: On an equivalence between PLSI and LDA. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 433–434 (2003)
17. Baeza-Yates, R., Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison-Wesley, Reading (1999)
18. Collins, M.: A new statistical parser based on bigram lexical dependencies. In: *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics*, Santa Cruz, CA, pp. 184–191 (1996)
19. British National Corpus, <http://www.natcorp.ox.ac.uk/>
20. Lodwick, K.L.: *The Chinese Recorder Index: a guide to Christian Missions in Asia, 1867–1941*. Scholarly Resources Inc., Wilmington (1986)
21. Noy, N.F., Ferguson, R.W., Musen, M.A.: The knowledge model of protégé-2000: Combining interoperability and flexibility. In: Dieng, R., Corby, O. (eds.) *EKAW 2000. LNCS (LNAI)*, vol. 1937, pp. 17–32. Springer, Heidelberg (2000)
22. Yeh, J.-h., Sie, S.-h.: Common Ontology Generation with Partially Available Side Information through Similarity Propagation. In: *Proceedings of the 2007 International Conference on Semantic Web and Web Services (SWWS 2007)*, Las Vegas, USA (June 2007)
23. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR* 163(4), 845–848 (1965)



# Towards Intelligent and Adaptive Digital Library Services

Md Maruf Hasan<sup>1</sup> and Ekawit Nantajeewarawat<sup>2</sup>

<sup>1</sup> School of Technology, Shinawatra University, Thailand

<sup>2</sup> Sirindhorn Int'l Inst. of Technology, Thammasat University, Thailand  
maruf@shinawatra.ac.th, ekawit@siit.tu.ac.th

**Abstract.** In this paper, we introduced a 3-Layer digital library architecture that facilitates intelligent and adaptive digital library services. We aimed at integrating DL contents with domain-ontology, user-profile and usage-pattern by means of intelligent algorithms and techniques. On top of an open-source digital library system, we developed required modules to capture and manipulate necessary data with the help of efficient techniques such as ontology-driven topic inference, collaborative filtering, single exponential smoothing, etc. We verified that our approach is capable of enhancing and adapting user profile dynamically with the help of ontology-driven topic inference and usage-pattern analysis. Usage pattern and content -based collaborative-filtering techniques are used in developing adaptive recommendation service. We also proposed a User Interest-Drift algorithm based on single exponential smoothing techniques. Our preliminary experimental results and exploratory analyses show that our approach has created positive user experience in a small digital library environment. Large scale deployment of the proposed digital library system along with further refinement of algorithms is also planned.

**Keywords:** User Modelling, Recommender System, Collaborative Filtering, Interest-drift Modeling, Ontology-based Topic Inference, Digital library.

## 1 Introduction

A digital library (DL) is a collection of documents in organized electronic form and accessible via searching and browsing interfaces [1]. Depending on the specific library, a user may be able to access magazine articles, books, papers, images, sound files, and videos. Typical DL Systems provide searching and browsing interfaces. Searching implies that the user knows exactly what to look for, while browsing should assist users navigating among correlated searchable terms to look for something new or interesting. Searching interface may range from basic keyword search to field-specific advanced search, etc. Browsing interface includes categorical navigation based on certain taxonomy and meta-data such as browse by author, category and the like [2], [3].

Information seeking in the DL context is inherently *different* from those of Web Search or Information Retrieval contexts. A common characteristic of the latter cases

is that they do not provide any personalized support to individual users, or poorly support it [3], [4]. In fact, what makes a digital library unique is the availability of contents in electronic form (which can be processed automatically and inferences can be made), and the availability of user profile and usage patterns. In contrast with the WWW, Google-like keyword search or Yahoo-like Directory is certainly not adequate for effective harnessing of information in a digital library.

The challenge of integrating DL contents with user-profile, usage patterns, etc. can be efficiently done using intelligent algorithms. In this research, we propose a *3-Layer DL Architecture* that facilitates intelligent and adaptive digital library services by integrating DL contents with domain-ontology, user-profile and usage-pattern by making use of intelligent algorithms and techniques in a unique fashion.

The outline of the paper is as follows. In the next section, we introduce the 3-Layer Digital Library Architecture, followed by the material and methods in Section 3. Further details about the domain-ontology, user profile and usage pattern capturing and analysis are also discussed in Section 3. We explain the exploratory evaluation along with experimental setup in Section 4. Finally, in Section 5, we summarize the contributions of this research and highlight further research directions.

## 2 The 3-Layer DL Architecture

There are several DL architecture and service-models found in relevant literature [5]. In the proposed 3-Layer DL Architectures, we focus on modularity and maintainability. In the core of the 3-Layer Architecture is a typical open-source DL system surrounded by a series of add-on Modules to capture and represent further information about contents, users and usages. The outer-most layer consists of services developed by making intelligent use of information represented and captured by one or more add-on modules in the middle layer, and with the help of efficient algorithms and techniques (cf. Figure 1).

We use a domain-ontology to annotate and organize DL items and to represent DL user's profile. We also capture user access logs and analyze them to obtain login frequency, search and browse history with timestamp etc. so that the temporal changes in access pattern can be computed easily. It should be noted that user initial profiles are mostly incomplete or inaccurate [3], [6]. By mapping and associating user-profile onto domain-ontology we refine the *initial* user-profile. Further enhancement of user profile is made continuously by considering the user's usage data. Through analysis of usage-pattern and by identifying temporal changes in user's interests (i.e., modeling interest-drift), we keep enhancing the profile continuously using weight-adjustment akin to *Spreading Activation* mechanism [7]. We outline the 3-Layer DL architecture in Figure 1.

As shown in Figure 1, user's interest drift is modeled using access log, user profile and bookshelf contents; dynamic profile is calculated with the help of profile data, bookshelf content and ontology-driven topic inference; and so on.

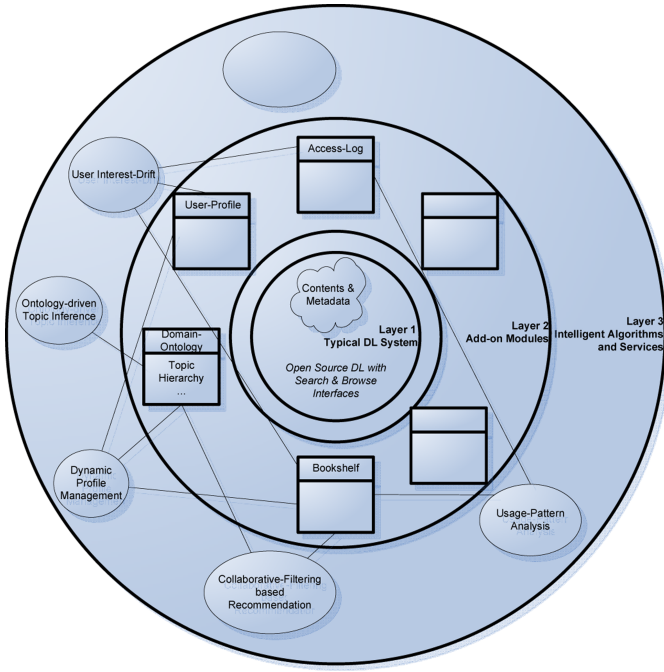


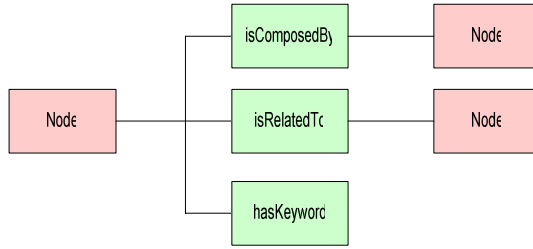
Fig. 1. 3-Layer DL Architecture for Intelligent and Adaptive Digital Library Services

### 3 Materials and Methods

In our research, we use Greenstone Digital Library System [8] in Layer 1. One of the major modules added at Layer 2 is the *domain-ontology*. For the purpose of our exploratory experiment, we make a small digital library collection with about 300 computer science related documents covering several sub-topics in computer science domain and organized/classified according to ACM Computing Classification System (CCS). Both browsing and search interfaces were made available for this small collection. We adopted a modified version of the ACM-CCS ontology [9] in organizing DL documents and representing DL user profile.

The schema of the ACM-CCS ontology is presented in Figure 2. It should be noted that the *hasKeyword* attribute is essentially useful in making topic-inference from usage-history (i.e., recording user keyword search) and through full-text analysis (i.e., automatic keyphrase extraction using natural language processing). We use automatic keyphrase extraction tool (KEA) to automatically extract keyphrases for our collection [10]. In doing so, we can *indirectly* associate one topic with the other topic by calculating keyword similarities.

Other key modules in Layer 2 includes the user access-log module which captures login frequency and other user activities such as number of search and browse interactions in each session (which is used as a parameter for interest-drift modeling as explained later). The bookshelf module captures information about the content of each user’s book-shelf, and the like.



**Fig. 2.** Domain Ontology: Schema for ACM-CCS Ontology based on CCS Taxonomy

Finally, in Layer 3, we define intelligent DL services which make use of the information captured in the lower layers as outlined below:

### 3.1 User Profile and Dynamic Profile Management

In the context of digital library, typically users are prompted to specify their interests as they sign-up for the *first* time. Users in our experimental study are students and faculty members of computer science and regular user of ACM digital library; and therefore, they are familiar with the computer science domain and ACM CCS taxonomy. We present the ACM-CCS topic hierarchy to the users at the time of signing-up and asked them to select their profiles by checking appropriate topics in the hierarchy as they deem appropriate. However, such *raw* user profile is usually incomplete or inaccurate; and often changes over time [3], [6].

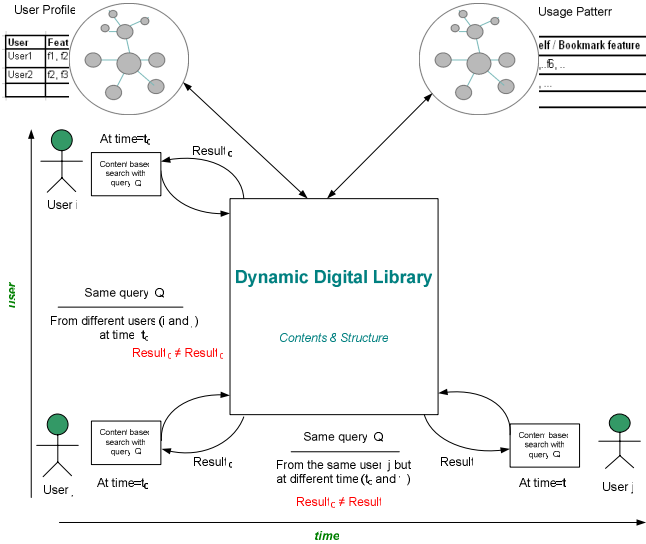
The immediate augmentation of a user's *initial* profile is made by normalizing the weights and propagating the weights using a spreading activation like mechanism [7] by considering topic interrelationships. For example, when a non-leaf topic is selected, we spread the normalized weight uniformly to all the child nodes and cross-references (if any). Such augmentations are rather rudimentary, and we refer to it as *raw user profile*. Nevertheless, by identifying and tracking content-relationships and usage history, we further refine user-profile continuously using a spreading activation mechanism. For example, user keyword searches are recorded and compared against topic-oriented keyword-list to identify the most relevant topic-area (relevancy is calculated using keyword similarity score). We call such user profiles *dynamic user-profile*. In our experiment, we re-compute user profile off-line after each login session ends. Such a dynamic profile arguably more accurately reflects a user's interest and context. By keeping extra information such as user access logs including user's keyword-sets and bookshelf-items along with timestamps, we also try to model user's interest-drift over *time* as well. The mathematical formalism of interest-drift is explained below.

### 3.2 User Interest-Drift Modeling

As a DL user undertakes new tasks or switches to new projects, an intelligent digital library should be able to adapt to such changes in context and preferences [11,12]. In another word a user searching for information with the *same keyword but on different point of time* should not be served with the same retrieval results (at least the ranking

of item should vary). Similarly, *two different users* searching the DL using the *same keyword at the same time* may not be served with the retrieval results. At least the ranking of the retrieval result should take into account of their profiles. We illustrate such adaptive digital library scenario in Figure 3.

In order to achieve such adaptivity, we use *Single Exponential Smoothing* technique [13] to model interest-drift in our digital library.



**Fig. 3.** Adaptive and Intelligent Service features in Digital Libraries that reflect User Interest-drift. (a) Horizontal Axis illustrates that search results for a particular user using the same keyword at different time yield different retrieval results. (b) Vertical Axis illustrates that search results for different users based on the same keyword may yield different retrieval results.

Single Exponential Smoothing technique has been widely used in financial forecasting and engineering applications where smoothing-based prediction is crucial. In a digital library, the user interests, information needs and information seeking behaviour change over time and therefore smoothing is inevitable to reflect such changes (interest-drift). We adopt a policy on assigning higher weights on recent usages since topics related to recent usages reflect user’s present context. We also assign exponentially weighted weights on older usage to reflect a user’s global context have to weight the user global context.

The smoothing is done using the following *iterative* formula (1):

$$S_t = \alpha Y_{t-1} + (1 - \alpha) S_{t-1} \quad 0 < \alpha \leq 1 \tag{1}$$

- where  $\alpha$  is the smoothing parameter, and
- $S_t$  is the smoothed vector (topic vector)
- $Y_{t-1}$  is the current session observations (also topic vector)

In the DL context, the frequency of a user’s login and user activities are directly related to the choice of the smoothing constant,  $\alpha$ . The  $\alpha$  values are therefore different

for each user at each session; and are estimated with the help of a normalized *activity measure*,  $f_i$ .

$$f_i = N_i / Max_N \quad (2)$$

where  $N_i$  = # of transactions for *user<sub>i</sub>*

$Max_N$  = # of transactions for the most active user during this interval

When smoothing constant,  $\alpha$  is close to 1, weight decreasing is quick and when smoothing constant,  $\alpha$  is close to 0, weight decreasing is slow. Therefore, based on the activity parameter,  $f_i$ , we choose the appropriate value of the smoothing constant,  $\alpha$  for a user.

### 3.3 Recommender Services via Usage-Pattern Analysis

Collaborative Filtering (CF) techniques have been intensively used in E-commerce. The idea is easily extendable in the context of digital library.

In the DL context, we can take advantage of both usage-pattern based and content-based collaborative filtering algorithms [14], [15], [16], [17]. It is interesting to note that in DL context, CF can be *multi-faceted* – that is, it is not only possible to recommend new DL items for a user by analyzing profile similarity and looking at their bookshelf items; but also it is possible to analyze bookshelf-item similarity (using keyword analysis) and augment user-profile. We use a *time-sensitive* CF algorithm (similar to [17] for our purpose) to incorporate adaptive effects on recommendations made.

## 4 Exploratory Study and Evaluation

In this section, we explain our exploratory study with a small collection of DL documents in CS domain and a group of users who are students or faculty in the computer science department.

### 4.1 DL Collection and Participating Users

A small digital library collection is prepared with about 300 articles in computer science domain. The articles are organized based on ACM CCS taxonomy, and full-text indexing is built using Greenstone DL software. Therefore, browsing and search interfaces are available for the collection as they are available in any typical DL system. This is our *baseline system* and kept in on the baseline server without any further augmentation.

The *same* collection is replicated to another system (hereafter, the *augmented system*) where we also implemented the Add-on modules (cf. Figure 1, layer 2 – sign-in/profile module, bookshelf, etc.) and value-added services (cf. Figure 1, layer 3 – Dynamic Profile, Collaborative Filtering, Interest-Drift Modules, etc.). A group of 10 users who are computer science students and faculty with their own research focus in computer science domain are requested to sign-up with the augmented system. They were also advised to periodically use the system over a period of 3 months, and to add as many relevant items in their bookshelf as they find them useful to their own profile and interests.

At the end of 3 month of trial period, we only found 4 users with *more* than 9 items in their bookshelves along with other usage data. In the exploratory study we invited these 4 users to perform the following experimental task.

## 4.2 The Primary Experimental Tasks

Each user is presented with 3 sets of *abstracts* consisting of 6 abstracts in each set (after deleting title, keywords, classification and other information). Each set of 6 abstracts includes 3 abstracts taken from his or her *own* bookshelf entries, and the other 3 were randomly selected from other user's bookshelf entries. There are no overlapping items in these 3 sets of abstract presented to each user. We asked all 4 users to perform the following 3 tasks with each set of 6 abstracts on the baseline system and on the augmented system as explained below. The users were told that both servers are equipped with *Camtasia* screen recording software so that we can capture their interactions for further experimental analysis. Since the screen recording is minimally intrusive to user activities, the users should therefore perform their tasks as naturally as possible without worrying about the experimental objectives but only considering the information seeking tasks<sup>1</sup>.

**Baseline Browsing Task:** For the Baseline Browsing Task, the users were asked to use the baseline system's *browsing interface* to locate the 6 target documents only through browsing. We used screen-recorder to capture user's interactions while they perform the task for further analysis. The average number of interactions each user made to locate their own 3 documents (supposedly, the familiar ones) and the other 3 documents are counted and listed in Table 1. It should be noted that the ACM CCS taxonomy consists of 4 levels (with some cross-references); and the entire DL collection consists of only about 300 items. In most cases the users succeeded to retrieve familiar documents in 4 or less interactions. However, for unfamiliar documents, the number of interactions was much higher since we consider tracking back as an additional interaction.

**Baseline Search Task:** For the Baseline Search Task, the users were given *another* set of 6 abstracts (3 new abstracts from their own bookshelf and the other 3 new abstracts were taken from the other's shelves). This time we requested the user to use keyword search on the baseline system to locate those documents via keyword base searching. Search results were presented as 5-documents/page (using Greenstone's default relevance ranking), and therefore, flipping a page is considered 1 interaction. Likewise, revision or editing of keywords is also considered as a new interaction. Same as before, we recorded user interactions passively for later analysis. The average number of interactions for each groups (own vs. others) are listed in Table 1. There was no regular pattern in the average number of interactions required to locate the documents regardless of familiar items (for documents from user's own bookshelf) or not (for documents selected from other's bookshelf).

---

<sup>1</sup> A warm-up session was conducted with 1 browsing and 1 searching task using a set of 2 abstracts for all 4 users.

**Augmented Search Task:** For the Augmented Search Task, the users were given *another* set of 6 abstracts (3 new abstracts from their own bookshelf and the other 3 new were taken from the other’s shelves). This time we requested the user to use keyword search on the *augmented* system where we already have their dynamic profile and usage histories recorded and DL contents are further augmented using keyphrase extraction, ontology-driven topic inference, etc. However, the search results in the augmented system are sorted against a user’s dynamic profiles (which in turn reflect their usage pattern; different from baseline search output). We asked the users to search for all 6 documents using keyword-bases search. Search results were presented as 5-documents/page (same as the baseline search task); and user interactions are recorded same as before. The average number of interactions for each groups (own vs. others) are listed in Table 1 for comparison. In most cases, the user’s located their own documents with a low average number of interactions while the average number of interactions required for unfamiliar documents was higher. This is due to the fact that the search result is ranked against user’s preferences and contexts and therefore flipping of pages and revision of search terms were required. The possibility of failing to use the right (combination of) keywords for unfamiliar documents can’t also be fully ruled out. However, in a small-scale exploratory experimental setting, we further explore the phenomenon by interviewing the participants.

**Table 1.** Average number of interaction in 3 different tasks. The value shows avg # of interactions for a subgroup of 3 documents.

User#	Base Browsing		Base. Search		Aug. Search	
	Own/3	Other/3	Own/3	Other/3	Own/3	Other/3
#1	3.3	6.3	3.7	4.0	2.0	5.3
#2	4.0	6.7	4.0	4.0	3.3	4.0
#3	3.7	5.0	4.0	5.3	2.3	5.0
#4	3.0	6.7	5.0	4.7	2.7	4.3

### 4.3 Other Experimental Evaluations

During the interview session, we revealed our experimental methods and objectives in detail to each user. We also presented the user Dynamic Profile (Topics with currents weights using a cut-off threshold) and their initial sign-up profile side by side and requested them to rate which profile reflect their interests better. There were both strong agreements and strong disagreements. However, when we further explained how the dynamic profile takes into account of their activities (such as keywords they used and items they chose for bookshelf) and their context (such as the, interest-drift), all 4 users tend to agree that the dynamic profile is in line with the usage and activities.

In the final phase of the exploratory evaluation, we used our recommender module to extract top 5 documents for each user. We requested the user to read the title and abstracts and validate how many of these documents they would possibly read further



if they are recommended by the system on their next login. The cumulative responses of 4 users showed that 14 out of 20 recommended documents are chosen as worth reading further.

## 5 Conclusions

In this paper we presented a 3-Layer Digital Library Architecture which targets for intelligent and adaptive digital library services. We attempted to integrate user profile, domain ontology, usage pattern and DL content analysis together in formulating intelligent and adaptive services for digital library users. Our present experimental setup and exploratory analysis using a small-prototype show that our approach and algorithms tend to integrate several facets of DL information seeking scenario; and therefore, we plan to further enhance our algorithms and integrate them seamlessly for a large-scale collection and users.

Information seeking in the context of digital library is unique and multi-faceted. Therefore, it is desirable that DL researchers will increasingly adapt recent developments if HCI-related research and user studies in the context of digital library. We therefore, plan to integrate mixed-interactions such as, integration of browsing and search, and integration of in-turn and out-of-turn interactions [18], [19] to facilitate enhanced information seeking experience in the digital library environment.

**Acknowledgment.** This research was supported by Thailand Research Fund (TRF) Grant No. MRG4880112 awarded to Dr. Md Maruf Hasan. The authors also like to thank all the participants who took part in the exploratory experiment, and Ms. Yenruedee Chanwirawong who developed the system.

## References

1. Chowdhury, G.G., Chowdhury, S.: *Introduction to Digital Libraries*. Facet Publishing, London (2003)
2. Feng, L., Jeusfeld, M.A., Hoppenbrouwers, J.: *Beyond Information Searching and Browsing: Acquiring Knowledge from Digital Libraries*, (Retrieved March 25) (2007), <http://citeseer.ist.psu.edu/421460.html>
3. Marchionini, G.: *Information Seeking in Electronic Environments*. Cambridge Series on Human-Computer Interaction. Cambridge University Press, Cambridge (1997)
4. Straccia, U.: *Collaborative Working in the Digital Library Environment Cyclades*, (Retrieved March 12) (2007), <http://dlibcenter.iei.pi.cnr.it/>
5. Hurley, B.J., Price-Wilkin, J., Proffitt, M., Besser, H.: *The Making of America II Testbed Project: A Digital Library Service Model*. The Digital Library Federation Washington DC (1999)
6. Brusilovsky, P.: *Adaptive Hypermedia*. *User Modeling and User-Adapted Interaction* 11(1-2), 87–110 (2001)
7. Crestani, F.: *Application of Spreading Activation Techniques in Information Retrieval*. *Artificial Intelligence Review* 11(6), 453–482 (1997)
8. Greenstone Digital Library Software. Project, Retrieved 2/2/2007, from <http://www.greenstone.org/>

9. ACM-CCS Add-on Ontology, University of Minho Web Site, (Accessed March 12, 2006), [http://dSPACE-dev.dsi.uminho.pt:8080/en/research\\_about.jsp](http://dSPACE-dev.dsi.uminho.pt:8080/en/research_about.jsp)
10. Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G.: KEA: Practical Automatic Keyphrase Extraction. In: Fourth ACM Conference on Digital Libraries DL 1999, pp. 254–255. ACM, New York (1999)
11. Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E., Breuel, T.: Personalized Search. *Communications of the ACM* 45(9), 50–55 (2002)
12. Dumais, S., Cutrell, E., Chen, H.: Optimizing Search by Showing Results in Context. In: ACM Conference on Human Factors in Computing Systems (CHI 2001), Seattle, WA, pp. 277–284. ACM Press, New York (2001)
13. Forecasting with Single Exponential Smoothing, NIST/SEMATECH e-Handbook of Statistical Methods. Retrieved 10/02/2007, from <http://www.itl.nist.gov/div898/handbook>
14. Liao, I.E., Liao, S.C., Kao, K.F., Harn, I.F.: A Personal Ontology Model for Library Recommendation System. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 173–182. Springer, Heidelberg (2006)
15. Middleton, S.E., De Roure, D.C., Shadbolt, N.R.: Capturing Knowledge of User Preferences: Ontologies on Recommender Systems. In: First International Conference on Knowledge Capture (K-CAP2001), pp. 100–107 (2001)
16. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based Collaborative Filtering Recommendation Algorithms. In: 10th International World Wide Web Conference (WWW 2001), Hong Kong, pp. 285–295 (2001)
17. Ding, Y., Li, X.: Time Weight Collaborative Filtering. In: 14th ACM International Conference on Information and Knowledge Management, pp. 485–492 (2005)
18. Olston, C., Chi, E.H.: ScentTrails: Integrating Browsing and Searching on the Web. *ACM Transactions on Computer-Human Interaction* 10(3), 177–197 (2003)
19. Perugini, S., Ramakrishnan, N.: Personalizing Web Sites with Mixed-Initiative Interaction. *IEEE IT Professional* 5(2), 9–15 (2003)

# Searching for Illustrative Sentences for Multiword Expressions in a Research Paper Database

Hidetsugu Nanba and Satoshi Morishita

Hiroshima City University, 3-4-1 Ozuka-higashi, Asaminami-ku  
Hiroshima, 731-3194, Japan  
nanba@its.hiroshima-cu.ac.jp  
NEC Micro Systems, 1-403-53, Kosugicho, Nakaharaku, Kawasaki  
211-0063, Japan  
morishita@nlp.its.hiroshima-cu.ac.jp

**Abstract.** We propose a method to search for illustrative sentences for English multiword expressions (MWEs) from a research paper database. We focus on syntactically flexible expressions such as “regard – as.” Traditionally, illustrative sentences that contain such expressions have been searched for by limiting the maximum number of words between the component words of the MWE. However, this method could not collect enough illustrative sentences in which clauses are inserted between component words of MWEs. We therefore devised a measure that calculates the distance between component words of an MWE in a parse tree, and use it for flexible expression search. We conducted experiments, and obtained a precision of 0.832 and a recall of 0.911.

**Keywords:** multiword expressions, a support system for writing technical documents, illustrative sentence, a research paper database.

## 1 Introduction

When non-English native speakers write or translate technical documents using English, they are often confused about how to choose proper expressions. Illustrative sentences shown with each entry word in dictionaries are useful for selecting the most appropriate expression from candidates. However, these sentences are not always useful when non-native speakers write technical documents, because while some expressions that are commonly used but not in a specific research domain are included in dictionaries, some technical expressions that are commonly used in the specific domain are not usually included in dictionaries. Therefore, a support system for writing technical documents is required. In this paper, we propose a method for searching for illustrative sentences of English multiword expressions (MWEs) from a set of research papers in a specific domain.

Nanba et al.[7,8] constructed a multilingual research paper database, “PRESRI”, by collecting more than 78,000 Postscript and PDF files published on the Internet. The database contains research papers in domains such as computer science, nuclear biophysics, chemistry, astronomy, material science and electrical engineering.

To collect research papers in a specific domain from PRESRI, we can use keyword search and citation analysis, such as bibliographic coupling [5] and co-citation analysis [10]. As PRESRI possesses information about the sources (journal titles or conference names) of research papers, we can collect illustrative sentences of MWEs that were commonly used in particular conferences or journals. We construct a system that searches for illustrative sentences of English MWEs from research papers from the PRESRI collection.

The remainder of this paper is organized as follows. In Section 2, we describe multiword expressions. In Section 3, we explain our method for searching for illustrative sentences of a given MWE. To investigate the effectiveness of our method, we conducted some tests. In Section 4, we report the results, and conclude in Section 5.

## 2 Multiword Expressions

Expressions that consist of multiple words are called Multiword Expressions (MWEs). Baldwin [3] classified MWEs as follows.

### 1. Lexicalized phrases

#### 1. Fixed expressions

Fixed strings that undergo neither morphosyntactic conversion nor internal modification (e.g., *ad hoc*).

#### 2. Semi-fixed expressions

Expressions that adhere to strict constraints on word order and composition, but undergo some lexical variation. For example, the word “oneself” in an MWE “prostrate oneself” has some variations, such as “himself” or “herself.” Compound nouns are also included in this category.

#### 3. Syntactically flexible expression

Expressions in which more than one word are inserted between their component words (e.g., *take the evidence way*).

### 2. Institutionalized phrases

In terms of syntax and semantics, these are considered as MWEs (e.g., *kindle excitement*).

In our work, we focus on searching for illustrative sentences of syntactically flexible expressions, because fixed expressions and semi-fixed expressions are easy to search for by conducting simple string matching after stemming words in target sentences. On the other hand, some restrictions are necessary when searching for illustrative sentences containing flexible expressions. In the next section, we will explain our method of searching for such sentences.

## 3 Searching for Illustrative Sentences of Flexible expressions

### 3.1 Related Works

Verb–particle construction (VPC) is a kind of “syntactically flexible expression” that consists of a verb and a particle, such as “hand in”. Baldwin [1,2] proposed the following methods to extract VPCs from texts.

1. Extract VPCs if the number of words between a particle and its governing verb is less than five.
2. Extract VPCs using method one with the restriction that the inserted words are nouns, prepositions, or verb chunks.
3. Extract VPCs using method two with a chunk grammar.

The limitation of “less than five words” has also been used for extracting collocations [9]. However, the limitation of “less than five words” does not ensure that we can search for illustrative sentences for MWEs other than VPCs comprehensively, because it is not uncommon that long phrases or clauses are inserted between component words of MWEs other than VPCs. To show the variety of illustrative sentences, we search for sentences in which more than four words are inserted between component words.

### 3.2 Our Method

Following is an illustrative sentence for the MWE “share – with.”

But Mr. Foley predicted few economic policy changes ahead, commenting that Mr. Major shares a very similar view of the world with Mr. Lawson.

The traditional method cannot detect this sentence, as there are seven words between the component words of the MWE. In Figure 1, we show a syntactic tree of this sentence. From this figure, we can find that “share” and “with” are close to each other on the tree. We therefore focus on syntactic trees for the detection of illustrative sentences.

Here, we define a measure for calculating distance between words on a syntactic tree. Figure 2 shows a flexible expression that is constructed from two words. CW and OW indicate component words of an MWE and other words, respectively. A hierarchical distance is defined as the number of nodes on the shortest path from one component word to another. In this example, as there are three nodes on the shortest path, which is shown as a bold line, the hierarchical distance is three. We extract all sentences with a hierarchical distance between components of an MWE that is smaller than a threshold value. Together with the hierarchical distance, we also use the following two definitions: “restriction of changing voice” and “insertion of a phrase”.

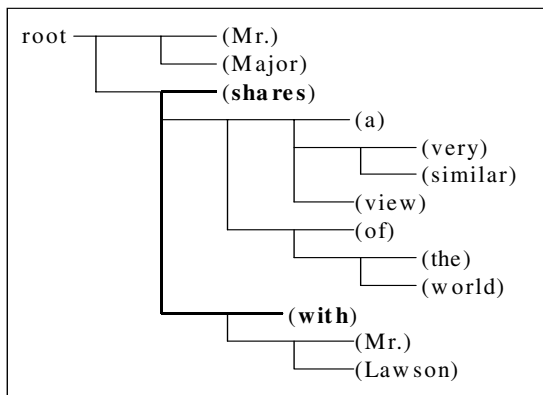
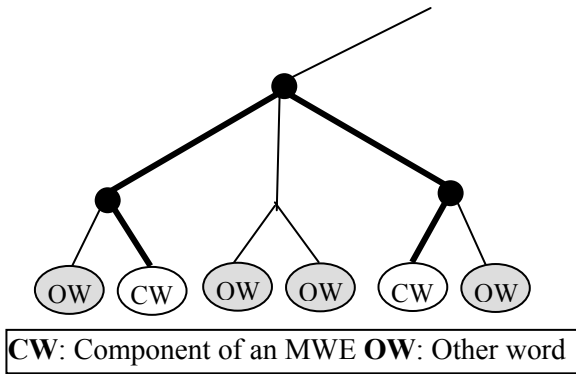


Fig. 1. A syntactic tree of an illustrative sentence for the MWE “share - with”



**Fig. 2.** An example of a hierarchical distance between component words

**Restriction of Changing Voice**

When an MWE contains a transitive verb, it is possible to change voice. However, it is considered that changing voice is not a general usage. If an MWE is generally used in the passive voice, such as “be attributed to”, the entry in dictionaries is also written in passive voice.

To confirm the validity of this assumption, we selected 21 MWEs that contain transitive verbs and investigated whether the voices of the MWEs are the same as those in illustrative sentences in three dictionaries: “Collins CoBUILD”, “Readers Plus” (Kenkyusha, Ltd.), and “New College English Japanese Dictionary” (Kenkyusha, Ltd.). The results are shown in Table 1.

**Table 1.** The ratio of illustrative sentences with voices that are different from those of entries in dictionaries

Dictionary	Ratio
Collins CoBUILD	0.11 (2/17)
Readers Plus (Kenkyusha)	0.13 (2/15)
New College English-Japanese Dictionary (Kenkyusha)	0.15 (2/13)
<b>Total</b>	<b>0.13 (6/45)</b>

There were 45 illustrative sentences for the 21 MWEs in the three dictionaries, and the voices of MWEs differ from those in the illustrative sentences in six cases (13%). Because the number of illustrative sentences used in this investigation was small, we cannot derive a concrete conclusion. However, the results do indicate that changing voice is not a general usage. Therefore, we do not search for sentences with voices that are different from MWEs.

**Restriction of Insertion of Clauses between Component Words of MWEs**

There are cases when clauses are inserted between component words of MWEs. To search for such illustrative sentences, we use the following restrictions. When a clause begins between the component words of an MWE and does not end in the same split

part, we consider that the clause is not a parenthetical clause, and eliminate the sentence from the candidates of illustrative sentences. Figure 3 shows an example in which a clause, shown as a shadowed rectangle does not end in the split part. In this case, we consider that this is not an illustrative sentence for an MWE “the same A as B”.

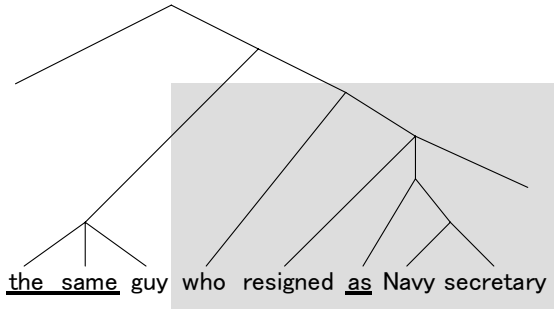


Fig. 3. An example in which a clause is not terminated between component words of an MWE

## 4 Experiments

To investigate the effectiveness of our method, we conducted some experiments.

### 4.1 Evaluation

#### Experimental Method

We used a syntactic parser [4] to search for illustrative sentences. The performance of the parser affects the search results directly; however, we could not estimate its effect. Therefore, we tested our method in two ways: (1) using manually annotated syntactic tags and (2) using the results from the syntactic parser and confirming the effects of parse errors by comparing their results.

To confirm the effects of parse errors, we used Penn Treebank (PTB)<sup>1</sup> [6]. PTB is a large corpus of *Wall Street Journal* material, in which 74,000 sentences are manually annotated with part-of-speech tags and syntactic tags. First, we tested our methods using PTB with manually annotated syntactic tags. Second, we tested using PTB with the results from the syntactic parser. Finally, we tested using 18,000,000 sentences in PRESRI with the results from the syntactic parser.

#### Alternatives

We conducted tests using the following five methods.

#### **Our methods**

- (A) Using a hierarchical distance. The maximum distance was four.
- (B) (A) + restriction of changing voice.
- (C) (B) + restriction of insertion of a clause.

<sup>1</sup> <http://www.cis.upenn.edu/~treebank/>

### Baseline methods

- (i) String matching. The number of words in a split area was not limited.
- (ii) String matching. The maximum number of words in a split area was three.

Here, we experimentally determined the threshold value as four in method A, using the data for making rules that we will describe later. In the same way, the threshold value for baseline method ii was determined as three.

### Test Collections

We manually selected 53 flexible expressions from nine books about technical writing for Japanese. We use 42 MWEs for making rules and 11 for evaluation.

We constructed test collections using the following three steps:

1. Convert all words in MWEs and in all sentences into their original forms using LimaTK [11];
2. Collect all sentences using simple pattern matching;
3. Manually identify whether the sentences collected in Step 2 are valid illustrative sentences for the given MWEs.

Table 2 shows the data that we used in our examinations.

**Table 2.** Data for the examinations

		The number of MWEs	The number of sentences for search	The number of correct sentences
PTB	For making rules	42	662	429
		42	351	219
PRESRI	For evaluation	53	2466	1720

### Evaluation Measures

We evaluate our methods and baseline methods using the following equations.

$$Precision = \frac{\text{The number of sentences that a system detected correctly}}{\text{The number of sentences that a system detected}} \quad (1)$$

$$Recall = \frac{\text{The number of sentences that a system detected correctly}}{\text{The number of sentences that should be detected}} \quad (2)$$



## 4.2 Results

In Table 3, we show the experimental result using the data of PTB with manual parse trees. As Table 3 shows, our methods are superior to both baseline methods.

**Table 3.** Results of searching for illustrative sentences using PTB (Manual)

		Precision	Recall
Baseline methods	i	0.624 (219/351)	1.000 (219/219)
	ii	0.708 (155/219)	0.708 (155/219)
Our methods	A	0.796 (207/260)	0.945 (207/219)
	B	<b>0.868</b> (204/235)	0.932 (204/219)
	C	<b>0.868</b> (203/234)	0.927 (203/219)

We also show the results using the data of PTB with parse trees by the parser. The results using the statistical parser (Table 4) are better than those using manual parse trees (Table 3), because most of the sentences that could not be analyzed by the parser happened to be incorrect as illustrative sentences.

**Table 4.** Results of searching illustrative sentences using PTB (Parsing)

		Precision	Recall
Baseline methods	i	0.624 (219/351)	1.000 (219/219)
	ii	0.708 (155/219)	0.708 (155/219)
Our methods	A	0.880 (205/233)	0.936 (205/219)
	B	0.887 (204/230)	0.932 (204/219)
	C	<b>0.889</b> (201/226)	0.918 (201/219)

**Table 5.** Results of searching illustrative sentences using PPRESRI (Parsing)

		Precision	Recall
Baseline methods	i	0.697 (1720/2466)	1.000 (1720/1720)
	ii	<b>0.870</b> (1140/1311)	0.663 (1140/1720)
Our methods	A	0.841 (1303/1549)	0.758 (1303/1720)
	B	0.849 (1277/1505)	0.742 (1277/1720)
	C	0.862 (1248/1447)	0.726 (1248/1720)

Finally, we show the result using the data of PRESRI with parse trees by the statistical parser. Baseline methods ii is superior to others, while the recall of this method is the worst.

## 4.3 Discussions

### Comparison of Baseline Method ii and our Methods

The gap of precision between method ii and our methods is more than 0.1 in tests using PTB data (Tables 3 and 4), while the gap was almost the same in the test using

PRESRI data (Table 5). This is caused by the low performance of the syntactic parser with the PRESRI data. As the syntactic parser was trained using PTB, we cannot obtain the same performance for PRESRI as for PTB.

### Effectiveness of Our Methods

Among our three methods, the precision of method C is the best. However, the precision of baseline method ii is superior to method C, although recall is the worst because the method eliminated all illustrative sentences if more than four words were inserted between component words of MWEs. However, high recall is also required in terms of variety of illustrative sentences.

### Combination of Method C and Baseline Method ii

Method C can find illustrative sentences correctly, even when many words are inserted between the component words of the MWEs, while baseline method ii can also find sentences correctly when less than four words are inserted between component words of MWEs. Therefore, it is considered that these methods can find many different sentences, i.e. it is possible to improve recall by combining both methods.

We investigated the relations between recall and precision and threshold values of method C and baseline method ii. We show the results in Figure 4. The figure shows that the precision of baseline method ii decreases when the threshold value exceeds three. On the other hand, the precision of method C is the highest when the threshold value of the hierarchical distance is four, then decreases as the threshold value increases.

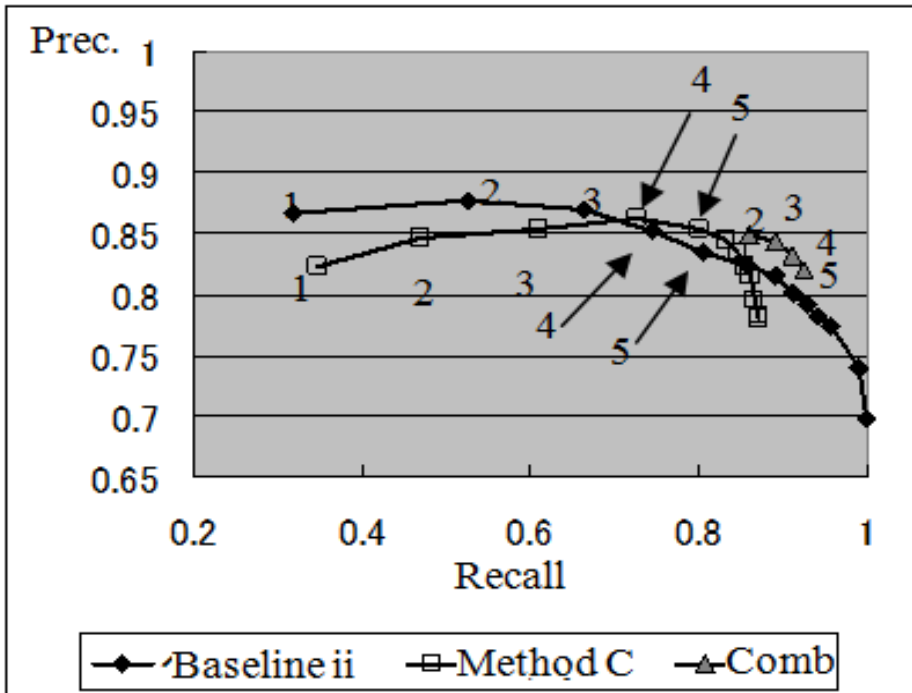


Fig. 4. Recall and Precision by baseline method ii, method C, and their combination method

We therefore combine baseline methods ii and C. When the threshold value of baseline method ii is smaller than a value  $n$ , we applied the baseline method, and when the value is larger than  $n$ , we applied method C. As a threshold value for a hierarchical distance, we used four.

We searched illustrative sentences using the combination method while changing the value of  $n$  from two to five. The results are shown as triangles in Figure 4. The figure shows that the combination method can improve recall while maintaining precision. When  $n = 4$ , we obtained the precision of 0.832 and recall of 0.911. From this result, we can conclude that the simple string matching method is useful when less than five words are inserted between component words and that using a hierarchical distance is also useful when more than four words are inserted between components.

## 5 Conclusions

We have proposed a method to search illustrative sentences of flexible expressions from the research paper database PRESRI. We conducted tests, and obtained the precision of 0.832 and recall of 0.911. From the results of the experiments, we can conclude that the simple string-matching method is useful when less than five words are inserted between component words, and that using a hierarchical distance is also useful when more than four words are inserted between components.

## References

1. Baldwin, T., Villavicencio, A.: Extracting the Unextractable: A Case Study on Verbparticles. In: Proceedings of the 6th Conference on Natural Language Learning 2002, pp. 98–104 (2002)
2. Baldwin, T., Bannard, C., Tanaka, T., Widdows, D.: An Empirical Model of Multiword Expression Decomposability. In: Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pp. 89–96 (2003)
3. Baldwin, T.: Multiword Expressions. Advanced course at the Australasian Language Technology Summer School (2004)
4. Bikel, D.M.: A Distributional Analysis of a Lexicalized Statistical Parsing Model. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, a Meeting of SIGDAT, pp. 182–189 (2004)
5. Kessler, M.M.: Bibliographic Coupling between Scientific Papers. In: Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research, pp. 68–81 (1997)
6. Marcus, M., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.: The Penn Treebank: Annotating Predicate Argument Structure. In: Proceedings of the Human Language Technology Workshop, pp. 114–119 (1994)
7. Nanba, H., Abekawa, T., Okumura, M., Saito, S.: Bilingual PRESRI: Integration of Multiple Research Paper Databases. In: Proceedings of the RIAO 2004, pp. 195–211 (2004)
8. Nanba, H., Kando, N., Okumura, M.: Classification of Research Papers using Citation Links and Citation Types: Towards Automatic Review Article Generation. In: Proceedings of the American Society for Information Science (ASIS) / the 11th SIG Classification Research Workshop, Classification for User Support and Learning, pp. 117–134 (2000)

9. Smadja, F.: Retrieving Collocations from Text: Xtract. *Computational Linguistics* 19(1 ), 143–177 (1993) (Special Issue on Using Large Corpora)
10. Small, H.: Co-citation in the Scientific Literature: A New Measure of the Relationship between Two Documents. *Journal of the American Society for Information Science* 24, 265–269 (1973)
11. Yamashita, T., Matsumoto, Y.: Language Independent Morphological Analysis. In: *Proceedings of the 6th Conference on Applied Natural Language Processing*, pp. 232–238 (2000)

# Query Transformation by Visualizing and Utilizing Information about What Users Are or Are Not Searching

Taiga Yoshida, Satoshi Nakamura, Satoshi Oyama, and Katsumi Tanaka

Department of Social Informatics, Graduate School of Informatics, Kyoto University  
Yoshida-Honmachi, Sakyo, Kyoto 606-8501 Japan  
{yoshida, nakamura, oyama, tanaka}@dl.kuis.kyoto-u.ac.jp

**Abstract.** The usage of search engines to obtain necessary information from the WWW has become popular. It is hard for users to make proper queries when they want to retrieve ambiguous information or topics that they are not familiar with. In this paper, we propose a system that helps users to retrieve information by visualizing efficient keywords. The system extracts important terms from search results and plots them on a two-dimensional graph, and the user can look down upon the tendency of search results. The system enables the user to re-rank and re-search search results dynamically by moving terms displayed on the graph. In addition, we attempt to extend the searching area by presenting complementary information related to the query. We verified the usefulness of our method by applying it to a web page search and a digital library.

**Keywords:** information retrieval, visualization, interactive operation, query transformation.

## 1 Introduction

If we want to search web pages which contain information we want, we usually use one of two methods. One way is to trace links in a Links-page which is a collection of links to domain specific pages such as social bookmark services. Another way is to retrieve information with a web search engine. Many Links-pages are constructed by human inputs, and they usually have descriptions of target web pages, therefore we can retrieve information efficiently. However, it is rare that we know about Links-pages beforehand, so we usually use a web search engine in order to find them.

If we use one keyword as a query, there would be many topics in pages of search result. Therefore, we must construct a query which can confine topics in result pages. However, it is usually difficult for us to choose keywords and construct a query connecting keywords with OR or NOT operators. For example, when a user executes “iPod” as a query with a search engine in order to buy iPod, the result is composed of many pages concerned with learning about iPod and a few pages about buying iPod. This is because “iPod” as a query is not sufficiently effective at confining the target topic. If the user wants to improve the search query to the target topic, he/she should add keywords such as “buy”, “order” or “shopping” and so on. However topics of search result pages will drastically change after changing the query. Therefore, he/she must try many queries in order to find appropriate pages.

*Google suggest*[1] is a system which attempts to resolve this problem. This system supports users in searching for their ideal pages by providing additional keywords which are frequently used with a query. However this method depends on query logs and does not consider topics in result pages. Therefore the system cannot identify what topics will be retrieved after a query transformation, and often provides keywords which do not allow users to find appropriate pages efficiently.

Therefore we attempt to allow users to understand topics in search results extracting important keywords in search result pages as topic terms, and plotting these topic terms and their relation on a graph. In addition, we enable users to re-rank search results and re-search by moving topic terms.

## 2 Related Work

There are many systems which visualize web pages or information[2]. *KeyGraph*[3] is a system which visualizes important words in text data by network graph in which important words as nodes connected each other. But *KeyGraph* is intended for one text file and not intended for a set of documents such as web search results.

*Natto view*[4] is a system which visualizes web space by 3D graphics. In this system, if a user lifts a node with a mouse, related nodes will lift together. So, the user can easily see nodes connected strongly. However, *Natto view* doesn't focus on filtering information or re-ranking search results by graph operations.

Clustering[5][6] is a method for categorizing documents. *Clusty*[7] is one of the clustering search engines. When a user executes a query, the system presents search results and some categorized groups with labels. By looking these labels, the user can narrow down search results. When using a clustering search engine, however, the user cannot decide how the system classify search results.

*Yahoo! Mindset*[8] is a system which reflects user's intention. This system enables a user to re-rank web pages according to whether they are for shopping or for researching. *I2Ir*[9] is a system with which a user can re-rank search result pages by operating axes presented in a radar chart. The system re-ranks result pages according to values of axes. These systems are same as our system in point of re-ranking search result pages by users' operation. However these systems don't generate new axes dynamically. And these systems do not designed for re-searching another query.

Matsuike et al.[10] made a system which assists Web search by presenting terms extracted from result pages. This system supports users' discovery of knowledge and transformation of queries by visualizing keywords in the form of a tree structure. However, types of queries generated by the system are confined. Our system is different from this system in point of reflecting users' operation flexibly.

## 3 Overview

In our research, we propose a system which helps users to understand topics in search result pages and thus find appropriate pages. The system displays important terms in result pages in a two-dimensional graph named "keyword map" instead of using a list of search result items because in this way the user can understand the results at one

glance. In this paper, we define important terms in result pages as topic terms. The user can transform queries and re-rank result pages by operating topic term nodes displayed on the keyword map.

### 3.1 Topic Terms and Co-occurring Terms

A page of search result returned by a search engine consists of the form of a textual list of search results. A search result is constructed by title of the web page, URL to the page and snippet which summarizes the web page. There are many topics in the page of search result, and a main topic is referred to in many snippets. However, a user must read almost all the snippets in the page of search result to understand what topics are described in the result pages.

In our research, we propose a system by which a user can find out a tendency of topics in result pages without looking at a list of search result items. Among terms that appear in snippets, there are some terms that are strongly associated with a topic in search result pages. For example, when we execute a query “jaguar” with a search engine, if there are keywords like “animal” or “zoo” in a result page, we will realize the page concerns the animal “jaguar”. Meanwhile, if there are some keywords like “car” or “ford”, we will realize the page concerns the car “jaguar”. We define these separable keywords as “topic terms” because they are involved in making topics. We visualized the relationship between query keywords and topic terms on a keyword map.

Moreover, a user will easily find topics in a page listing search result items by knowing what keywords are closely connected with topic terms. Therefore, we define these surrounding keywords as “co-occurring terms” and visualize them along with topic terms.

### 3.2 External Topic Terms

When a user searches for certain information with a search engine, he/she would be interested in relative information which he/she has not intended to search. For example, the user who executes a query “iPod” would be interested in other portable mp3 players such as “ZUNE”, “walkman” and so on. However, there are very few pages about “ZUNE” or “walkman” in search result items of a query “iPod”.

Our system automatically searches complementary information by constructing a new query consists of topic terms. And the system extracts some keywords about complementary information from snippets in the search result. We define these keywords as “external topic terms”.

### 3.3 Visualization and Manipulation

It is difficult to grasp the relationship between each topic term and the topic's weight merely by looking at a list of search results. Tag cloud[11] is a method which shows the importance of terms by changing the font size of them. However, Tag cloud does not consider about relationship between terms.

This affected how we chose the form of network structure for the system interface. The system plots query keywords, topic terms, and co-occurring terms, connecting them to each other with lines. In this paper, we call this network structure a keyword map. We call terms on a keyword map “nodes”, and lines between nodes “edges”.

On the keyword map, the system plots a query as “query node”. The system plots topic terms as “topic nodes”. Query node and topic nodes are connected with edges and a user can move every node by drag-and-drop operation. Terms which co-occur with topic terms are plotted as “co-occurrence nodes”. They are plotted around topic nodes. An image of nodes on a keyword map is shown in Fig.1.

Using a conventional search engine, a user must make a multi-keyword query in order to reduce useless pages. And desirable pages will not be obtained unless a user chooses appropriate keywords. If a user uses OR operators in a query, he/she might be able to find appropriate pages more flexibly. However, it is difficult to make effective queries with OR operators as described in the paper written by White et al.[12].

In our research, we propose a system with which a user can weight each keyword smoothly by operating keyword candidates which are presented on the screen with a mouse. Our system not only re-ranks but transforms a query if a user weights a keyword strongly.

A user can generate a new query using AND or NOT operators and re-rank result pages by changing a distance between a query node and a topic term node in the keyword map. The behaviors of our system depend on a distance between a query node and a topic node as follows:

- AND search if the distance is shorter than a threshold  $L_1$
- NOT search if the distance is longer than a threshold  $L_2$
- Re-ranking according to the distance if it is between  $L_1$  and  $L_2$

Fig.2. shows how the system re-ranks search results or reconstructs queries according to a layout of nodes on a keyword map. When a query changes, topics in search result pages also change. So if a query changes, our system calculates new topic terms and co-occurrence terms, and plots a keyword map again.

If a user wants to find information about “external topic terms”, the user can use an “external topic node” as a normal “topic node” by drag-and-dropping the node into a keyword map.

## 4 Design and Implementation

The steps of using the system are given below.

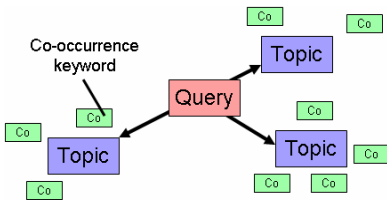


Fig. 1. Visualization of terms

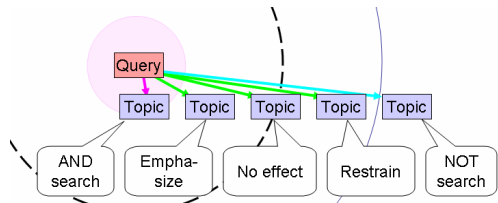


Fig. 2. Query transformation

1. A user inputs a query to the system.
2. The system presents a query node, topic nodes and co-occurrence terms of a page listing search result items on a keyword map.



3. The user operates nodes on the keyword map according to the user's intention.
4. The system re-ranks a list of search result items according to the layout of keyword map.
5. The user repeats operations until appropriate pages are found.

Fig.3. is a system image when a user executed a query. A keyword map is presented on the left side of the window. The system shows topic terms of the query as topic nodes in a keyword map.

Query transformation and re-ranking are executed by approximating important topic terms and keeping away unnecessary topic terms. A list of re-ranked pages is shown in the right section of the system. There are titles, snippets, and URLs of result pages.

#### 4.1 Extracting Topic Terms

When a user inputs a query and presses the “Search” button, the system executes the query with Yahoo web search and some result pages are obtained. The system extracts terms from snippets of the result pages and calculates their DF value.

Then the system chooses topic terms from these extracted terms. We defined topic terms as terms which can confine topics in pages by adding the term to the query. The system extracts topic terms by the method shown below.

1. Sort terms according to DF values in descending order and name  $T_i$  ( $i = 1 \sim n$ ) in order. Continue from 2. to 5. for each term until 10 topic terms are extracted.
2. Divide result pages into two groups  $P_i$  and  $N_i$ . If a page contains term  $T_i$ , categorize the page into  $P_i$ . Otherwise categorize it into  $N_i$ .
3. Calculate DF value for pages in  $P_i$  and make a DF list for all  $T_i$ . Name the list  $DF_P = \{DF_{P_i} \mid i = 1 \sim n\}$ . In the same way, make a DF list from pages in  $N_i$  and name the list  $DF_N = \{DF_{N_i} \mid i = 1 \sim n\}$ .
4. Calculate the cosine similarity between  $DF_P$  and  $DF_N$ .
5. If the cosine similarity is less than the threshold (= 0.6), add the term to the keyword map.

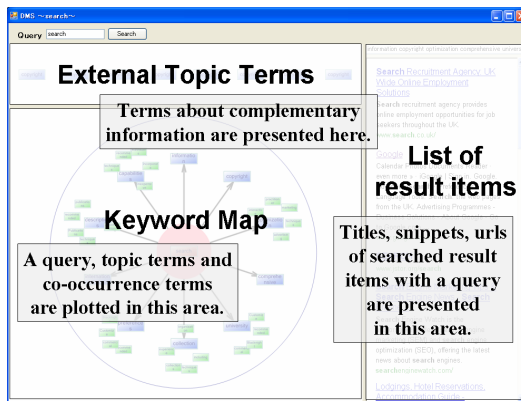


Fig. 3. System image

## 4.2 Extracting Terms that Co-occur with Topic Terms

When a user confines result pages to pages which contain query keywords and an extra topic term, topics in pages change radically. Therefore terms in snippets will change too.

The system presents some terms around topic terms on a keyword map. These terms are chosen by selecting terms whose frequencies of appearance increase when pages are confined to those which contain query keywords and the topic term. A user can decide whether or not to emphasize the term by looking at terms around the topic terms on the keyword map.

Terms co-occurring with topic terms are extracted by the method below.

1. For all topic terms, continue from 2. to 6.
2. Divide result pages into two groups. One group is the "Positive group". Pages in this group contains the topic term in their snippets. If a page does not contain the topic term, the page is classified into the "Negative group".
3. Count the aggregate number of pages in the "Positive group" and call it  $n_P$ . Then count the aggregate number of pages in the "Negative group" and call it  $n_N$ .
4. For all  $T_i$  ( $i = 1 \sim n$ ), calculate the DF value for pages in the "Positive group" and call the value  $DF_{PT_i}$ . In the same way, calculate the DF value for pages in the "Negative group" and call the value  $DF_{NT_i}$ .
5. Calculate the chi-square value using  $n_P$ ,  $n_N$ ,  $DF_{PT_i}$ ,  $DF_{NT_i}$ .
6. If the chi-square value is more than the threshold, adopt the term  $T_i$  as a co-occurrence term.

A Chi-square value  $S_i$  is calculated by the formula below. If the value is more than the threshold, it can be said that an occurrence rate of the term in the "Positive group" is different from the frequency in the "Negative group."

$$S_i = \frac{(n_P + n_N) \{ DF_{PT_i} (n_N - DF_{NT_i}) - DF_{NT_i} (n_P - DF_{PT_i}) \}}{DF_{PT_i} (n_P - DF_{PT_i}) DF_{NT_i} (n_N - DF_{NT_i})} \quad (1)$$

## 4.3 Extracting External Topic Terms

To support a user to find information which he/she does not intend to search, the system presents him/her some keywords relevant to complementary information about the query and topic terms. We call these terms "external topic terms". The system extracts external topic terms by the method shown below.

1. Choose three terms  $T_1$ ,  $T_2$  and  $T_3$  by selecting the most frequently emerging terms from topic terms.
2. Construct a query  $Q = (T_1 \text{ or } T_2 \text{ or } T_3)$  not  $Q_o$ .  $Q_o$  means the query which is inputted by the user.
3. Execute the query  $Q$  and extract some terms by the method similar to extracting topic terms.

#### 4.4 Weighting Topic Terms

In our system, the system re-ranks search results according to a layout of topic nodes on a keyword map. A user can re-rank pages by moving topic nodes with a mouse.

Each page has its own score. The score  $S_j$  of a page is calculated by the formula shown below.  $d_i$  means a distance between a query node and a term node  $T_i$ . And  $x_{ji}$  is a value if a snippet of a page contains term  $T_i$ ,  $x_{ji} = 1$ ; otherwise  $x_{ji} = 0$ .  $\theta$  is a threshold.

$$S_j = \sum_{i=1}^n \left( \frac{x_{ji}}{d_i} - \theta \right) \quad (2)$$

If the user moves a topic node toward the center of a keyword map, the score of pages which contain the topic node in their snippets increase. Searched result pages are sorted in descending order by  $S_j$ , and some pages which have high score are presented as a list of pages on the window.

#### 4.5 Extension for Google Scholar

We designed a system for *Google scholar*[13]. First we explain about *Google scholar*. *Google scholar* is a web site which is used for searching papers. A user can search for papers by inputting free keywords, author names, publication names or publication date as a query. Titles, summaries and conference information etc. about papers are given as search results. When a user uses *Google scholar*, he/she can retrieve papers by many attributes. However, many users find it hard to make full use of these.

When using our system, a user inputs keywords in a text box as a query. If he/she wants to input author name, publication name or publication date, he/she can use those attributes as a query. The user can execute the query by clicking on the “Search” button. After clicking the button, the search results of *Google scholar* are presented on the right side of the system.

In conjunction with presenting search results, the system extracts some terms from authors, titles, summaries, conference names, years and publication names of the search results. These terms are presented on a keyword map. Some common, short words such as “is” or “in” are excluded from the extracted terms. The user can change what attribute terms to plot by clicking a button presented under the keyword map.

On the keyword map, term nodes are plotted reflecting the frequency of terms in a search result. Therefore, we can regard a paper that contains many terms plotted in the vicinity of the center of the keyword map as a typical paper for the query. The system calculates scores for every paper based on the frequency of each term and the distance from the position of each term to the center of the keyword map. Each paper is re-ranked in descending order of the calculated score. In this way, searched papers are re-ranked in order from papers strongly related to the query to papers that are not related to the query. Some of the highly ranked papers are presented as a list of papers on the right side of the window.

A user can grasp the term tendency of search results by looking at the keyword map. If there is a term which he/she thinks is related to the topic he/she wants to browse,

he/she can find an appropriate paper by moving the term toward the center of the keyword map. On the other hand, if a user does not want to browse papers which contain a certain term, he/she can remove such papers by moving the term toward the outer side of the keyword map. When a user moves a term node, the system automatically calculates the scores of papers and re-ranks these papers. “External topic nodes” are plotted around a circle in a keyword map. They usually have no effect on search results, but they act as well as normal “topic nodes” when they are drag-and-dropped into a circle in the keyword map. Fig.4. shows a system image for *Google scholar*.

## 5 Evaluation and Discussion

To validate a feasibility of the system, we performed some experimentation. A task of the experimentation is identification of persons who is sharing his/her name. In the experiment, we ascertained whether the system can present pages about a specified person when we moved topic terms related to the person on a keyword map.

We performed this experimentation for 10 names. In each trial, we decided one person from some persons who share the same name, and chose 3 terms from topic term nodes on a keyword map. And we counted the number of correct pages in the top 20 pages after re-ranking.

Before moving topic terms, the average number of correct pages in the top 20 pages was 6.6. The number of correct pages increased to 11.0 after moving 1 node, 12.8 after moving 2 nodes, and 13.1 after moving 3 nodes. This experimentation shows that the system can provide users with proper pages which match the user’s intention by simple operations.

A difference between our system and a conventional search engine is the method of displaying search results. A conventional search engine presents search results as a list of their titles and snippets. On the other hand, our system visualizes keywords in search results as a keyword map. By viewing the keyword map, a user can understand the tendency of topics in search results. So, our system is good for searching information when there are many topics in search result items.

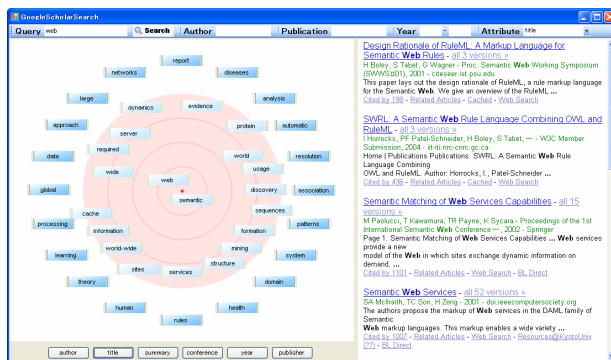


Fig. 4. System for Google scholar

Searching papers with *Google scholar* is different from searching web pages with a search engine in some points. At first, there is a large number of attributes in each result papers. When a user search papers, these attributes(conference, year, publisher) are attached importance if the user know the attribute values of papers he/she want to find. In a *Google scholar*, however, a user can communicate these attribute values to the system only as query keywords. So the user cannot specify attribute values in detail. Our system resolves this problem. The user can grasp the tendency of attribute values by looking at a keyword map and can specify attribute values even if he/she wants to emphasize the keyword after looking search results.

Second, topics in each paper are more concrete than topics in web pages, and usually their topics are different from each other. So it is difficult for users to make queries if a user want to find papers which the user does not know well about their attribute values(titles, authors etc.). Our system is adapted for searching papers when the user wants to indicate of attribute values vaguely. For example, when a user wants to find papers about database, papers obtained by a query “database” contain many topics. In this case, the user can pick out papers in which he/she has an interest with our system by operating a keyword map. The system is also adapted for a situation when the user wants to specify some keywords as author names but he/she is not sure whether all names are in the author list of the paper.

The system for *Google scholar* is an instance of our method. The availability of the method does not depend on what information to search, so the method can be applied to other search engines. For example, a user can find ideal information by our method when he/she searches some products at online shopping sites. When applying the method to social bookmarking services, a user can search pages intuitively and flexibly by moving topic nodes which correspond to tags.

## 6 Conclusion

In this paper, we proposed a system that visualizes search results and re-ranks those results according to user operations. The system extracts topic terms from search results, and proposes them in a two-dimensional graph named a keyword map. A user can find appropriate results just by moving terms with the mouse.

Generally, a user uses nouns as keywords for a query. Therefore we mainly extract and plot nouns on a keyword map in our system. However, there are times when we want to know about the reputation of a certain product, or want to know about something but we cannot conceive concrete keywords. In these cases, we may want to input adjective keywords as a query. However, it is hard to find results which meet the purpose because it is rare that these results contain the keywords, even if the contents of the results are related to the keywords. We plan to improve the system so as to present adjective keywords to a user and enable him/her to re-rank search results by moving them. For this purpose, we have to conceive a method of associating adjective terms with search results which do not contain those keywords.

When a system proposes topic terms, it is important to enhance the coverage of topics. In order to present various keywords, we will extend the system to change presented terms dynamically according to user operations. For example, when a user moves a certain keyword to the outer side of the keyword map, it means he/she does not

want to browse pages or papers about the keyword. Therefore, the system removes some keywords related to the keyword and presents new keywords instead of those just removed. Then, a user can browse many topics by removing keywords which he/she is not interested in.

Our system enables a user to construct queries and re-rank search results by operating a keyword map. When using our system, a user does not need to construct complicated queries, so he/she can find appropriate results using a simple process. To make a system which can be operated more intuitively, we also implement a system that presents information about search results intelligibly.

## Acknowledgments

This work was supported in part by "Informatics Education and Research Center for Knowledge-Circulating Society" (MEXT Global COE Program, Kyoto University), and by MEXT Grant-in-Aid for Scientific Research in Priority Areas entitled: "Content Fusion and Seamless Search for Information Explosion" (A01-00-02, Grant#: 18049041) and "Design and Development of Advanced IT Research Platform for Information" (Y00-01, Grant#: 18049073).

## References

1. Google suggest, <http://www.google.com/webhp?hl=en&complete=1>
2. Chen, C.: Information Visualization. Springer (2004)
3. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor. Research and Technology Advances in Digital Libraries (1998)
4. Shiozawa, H., Nishiyama, H., Matsushita, Y.: The Natto View: An Architecture for Interactive Information Visualization. IPSJ Journal (1997)
5. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Advanced Reference Series. Prentice-Hall, Englewood Cliffs (1988)
6. Cutting, D.R., Pedersen, J.O., Karger, D., Tukey, J.W.: Scatter/gather: A cluster-based approach to browsing large document collections. In: Proc. of SIGIR 1992, pp. 318–329 (1992)
7. Clusty, <http://www.clusty.com/>
8. Yahoo! Mindset, <http://mindset.research.yahoo.com/>
9. 121r(one to one ranking system), <http://www.kbmj.com/service/products/121r.html>
10. Matsuike, Y., Zettu, K., Oyama, S., Tanaka, K.: Approximate Intentional Representation showing the Outline and Surrounding of Web Search Results and its Visualization, DBWS 2004 (2004)
11. Sinclair, J., Cardew-Hal, M.: The folksonomy tag cloud: when is it useful? Journal of Information Science (2008)
12. White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: Proc. of SIGIR 2007, pp. 255–262 (2007)
13. Google scholar, <http://scholar.google.co.jp/schhp?hl=en>

# Language Independent Word Spotting in Scanned Documents

Sargur N. Srihari and Gregory R. Ball

Center of Excellence for Document Analysis and Recognition (CEDAR)  
Department of Computer Science and Engineering  
University at Buffalo, The State University of New York  
Buffalo, New York 14228, USA  
srihari@cedar.buffalo.edu

**Abstract.** Large quantities of scanned handwritten and printed documents are rapidly being made available for use by information storage and retrieval systems, such as for use by libraries. We present the design and performance of a language independent system for spotting handwritten/printed words in scanned document images. The technique is evaluated with three scripts: Devanagari (Sanskrit/Hindi), Arabic (Arabic/Urdu) and Latin (English). Three main components of the system are a word segmenter, a shape based matcher for words, and a search interface. The user gives a query which can be (i) A word image (to spot similar words from a collection of documents written in that script) or (ii) text (to look for the equivalent word images in the script). The candidate words that are searched in the documents are retrieved and ranked, where the ranking criterion is a similarity score between the query and the candidate words based on global word shape features. For handwritten English, a precision of 60% was obtained at a recall of 50%. An alternate approach comprising of prototype selection and word matching, that yields a better performance for handwritten documents is also discussed. For printed Sanskrit documents, a precision as high as 90% was obtained at a recall of 50%.

## 1 Introduction

Large quantities of scanned handwritten and printed documents are rapidly being made available for use by information storage and retrieval systems, such as for use by libraries. Such documents come both from historical and contemporary sources worldwide and are available in many languages. Unfortunately, indexing such documents for use by conventional search engines is time consuming and expensive.

One common task for such databases of scanned documents is search. Users generally need a way to access the interesting documents by providing a keyword or phrase, much as they search for webpages using search engines. However, search technology is generally predicated upon having complete transcripts of text. Obtaining such complete transcripts for scanned documents is often infeasible due to the time and expense.

To this end, we describe a word spotting method useful in searching scanned documents based on two important criteria: (i) language independence and (ii) non-reliance on complete transcripts. Word spotting is a content-based information retrieval task to find relevant words within a repository of scanned document images. Such retrieval tasks are present wherever such databases of scanned documents are present—digital libraries, historical document processing, forensics, personal records, medical records, etc. All these applications have in common an interest in retrieving a subset of documents from a large database of documents based on the visual and the textual content.

Spotting handwritten words in documents written in the Latin alphabet has received considerable attention [1], [2] and [3]. A systematic comparison of seven methods was made in [4], who showed that the highest precision was obtained by their method based on profiles and dynamic time warping (DTW). A method based on word shape was shown to perform better than the method based on DTW, both in terms of efficiency and effectiveness, in [5].

This paper discusses the performance of the word shape method presented in [5] when applied to other scripts (Devanagari, Arabic) and languages arising from these scripts such as Sanskrit/Hindi and Arabic/Urdu. The paper also compares the retrieval performances in these scripts and those between printed and handwritten documents.

The word spotting method presented here involves the indexing of documents that involves (i) line and word segmentation and (ii) feature extraction—global word shape features known as GSC features are extracted for each of the words in the documents. The similarity between word images is measured using a correlation similarity measure. The system that is mentioned in this paper includes functionalities obtained from the CEDARABIC system [6], and the CEDAR-FOX system, which was developed for forensic examination [7], [8].

The remainder of this paper describes the word spotting technique, including the indexing of the documents and the computation of the similarity between words, the datasets used and the experiments conducted, along with the retrieval performance for word spotting in each of the three languages.

## 2 Word Spotting Technique

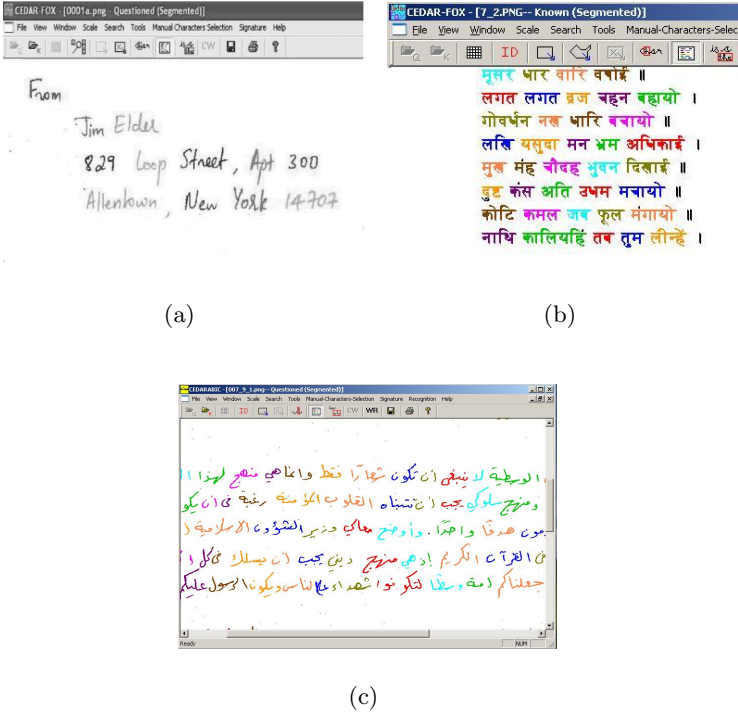
The word spotting technique involves the segmentation of each document into its corresponding lines and words. Each document is automatically indexed by the visual image features of its words. The indexing technique used was the same for all the languages evaluated (English, Sanskrit and Arabic).

### 2.1 Preprocessing

In the image preprocessing stage, the scanned image of each of the documents is subjected to several processing functions that leads to the segmentation of the



document into lines and the words in each of these lines. Figure 1 shows the line and word segmentation for handwritten English, Devanagari text, and handwritten Arabic. The line segmentation is performed using a clustering method. For word segmentation, the problem is formulated as a classification problem as to whether or not the gap between two adjacent connected components in a line is word gap or not. An artificial neural network with features characterizing the connected components was used to make a decision on this classification problem.



**Fig. 1.** Segmentation examples where segmented words are colored differently: (a) handwritten English (b) printed Sanskrit and (c) handwritten Arabic

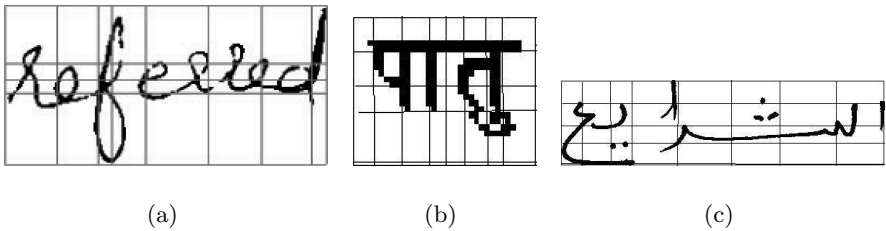
Although the microfeatures needed major revision as compared with the English-based system, the macrofeatures remained largely similar. One major exception was the distance between words. Earlier, we developed and presented [9] an automatic word segmentation, now part of the CEDARARABIC system the word distance. This algorithm takes into consideration various intricacies of handwritten Arabic and thereby achieves higher accuracy on Arabic handwritten documents than the word segmenter found in CEDARFOX. It is based on

taking several features on either side of a potential segmentation point and using a neural network for deciding whether or not the segmentation is between two distinct words. Some of the differences between the tasks of segmenting Arabic script and segmenting Latin script are the presence of multiple dots above and below the main body in Arabic and the absence of upper case letters at the beginning of sentences in Arabic. The method presented was found to have an overall correctness of about 60% when applied to the documents with correctly segmented lines using a set of seven segmentation features.

The process of word segmentation begins with obtaining the set of connected components for each line in the document image. The connected components are grouped into clusters, by merging minor components such as dots above and below a major component. Also particular to Arabic, many words start with the Arabic character “Alef”. The presence of an “Alef” is a strong indicator that there may be a word gap between the pair of clusters. The height and width of the component are two parameters used to check if the component is the character “Alef”. Every pair of adjacent clusters are candidates for word gaps. 9 features are extracted for these pairs of clusters and a neural network is used to determine if the gap between the pair is a word gap.

## 2.2 Feature Extraction

The next step in the word spotting technique involves the computation of features for each of the words identified. These features form the key index for image search. The features used here for each word image, are the gradient, structural and concavity (GSC) features which measure the image characteristics at local, intermediate, and large scales and hence approximate a heterogeneous multi-resolution paradigm to feature extraction. The features are extracted under a  $4 \times 8$  division. Figure 2 shows examples of the region division for words in all three languages.



**Fig. 2.** Examples of word images with  $4 \times 8$  division for feature extraction: (a) English (b) Sanskrit and (c) Arabic

## 2.3 Measure for Word Spotting

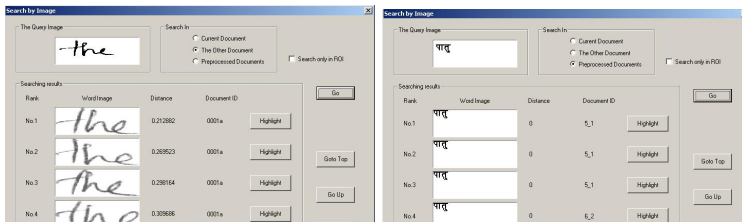
The distance between a word to be spotted and all the other words in the documents being matched against is computed using a normalized correlation

similarity measure. This similarity measure is used to measure the similarity between two word images whose shapes are represented using the 1024-bit binary feature vectors described above.

## 2.4 Word Spotting

**Image as query:** Here the query is a word image and the goal is to retrieve documents which contain this word image. For this the system supports three kinds of word-spotting searches : 1) search for all relevant words within the same document 2) search for all relevant words in another document which is currently open 3) search for all relevant words from a collection of preprocessed documents. Each of the retrieved word images is also linked with its corresponding document ID, which allows the user to easily retrieve its location and the document it belongs to. The queried word is selected as an image from the document containing it. The results returned are word images and they are ranked according to their distances to the queried image.

Figure 3 shows screen shots of the system after doing word spotting for words in English and Sanskrit.

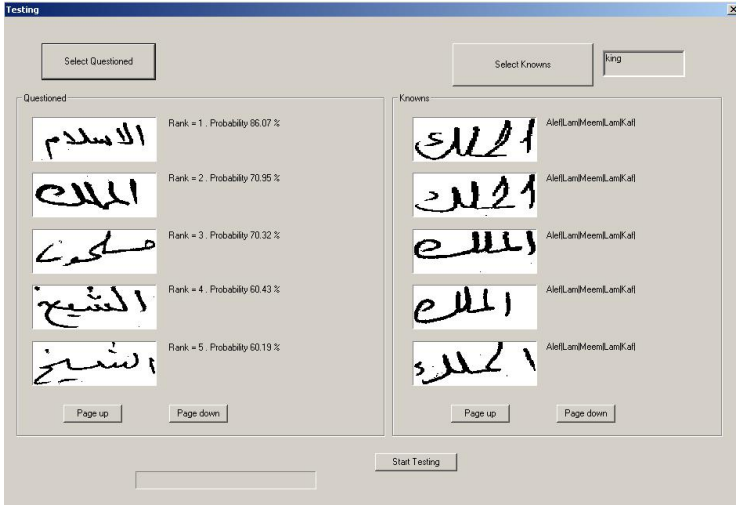


(a) English Search with an image query of the word “the”

(b) Sanskrit Search with an image query of the word “paathu”

**Fig. 3.** Screen shot of the system when performing word spotting search in (a) English and (b) Sanskrit

**Text as query:** This method of retrieval was used for spotting Arabic words, with the idea of using the English equivalent meaning of that Arabic word as the query. Here the query is text and the goal is to retrieve the documents containing the word. The search phase has two parts: (i) Prototype selection: Here the first part of the typed in query is used to generate a set of query images corresponding to handwritten versions of it. (ii) Word Matching: Here each query image is matched against each indexed image to find the closest match. The purpose of such a two step search is the following (i) The prototype selection step provides



**Fig. 4.** Arabic Search using Text Query. The prototype word images on the right side, obtained in the first step, are used to spot the images shown on the left. The results are sorted in rank by probability.

for a set of handwritten versions of the query, to account for the different ways in which the word can be written. And hence this improves the word matching capability across different styles of writing. (ii) Also it enables the query to be in text and in a different language such as English. The two step approach is described below.

1. Prototype selection: Prototypes which are handwritten samples of a word are obtained from an indexed (segmented) set of documents. These indexed documents contain the truth (English equivalent) for every word image. Synonymous words if present in the truth are also used to obtain the prototypes. Hence queries such as “country” will result in selecting prototypes that have been truthed as “country” or “nation” etc... A dynamic programming Edit Distance algorithm is used to match the query text with the indexed word image’s truth. Those with distance as zero are automatically selected as prototypes. Others can be selected manually. This step can be thought of as a query expansion step where the text query is mapped into a number of query word images.
2. Word matching: For word matching, each word image in the test set of documents is compared with every selected prototype and a distribution of similarity values is obtained. Figure 4 shows the word matching step.

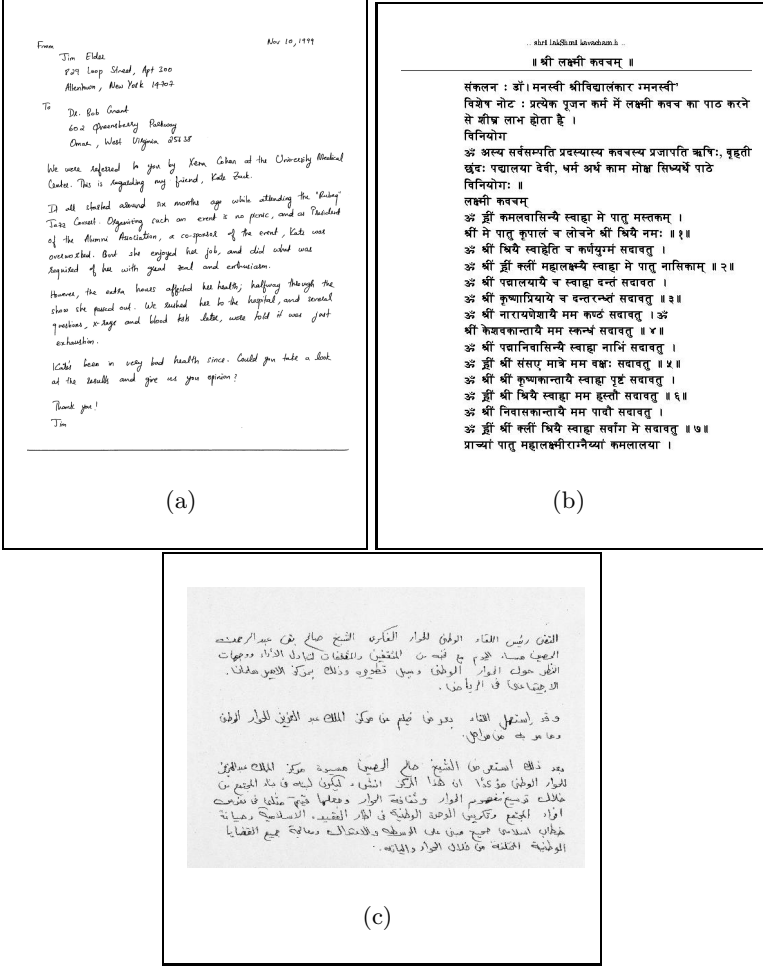


Fig. 5. Sample documents: (a) English, (b) Sanskrit and (c) Arabic

### 3 Experiments and Results

#### 3.1 Dataset

The data sets used for English are subsets of a document image collection consisting of 3000 samples written by 1000 writers, where each writer wrote 3 samples of a pre-formatted letter called “CEDAR Letter”.

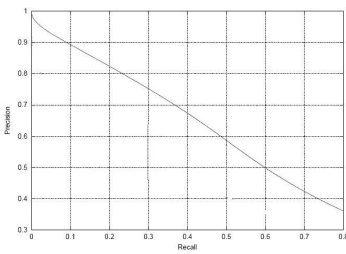
The dataset used for Sanskrit consists of 18 documents with printed text randomly selected from <http://sanskrit.gde.to/>. Each document consists of a different text and the number of words vary from 45 to 250.

A document collection was prepared with 10 different writers, each contributing 10 different full page documents in handwritten Arabic. For each of the 10 documents that were handwritten, a complete set of truth comprising of the alphabet sequence, meaning and the pronunciation of every word in that document was also given.

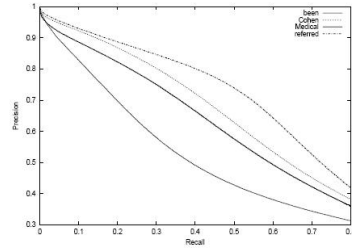
All the documents were scanned in 8-bit gray scale (in other words, 256 shades of gray) and using 300 dpi resolution. Figure 5 shows portions of three of the sample documents used. Precision-recall curves have been used to evaluate the word spotting performance.

**Table 1.** Performance of English Word Spotting

Rank of the word image Returned	Percentage correctly found ( $\frac{\text{Number of times}}{\text{total no of queries}} \times 100\%$ )
1	80 %
< 5	81 %
< 10	85 %
< 20	89 %
< 50	100 %



(a) Average Precision Recall Curve

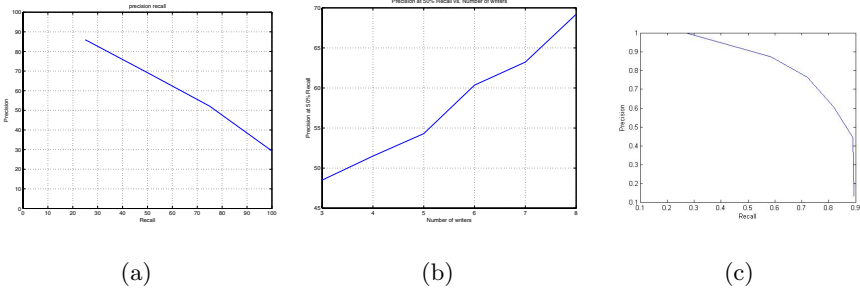


(b)

**Fig. 6.** (a) Average Precision/recall curve for English, (b) Precision/recall curve for the four English words

### 3.2 English Word Spotting

Word spotting in English was tested for by searching for several different word images. Each word image selected was searched for in another document written by the author of the queried image. A total of 100 queries from different documents were performed. Table 1 shows the evaluation results.



**Fig. 7.** (a) Retrieval of Arabic words. All results were averaged over 150 queries. The styles of 8 writers were used for training (by using their documents) to provide template word images. (b) Precision at 50% recall is plotted against different numbers of writers used for training. As more number of writers are used for training, the precision at a given recall increases. (c) Average Precision/Recall curve for Sanskrit words.

**Table 2.** Results for Word Spotting of Sanskrit words occurring 10 times

Words	Rank 1	Within Rank 2	Within Rank 5	Within Rank 10	Within Rank 20
yah	1	2	5	8	9
hoye	1	2	5	9	9
sadavatu	1	2	5	7	7
prabhu	1	2	5	8	8
tum	1	2	5	10	10

In addition, we also calculated the Precision-Recall values for four randomly selected English words which occurred atleast twice in each document, "referred", "Cohen", "been" and "Medical". Figure 6(a) shows the average Precision-Recall curve for these four words and Figure 6(b) shows the individual Precision-Recall curves for these four words.

### 3.3 Sanskrit Word Spotting

Testing for Sanskrit was done by searching for several randomly selected Sanskrit words in 18 documents. Each of the words tested occurred at least 5 times in the documents with varying frequencies. Table 2 displays retrieved results for 5 Sanskrit words which occurred 10 times in the selected documents. Figure 7(c) displays a precision recall curve of the average precision recall values of all the queried words. As always additional documents in the test set would have be good, but we were limited by time. In addition, we found the 18 documents were enough to give consistent results.

### 3.4 Arabic Word Spotting

The performance of the word spotter was evaluated using manually segmented Arabic documents. All experiments and results were averaged over 150 queries, utilizing all documents available in the CEDAR letter set. For each query, the set of documents were partitioned randomly for use as training and testing. Precision-Recall curves are shown in Figures 7(a). When the number of writers used to provide prototype styles are increased, the precision at a given recall can be seen to increase in figure 7(b). The reason for this is intuitive as more writers provide for more prototype images that correspond to learning a greater number of styles in which the query word can be written.

## 4 Conclusion

A comparison of a language independent method of word spotting was presented and evaluated on documents using Latin, Devanagari, and Arabic scripts. The performance of the system in these various types of searches has been presented. A high precision can be obtained for words from printed documents. A two step approach involving a query expansion step yields promising results for spotting words in handwritten documents.

## References

1. Kuo, S., Agazzi, O.: Keyword spotting in poorly printed documents using 2-d hidden markov models. *IEEE Trans. Pattern Analysis and Machine Intelligence* 16, 842–848 (1994)
2. Burl, M., Perona, P.: Using hierarchical shape models to spot keywords in cursive handwriting. In: *IEEE-CS Conference on Computer Vision and Pattern Recognition*, June 23–28, pp. 535–540 (1998)
3. Kolz, A., Alspector, J., Augusteijn, M., Carlson, R., Popescu, G.V.: A line-oriented approach to word spotting in handwritten documents. *Pattern Analysis and Applications* 2(3), 153–168 (2000)
4. Manmatha, R., Rath, T.M.: Indexing of handwritten historical documents-recent progress. In: *Symposium on Document Image Understanding Technology (SDIUT)*, pp. 77–85 (2003)
5. Zhang, B., Srihari, S.N., Huang, C.: Word image retrieval using binary features. In: *Document Recognition and Retrieval XI, SPIE, San Jose, CA* (2004)
6. Srihari, S.N., Srinivasan, H., Babu, P., Bhole, C.: Handwritten arabic word spotting using the cedarabic document analysis system. In: *Proc. Symposium on Document Image Understanding Technology (SDIUT 2005)*, pp. 123–132 (2005)
7. Srihari, S.N., Cha, S.-H., Arora, H., Lee, S.: Individuality of handwriting. *Journal of Forensic Sciences* 47(4), 856–872 (2002)
8. Srihari, S.N., Zhang, B., Tomai, C., Lee, S., Shi, Z., Shin, Y.C.: A system for handwriting matching and recognition. In: *Proceedings of the Symposium on Document Image Understanding Technology (SDIUT 2003)*, Greenbelt, MD (2003)
9. Srihari, S.N., Srinivasan, H., Babu, P., Bhole, C.: Spotting words in handwritten Arabic documents. In: *Document Recognition and Retrieval XIII: Proceedings SPIE, San Jose, CA*, pp. 606702–1–606702–12(2006)



# Focused Page Rank in Scientific Papers Ranking

Mikalai Krapivin and Maurizio Marchese

Dipartimento di Ingegneria e Scienza dell' Informazione (DISI), University of Trento, Italy  
krapivin@disi.unitn.it, marchese@disi.unitn.it

We propose Focused Page Rank (FPR) algorithm adaptation for the problem of scientific papers ranking. FPR is based on the Focused Surfer model, where the probability to follow the reference in a paper is proportional to its citation count. Evaluation on Citeseer autonomous digital library content showed that proposed model is a tradeoff between traditional citation count and basic Page Rank (PR). In contrast to basic Page Rank, proposed Focused Surfer model suffers less from the "outbound links" problem. We believe that FPR algorithm is closer to reality because highly cited papers are more visible and tend to attract more citations in future. This is in accordance with the one of the most significant principles of Scientometrics. No need for lexical analysis of the domain corpus and simplicity of implementation are among the strong points of the proposed model and make the proposed ranking technique attractive for academia digital libraries.

**Keywords:** Scientometrics, Page Rank, Focused Surfer, Citation-based metrics, Digital Libraries.

## 1 Introduction

Ten years ago Google corporation applied Page Rank (PR) algorithm [1] with great success to the problem of web-pages ranking. PR algorithm is purely statistical, and there is no need to analyze the content of each page lexically. It uses a "*Random Surfer*" model [1] in which the process of browsing through the web pages links is modeled by the stochastic Markov process, fully described by a Markov chain matrix. Recently Page Rank has been studied from several points of view including computational feasibility, modifications and adaptations to the different types of graphs and network models, probabilistic model, mathematical background [2]. Its popularity for ranking web-pages makes it popular in other domains, like ranking of scholarly publications.

The most intriguing question about PR is how to compute it for the whole web? Whole internet contains terabytes of information, and being represented as a graph it exceeds modern computers' memory. It is a creative engineering task to design fast access storage to compute PR. Let us briefly outline major methods for PR computation.

- 1) The simplest one is the cyclic PR computation for all nodes – one by one - in the graph, using recursive formula (1) until convergence [3]. This method takes unit vector as initial rank approximation.
- 2) PR authors, Brin and Page proposed polynomial convergence method [1], similar to Jacobi methods.

- 3) Method (2) was improved by Haveliwala in 1999 [4] using "block-based strategy", similar to implementations in relational database products.
- 4) In 2003 Langille [5] invented the procedure with reduction of the iterations number with lucky initial approximation.
- 5) In 2003 Kamvar et al.[6], proposed quadratic extrapolation method to accelerate PR convergence and evaluated their methodology under roughly 81 millions of pages.

Most of mentioned above works are related to the Web links ranking problem which usually deals with much larger graphs than scientific citing problem. So, the computation problem has been studied well enough and looks feasible.

A correlated research topic is related to promising PageRank modifications, for instance:

- 1) PR Computation with or without damp factor (see formula (2) below).
- 2) Personalized Page Rank with some initial personalization vector is more common for web-search engines. Here all pages have their own personal weights *before* PR calculation.
- 3) Focusing of PR, or redistribution of links to link probabilities in the stochastic Markov matrix. This means that core PR model of *Random Surfer* is no longer Random, it becomes focused. This model was successfully applied to the web pages ranking problem by Tony Abou-Assaleh *et al.* [7] and by Fuyong Yuan *et al.*[8] in 2007.
- 4) Double (or more) focusing of PR takes into account more deep properties of citation graph entities during stochastic Markov matrix composition. For example, it may first focus on site name and then on site content.

Ranking problem is also very important in the scholarly domain, where the main metrics of an article's contribution is the *citations count* [9]. Recently Chen and others [3] applied Page Rank idea for the scientific citations. The major result of this application is that some classical articles in Physics domain have small quantity of citations and very high Page Rank. Chen *et al.* called them "*scientific gems*". Existence of "*scientific gems*" is caused by PR model which captures not only the total citation count, but the rank of each of the citing papers.

Another Page Rank adaptation for the same problem was performed by Yan Sun *et al.*, 2007 [10]. They applied "personalization" modification from above, where personalized vector was taken in proportion to the publishing journals weight. Then the validity of the rank was estimated by the *cumulative gain* function [10].

Recently Page Rank was successfully applied to the problem of assessing papers, institutions, authors for really large scale problem (~billion of items) [12]. Both methods of assessing academia papers – the traditional citation count and the more recent PageRank and are based on the quantity of citations. Citation count advantages are I) simplicity of computation; II) it is a proven method which has been used for many years in scientometrics. Proven history of use is very important in the conservative academia domain. Page Rank has the following strong sides: I) it statistically analyses whole citations graph at once; II) it captures not just quantity, but also *quality* of citing papers. However, Page Rank algorithm introduces also computational artifacts like the "effect of outbound links" [13]: this means that if a paper *P* is cited many

times by papers with high rank but containing a large quantity of outgoing links — it may decrease  $P$ 's rank. Situation when a paper is highly cited but poorly ranked by PR looks strange for academia publications.

In this paper, we propose Focused Page Rank adaptation to reduce the "effect of outbound links" and to make a tradeoff between Page Rank and Citation Count. In our proposed model a "reader of an article"<sup>1</sup> may follow all references with different probabilities, so our random surfer model is getting focused. We take citation count as a measure of attractiveness of a reference inside a scientific paper.

## 2 Problem Statement

Let us briefly outline what the original Page Rank algorithm does. It performs ranking for the nodes of the oriented graph with  $N$  vertices. There are two different link types which may connect node to the neighbors: outbound links and inbound ones. The main measure of node's weight is inbound links quantity. When we apply this model to the scientific citations problem, we can establish the following similarities: papers are nodes of the graph; citations made by the other papers are inbound links; "references" section creates outbound links set. This is true for most of scientific papers. Rank of a node according to PR is given by the recursion formula (1):

$$P_i = \sum_{\substack{j \in D \\ i \neq j}} \frac{P_j}{S(j)}, \quad (1)$$

where  $S(j)$  is the quantity of references for paper  $P_j$ ,  $i, j \in \{1, \dots, n\}$  are paper sequence numbers in a graph and  $D_i$  is variety of all articles which cite article  $i$ . In the matrix form we can rewrite it as eigenvector problem (2):

$$\vec{r} = A \cdot \vec{r}, \quad (2)$$

where  $A$  is the transition matrix, or stochastic Markov matrix. This consideration exposes several potential problems in rank computation as discussed in [2],[5]. One of them is the presence of the papers which cite other papers but are not cited themselves. They are called dangling nodes and they may be treated as the most recent papers. In this case equation (2) may have no unique solution, or it may have no solution at all. It will lead to zero-rows occurrence in the transition matrix and uncertainty of the rank of dangling nodes. Such problem may be resolved with the introduction of a damp-factor  $d$ . The damp (or decay) factor is a positive number  $d$ , such that  $0 < d < 1$  and we illustrate it in formula (3):

$$P_i = (1 - d) \cdot \sum_{\substack{j \in D \\ i \neq j}} \frac{P_j}{S(j)} + \frac{d}{N} \quad (3)$$

Damp factor was proposed by PR inventors Page and Brin and widely used in different Page Rank computations. It helps to achieve two goals at once: 1) faster convergence using iterative computational methods; 2) problem becomes solvable for sure since all nodes have a possibility to be visited by a Random Surfer.

---

<sup>1</sup> "Reader of an article" is a Focused Surfer.

## 2.1 Scientific Citations Graph Specific Characteristics

When considering the scientific citation problem we may avoid the mentioned above problems in a very natural way because of the following peculiarities of our specific domain (i.e. scientific papers):

- I) After an article is published, it cannot cite anymore.
- II) If the number of articles in the graph is  $N$ , each paper may potentially have from  $1$  to  $N-1$  ingoing links and the same quantity of outgoing ones. Since  $N \gg 1$ , in real life the citation graph is extremely sparse. Indeed, articles normally have from 5 to 20 citations inside, comparing average quantity of citations per article  $m$  with quantity of papers in graph  $N$  it is obvious that  $m \ll N$ .

First condition simplifies highly the problem because citations graph becomes unidirectional. We assume (and experimentally prove) that citation graph is free of loops, cliques or some other complex structures.

Situation with a loop when paper  $A$  cites paper  $B$  and paper  $B$  cites  $A$  is theoretically possible, for example if authors exchange their deliverables and cite not yet published but already accepted for publication papers. However, according to Glänzel [9] traditional scientometrics does not consider such citations as the valid ones.

## 2.2 Focused Surfer

The Random Surfer model is the basis of PR algorithm. Page Rank of the certain node is proportional to the probability to reach this node by randomly riding the graph. At each step rider randomly chooses the link to follow. Focused Surfer decides which path is more preferable for him. Formula (4) expresses this mathematically:

$$P_i = (1 - d) \cdot \sum_{\substack{j \in D \\ i \neq j}} P_j \cdot s(j | i) + \frac{d}{N}, \quad (4)$$

where  $s(j | i)$  is the probability to follow the reference  $i$  being at the place  $j$ .  $s$  is a function that may be arbitrary. We propose to use the simplest variant of it, which we show in formula (5):

$$s(j | i) = \frac{C(i)}{\sum_{k \in D} C(k)}, \quad (5)$$

where  $C(m)$  is paper  $m$  citations count, and  $D$  is the set of all references in paper  $C(j)$ . This means that more cited nodes have advantage and they are more visible and attractive for further citation.

## 3 Evaluation and Experimental Methodology

In our evaluation, we explore 266788 papers published in ACM conferences or journals starting from 1950 and till 2007 with the majority of papers around 2002-2005.

This dataset may be completely matched to ACM portal<sup>2</sup> and was crawled by the Citeseer<sup>3</sup> digital library.

### 3.1 Plotting the Difference

We introduce here our proposed experimental methodology. The obvious approach to exploring the effect of using PR vs citation count (CC) in evaluating papers is to plot these values for the different papers. The density of points (points cloud) that have a high CC and low PR (or vice versa) would provide an indication of how often these measures can give different quality indication for a paper. However, this leads to charts difficult to read in many ways: First, points overlap because many papers have the same CC, or the same PR, or both. Second, it is hard to get a qualitative indication of what is “high” and “low” for CC or PR. This is why we divide CC and PR axis in bands.

Ideally we would have to split the axes into 10 (or 100) bands. We put in the first band the top 10% (top 1%) of the papers based on the metric, to give qualitative indications so that the presence of many papers in the corners of the chart would denote a high divergence. However, the overlap problem would remain, and it would distort the charts in a significant way since the measures are discrete. For example, the number of papers with 0 citations is well above 10%. If we neglect this issue and still divide in bands of equal size (number of papers), papers with the same measure would end up in different bands.

Finally, the approach we took (Fig. 1, Fig. 2) is to divide the X-axis in bands where each band corresponds to a different - discrete - citation count. With this separation we built 290 different bands, since there are 290 different values for CC (even if there are papers with much higher CC, there are only 290 different CC values in the set). For the Y-axis we leverage mirrored banding, i.e., the Y-axis is divided into as many bands as the X-axis, also in growing values of PR. Each Y band contains the same number of papers as X. In other words, the vertical rectangle corresponding to band  $i$  in the X axis contains the same number of papers  $q_i$  as the horizontal rectangle corresponding to band  $i$  of the Y-axis. We call a point in this chart as a square, and each square can contain zero, one, or many papers (not ranks, because the zone number represents the actual PR or CC).

The reasoning behind the use of mirrored banding is that this chart emphasizes divergence as distance from the diagonal. At an extreme, plotting a metric against itself with mirrored banding would only put papers in the diagonal. Since the overlap in PR values is minimal (there are thousands of different values of PR and very few papers with the same PR values, most of which having very low CC and very low PR, and hence uninteresting), it does not affect in any qualitatively meaningful way the banding of the Y-axis. To realize what are the real value of PR and CC is behind of each zone please take a look at the Table 1.

---

<sup>2</sup> <http://portal.acm.org/>

<sup>3</sup> <http://citeseer.ist.psu.edu/>

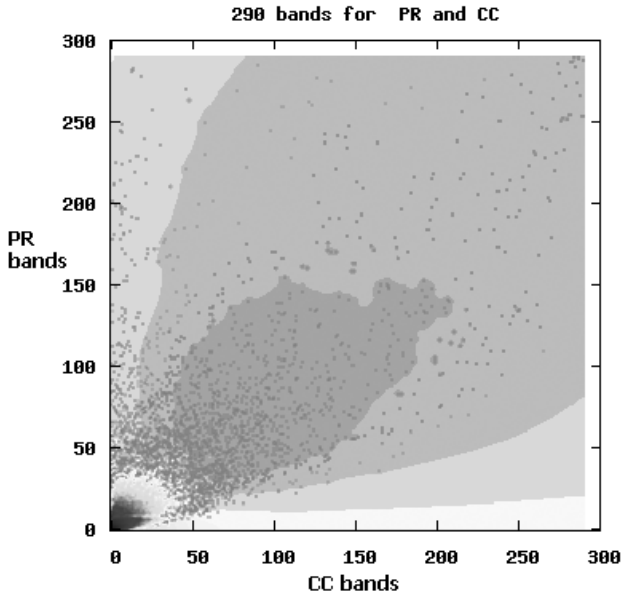
**Table 1.** The mapping between real CC and PR and the band number

Number of band both for CC and PR	CC	PR
50	50	6.23
100	100	14.74
150	151	26.57
200	213	38.82
250	326	58.86
280	632	113.09
290	1736	224.12

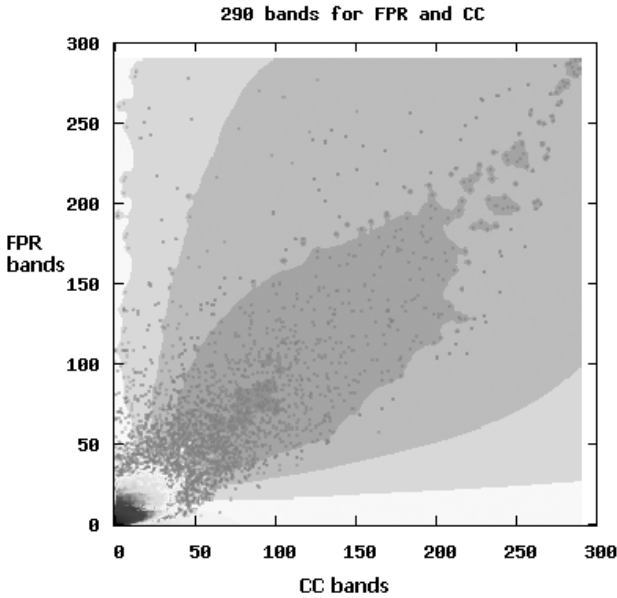
### 3.2 Evaluation

The described analysis and visualization methodology gives the overall picture for all 266788 papers on one chart (Fig. 1). The points are strongly biased around the main diagonal. This biasing shows the *diversity*, or difference between PR and CC. There are some papers with extremely low citation count but very significant Page Rank, or “scientific gems” following Chen *et al.* [3]. They are the papers cited by several heavily cited papers. Being cited by just *one* extremely high ranked paper may be enough to improve PageRank drastically. Fig. 3 represents a piece of full citation graph, where there is a real paper with  $PR \gg CC$  and just 14 citations from other papers in the graph.

In contrast to “scientific gems” there are some other papers below the main diagonal, located in the bottom-left part of Fig. 1 and Fig. 2, when CC band is greater



**Fig. 1.** Diversity of Page Rank (PR) and Citation Count (CC). White and black points in the bottom-left corner does not mean absence of papers. This is a grayscale of colored map, where the major quantity of papers has small number of CC, and since lie exactly in the bottom-left corner and it is nearly the same for the both plots. The plot is mirror-like banded.



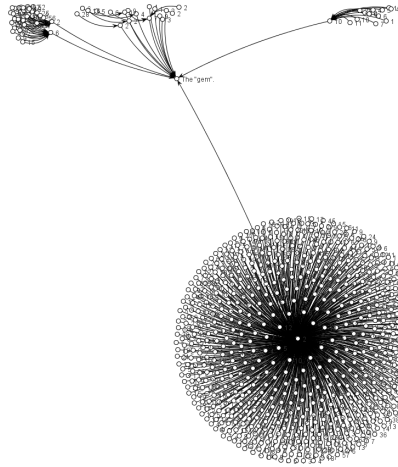
**Fig. 2.** Diversity of Focused Page Rank (FPR) and Citation Count (CC). Bottom-left corner distribution explained on the Fig. 1 description. Again the plot is mirror-like banded.

than 50. Papers in that region have significantly high CC and small PageRank. This is caused by “outgoing links effect”. To understand the nature of this effect let us see the formula (1). Denominator  $S$  in (1) represents the probability to follow the link, and being a big number it reduces the Page Rank of a paper. This denominator  $S$  is the corner-stone of Random Surfer model, and it reflects the fact that all references are completely equal from probabilistic point of view. Thus if a paper is cited many times by papers with large quantity of *outgoing links* (papers with long “references” section) it may have much lower PageRank than the other papers with the same citation count. The example of such a paper is plotted in the middle of Fig. 4, this paper has 55 citations and more than 100 times lower PR than “scientific gem” plotted in Fig. 3.

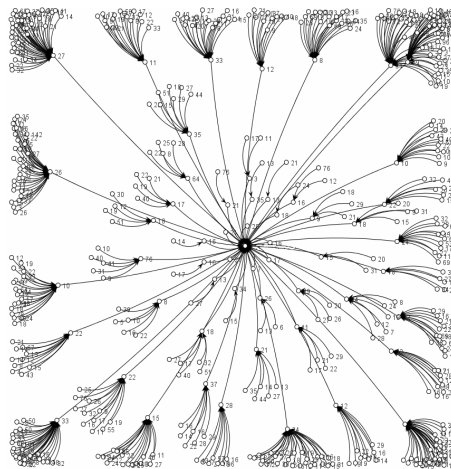
Fig. 2 illustrates the Focused Surfer model and FPR algorithm instead of PR. Focused Surfer model gives better chances to more cited papers, at the same time stealing the part of the weight from their poorly cited neighbors. This idea leads us to the conclusion that in general total FPR rank remains the same as PR, it just gets *re-distributed*. This idea is supported by computation of average FPR and PR which are nearly the same:  $\langle FPR \rangle = 0.603$  and  $\langle PR \rangle = 0.602$ .

Now let us observe effects present on Fig. 2. The points are located closer to the main diagonal<sup>4</sup> (comparing with Fig. 1) and there is significantly less papers with big CC and small PR (reducing of the effect of outbound links). On the other hand we see that “gems”-effect is still noticeable.

<sup>4</sup> If all the points lie exactly on the main diagonal we would have 100% match of CC and PR.



**Fig. 3** “Scientific gem” in the center. Cited by heavily cited paper (in the bottom).

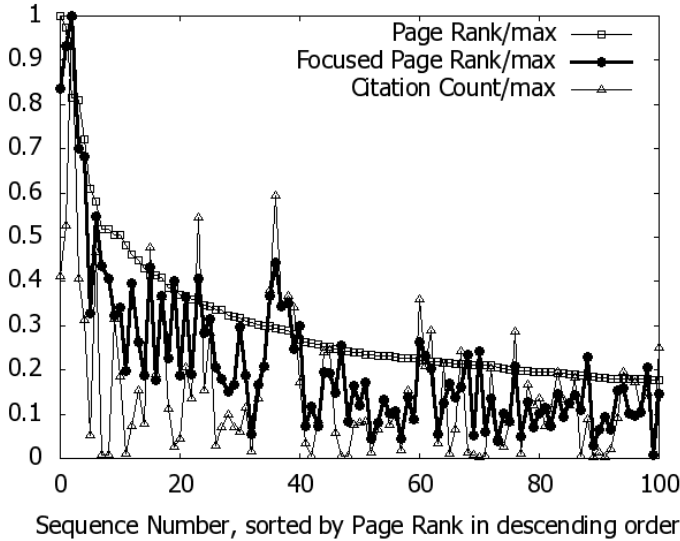


**Fig. 4.** The opposite to “scientific gem” paper (in the middle)

This means that FPR tends to reduce “outgoing links” effect and tends to make FPR closer to the CC. Or in other words it tends to shift the points in Fig. 2 towards the main diagonal. This effect of *shifting* the points towards the main diagonal may be numerically evaluated. We compute the difference  $\Delta_{gems} = PR - FPR$  for each *square* in the “gems zone”, where CC band < 10 and PR band > 10. Then we do the same for the opposite “popular papers” zone where CC band > 10 and PR band < 10. It would be  $\Delta_{popular\ papers} = PR - FPR$ . We notice that  $\Delta_{gems} = 3 \Delta_{popular\ papers}$ , which means that focusing eliminates “popular papers” 3 times greater than “gems”. So Focused Page Rank tends to keep “gems” while correcting “popular papers” ranks.



The last plot in Fig. 5 shows the top 100 papers with the biggest CC. There are 3 curves there: PR, CC and FPR. It is clear from Fig. 5 that FPR is a tradeoff between PR and CC in highly cited region.



**Fig. 5.** Top 100 papers with the highest CC. Bold line is the Focused Page Rank. All ranks are normalized by their maximum value, and thus comparable.

## 4 Conclusion

Focused Page Rank has been proposed for the problem of scientific citing. Our major strong points are:

1. *It is the tradeoff between Page Rank and Citation Count.* So it may serve as an agreement between the followers of pure citation count and Page Rank followers.
2. The proposed Focused Page Rank suffers less from the effect of outbound links. Therefore, it is capable to better capture one of the fundamental principles of Scientometrics, first time formulated by de Solla Price in 1976 [11]:

*“Success seems to breed success. A paper which has been cited many times is more likely to be cited again than one which has been little cited. An author of many papers is more likely to publish again than one who has been less prolific. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use”.*

3. It captures the power of Page Rank, where not only the quantity of citations, but also the quality of ones counts.

## Acknowledgements

Author would like to thank Prof. Fabio Casati, Andrei Yandratsau and Alexander Autayeu for useful discussions, and Prof. Lee Giles for providing high quality dataset.

## References

1. Page, L., Brin, S.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the Seventh International Web Conference, pp. 107–117 (1998)
2. Diligenti, M., Gori, M., Maggini, M.: Web Page Scoring Systems for Horizontal and Vertical Search. In: WWW 2002, pp. 84–89. ACM Press, New York (2002)
3. Chen, P., Xie, H., Maslov, S., Redner, S.: Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics* 1(1), 8–15 (2007)
4. Haveliwala, T.: Efficient Computation of PageRank. Technical report, pp. 84–89 (1999), <http://dbpubs.stanford.edu/pub/1999-31>
5. Langville, A.N., Meyer, C.D.: Deeper Inside PageRank. *J. Internet Mathematics* 15(5), 335–380 (2004)
6. Kamvar, S., Haveliwala, T., Manning, C., Golub, G.: Extrapolation Methods for Accelerating PageRank Computations. In: WWW 2003. ACM, New York (2003)
7. Abou-Assaleh, T., Das, T., Weizheng, G., Yingbo, M., O'Brien, P., Zhen, Z.: A Link-Based Ranking Scheme for Focused Search. In: WWW 2007. ACM Press, New York (2007)
8. Fuyong, Y., Chunxia, Y., Jian, L.: WImprovement of PageRank for Focused Crawler. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, pp. 797–802 (2007)
9. Glänzel, W.: Bibliometrics as a research field, A course on theory and application of bibliometric indicators, Magyar Tudományos Akadémia, Course Handouts (2003), [http://www.norslis.net/2004/Bib\\_Module\\_KUL.pdf](http://www.norslis.net/2004/Bib_Module_KUL.pdf)
10. Sun, Y., Giles, C.L.: Popularity Weighted Ranking for Academic Digital Libraries (2007)
11. de Solla Price, D.J.: Little Science - Big Science. Columbia Univ. Press, New York (1963)
12. Bollen, J., Van de Sompel, H., Balakireva, L., Chute, R.: A ranking and exploration service based on large-scale usage data. In: JCDL, ACM/IEEE, poster. IEEE Computer Society Press, Los Alamitos (2008)
13. Sobek, M.: The effect of outbound links (2003), <http://pr.efactory.de/e-outbound-links.shtml>

# Scientific Journals, Overlays and Repositories: A Case of Costs and Sustainability Issues

Panayiota Polydoratou and Martin Moyle

University College London,  
Library Services,  
DMS Watson Building, Malet Place, WC1E 6BT  
lib-rioja@ucl.ac.uk

**Abstract.** Publishing scientific research is an area of study that attracts interest from various stakeholders such as publishers, academic and research staff, libraries and funders. In the past decade increased journal subscription fees prompted calls for cheaper and more efficient means of accessing the scientific literature. Factors such as the expansion of digital repositories, the introduction of open source journal management software, an increasing awareness within the scholarly community at large of the issues around open access, and an increasing readiness within the publishing community to experiment with new models, suggest that the circumstances may now be right for new models of scientific publishing to be explored, as well as potential business models and sustainable solutions around them. This paper explores some of the issues around the costs and sustainability of a prospective journal model known as the overlay journal. We present estimates of initial start up costs for such a model, discuss the factors that would influence scientists in deciding whether to publish in a journal overlaid onto a public repository; and report their views on the relative importance of different features and functions of a journal in terms of funding priorities.

## 1 Introduction

This paper presents findings from the JISC (Joint Information Systems Committee, UK) funded RIOJA (Repository Interface for Overlaid Journal Archives) RIOJA (Repository Interface for Overlaid Journal Archives) project (<http://www.ucl.ac.uk/lis/rioja>) which aimed to address the issues around the development and implementation of a new publishing model, the overlay journal. For the purposes of this paper, an overlay journal is defined as an open access journal whose content is deposited to and resides in one or more open access repositories. It utilizes quality certification, is sustainable and adheres to preservation standards.

The impetus for the RIOJA project came directly from academic users of the arXiv (<http://arxiv.org>) subject repository. For this reason, arXiv and its community is the testbed for RIOJA. arXiv was founded in 1991 to facilitate the exchange of pre-prints between physicists. It now holds over 495,000 scientific papers, and in recent years its coverage has extended to mathematics, nonlinear sciences, quantitative biology and

computer science in addition to physics. arXiv is firmly embedded in the research workflows of these communities.

## 2 Background and Statement of the Problem

The overlay concept, and the term "overlay journal" itself, appear to be attributed to Ginsparg (1996). Significant contribution to the concept of overlay journals has been conducted by J W T Smith (1999) who discussed and compared functions of the existing publishing model what he referred to as the "deconstructed journal".

Halliday and Oppenheim (1999), in a report regarding the economics of Digital Libraries, recommended further research, in the field of electronic publishing in particular. Specifically, they suggested that the costs of electronic journal services should be further investigated, and commented that the degree of functionality that users require from electronic journals may have an impact on their costs. In a JISC funded report, consultants from Rightscom Ltd (2005) suggested that commercial arrangements for the provision of access to the published literature are made based on the nature of the resource and the anticipated usage of the resource. Cockerill (2006) indicated that what is regarded as a sustainable publishing model in the traditional sense (pay for access) is actually supported by the willingness of libraries to pay [... "even reluctantly", p.94] large amounts of money to ensure access to the published literature. He suggested that as open access does not introduce any new costs there should not be any difficulty, in theory, in sustaining open access to the literature. Waltham (2005) raised further questions about the role of learned societies as publishers as well as the overall acceptance of the 'author pays' model by the scientific community.

Self-archiving and open access journals have been recommended by the Budapest Open Access Initiative (<http://www.soros.org/openaccess/read.shtml>) as the means to achieve access to publicly-funded research. The overlay model has the potential to combine both these "Green" (self-archiving) and "Gold" (open access journal) roads to open access. Hagemmann (2006) noted that "...*overlay journals complement the original BOAI dual strategy for achieving Open Access...*" and suggested that the overlay model could be the next step to open access. In support of open access to information the BOAI published guides and handbooks on best practice to launching a new open access journal, converting an existing journal to open access, and business models to take into consideration [Crow & Goldstein, 2003a-c).

Factors such as the expansion of digital repositories, the introduction of open source journal management software, an increasing awareness within the scholarly community at large of the issues around open access, and an increasing readiness within the publishing community to experiment with new models, suggest that the circumstances may now be right for an overlay model to succeed. Part of the RIOJA project was to test the reaction of one research community, selected for its close integration with a central subject repository, to this prospective new model.

### *arXiv and the publishing process*

Despite the everyday importance of arXiv to researchers, depositing papers to the repository remains a supplement to the traditional publishing process, rather than a replacement for it. Peer review is as important to arXiv-depositing researchers as to

those in other scientific disciplines, and, to achieve peer acceptance, papers continue to be submitted for publication in the traditional way. Once a paper is accepted for publication, an author will typically update the corresponding arXiv version to denote the publishing journal title and the date of acceptance. These annotations, indicating acceptance for publication, serve as badges of quality for arXiv deposits.

Prosser (2005), quoting GermanMan (199x) who must get credit, notes that journals are traditionally held to perform four "first order" functions:

- Registration: an author wishes to be acknowledged as the person who carried out a specific piece of research and made a specific discovery
- Certification: the author's claims are tested through independent peer review, and it is determined that they are reasonable
- Awareness: the research is communicated to the author's peer group
- Archiving: the research is retained for posterity

It is clear that arXiv already provides three of these functions:

*Registration* occurs when a research paper is received by arXiv, at which point it is assigned a unique identifier and date stamp. It is commonplace for papers to be cited thereafter by arXiv reference number, illustrating the acceptance of the arXiv registration process.

Once registered, a paper can appear in the public domain on the same day. It is openly and freely available, without barriers to access. arXiv also offers email alerting to new papers and is compliant with OAI-PMH. It fulfils the *Awareness* function: many researchers clearly consult the repository in preference to traditional journals.

arXiv also satisfies the *Archiving* function, with an emphasis on stable and portable formats at ingest, and the retention for public scrutiny of version-controlled superseded papers alongside the most recent update.

arXiv, therefore, provides three of the four "first order" functions of the traditional journal. It does not yet provide *Certification*. To achieve a quality stamp, researchers from arXiv's subject communities and their institutions must engage with the full, protracted and costly machinery of formal publication. This can involve delays, page charges, author/funder charges, restrictive copyright transfer agreements, version control issues between the arXiv holdings of a paper and its published counterpart, and post-publication barriers to access because of subscription and licensing arrangements; and yet the content of the resulting journal productions has often already been disseminated via arXiv and consumed by researchers. During the development and implementation of the RIOJA tool (see below) we were able to estimate some initial start up costs which alongside the surveys' findings allowed to draw some cost projections for the overlay journal model.

### 3 Methodology

This paper builds on the results from two community surveys which were undertaken to explore the views of scientists in the fields of astrophysics and cosmology concerning the feasibility of an overlay journal model. The community surveys comprised of:

- An online questionnaire survey targeting more than 4000 scientists from the top 100 universities and 15 non academic institutions in science (yielded response by 683 scientists, 17% response rate),
- Interviews with publishers and members of editorial boards of peer-reviewed journals. These complementary studies were intended to enable a more rounded understanding of the publishing process, and to help the project to explore whether an overlay journal model in astrophysics and cosmology could be viable in the long term<sup>1</sup>.

In addition, the authors undertook desktop research to identify studies on the costs of publishing scientific journals and to compare, where possible, their findings against the development and implementation of the RIOJA toolkit described below.

## 4 The RIOJA Toolkit

The technical part of the project dealt with the development of XML-based APIs for the exchange of data between digital repositories and journals to facilitate the overlay of academic journals onto separate digital repositories. It was assumed that: a) the repository provides the registration, awareness and archiving functions of a journal and b) the journal provides only the certification (peer review) and additional awareness functions. All versions of a paper are stored in the repository, from the original submission to the published version and beyond. The repository can tag papers with their status, so end users can, if desired, filter papers to see only submitted, accepted or published papers as they prefer. The journal tracks different versions of the repository paper, and applies its final "published" quality stamp to one particular "final" paper version. The repository may, however, allow updates to a paper after publication, allowing easy access to a corrected version as well as the "published" version. The APIs are implemented in the RIOJA project's test bed, and (partially) in the arXiv subject repository (Lewis, 2007).

### 4.1 Initial Start Up Costs

The RIOJA toolkit saw the development of a module specification to support automated interactions with repositories. In full, the technical work comprised:

- Development of open API for communication between repositories and journals
- Development of software for hosting overlaid journals using the API
- Demonstration journal software, using the RIOJA API implemented on arXiv.org repository
- Version of ePrints repository software to incorporate RIOJA APIs for application in any subject area (N.B. still in progress)

---

<sup>1</sup> Published results from the RIOJA project community surveys can be found at <http://www.ucl.ac.uk/ls/rioja/dissem/>

Start-up costs included the fee to the company to which the development was outsourced. Overall, initial development and implementation costs, excluding person power, did not exceed £5000 (\$7500)<sup>2</sup>. Indicative amounts are listed below:

- Software development costs ~ 4000 (\$6500)
- OJS developed LaTeX plugin ~200 (\$400)
- Web hosting ~200/per year (\$400)

## 4.2 Fixed and Variable Costs

The term fixed costs is used in the literature to identify those costs associated with the publishing process that remain the same regardless of circulation (King, 2007; SQW Limited 2004). By contrast, variable costs refer to those that change with the number of subscriptions (e.g. cost of reproduction, subscription maintenance, etc.). Some of these costs are associated with particular business models (e.g. subscription based model) and publishing media (e.g. print versus electronic) which raises the question as to whether a cost recovery model such as the “author pays” could be cheaper to sustain in an electronic environment and using the overlay journal as a model.

Some of the costs referring to the registration stage concern submission. Costs at submission level include both rejected and accepted papers, are in general fixed costs and include what is usually addressed as first copy costs. Those include costs linked to article processing such as the work of the editor and editorial board, system support (administrative and managerial aspects, the organisation of the peer review process, staff involved in the system, etc). In those costs should be included those that refer to non-article processing. The average cost of first copy production varies widely in different sciences. King (2007) presents findings from previously reported first copy costs, ranging from \$450 to \$2500 for article processing and reaching to \$10000 in some disciplines. Consultants in SQW Limited (2004) reported that first copy costs for a good to high quality journal are estimated at around \$1500 (\$1650 including first copy and fixed costs). However, distribution costs do not vary with the number of subscriptions and are in the majority fixed rather than variable. Furthermore, it is even easier to separate and control submission costs if a submission fee and a publication fee is set separately.

## 4.3 Community Uptake

The community surveys received responses from 683 scientists (17% of 4012 contacted), and representatives from publishing houses and members of editorial boards from peer-reviewed journals in astrophysics and cosmology. Results indicated that more than half of the respondents (53%) were favourably disposed to the idea of overlay journal as a potential future model for scientific publishing. Over three quarters (80%) of the respondents were, in principle, willing to act as referees in an arXiv-overlay journal.

The most important factors which would encourage publication in a repository-overlaid journal were the quality of other submitted papers (526 responses), the transparency of the peer review process (410) and the reputation of the editorial board (386). Respondents also provided a range of other factors that they considered important, among them the reputation of the journal; its competitiveness measured against

<sup>2</sup> Exchange rate of 1 GBP = 1.98 USD (13/08/2008)

other journals under the RAE (the UK's Research Assessment Exercise); the quality both of the journal's referees and of its accepted papers; a commitment to using free software; a commitment to the long-term archiving and preservation of published papers; relevant readership; and its impact factor, (which, it was noted, should only take into account citations to papers after final acceptance and not while residing on arXiv prior to "publication").

The interviews with publishers and editors did not reveal any substantial information about costings that have not already been reported in the literature (King, 2007, SQW Limited, 2004; Waltham, 2004) or are available on some publishers' websites, e.g. PhysMath Central (<http://www.biomedcentral.com/info/about/apcfaq>). Interviewees suggested that the processing price per article varies by journal, discipline and usage. However, it was noted that community uptake and in particular the interest of academic and research staff in new publishing models is the prime driver for their adapting to technology challenges. For example, one of the publishers interviewed stated that one of their most successful journals, both in terms of revenue to the publisher and in terms of perceived quality and acceptance by the scientific community, was converted to open access (the 'author pays' model) purely because of community demand.

#### 4.4 Journal Functions

Meanwhile, a question included in the questionnaire survey concerning how expenditure should be apportioned towards particular functions of a journal was subject to criticism: respondents queried whether a scientist has adequate knowledge of the publishing process and its associated costs to make any useful observations. It was also observed that the publishing process entails more than the distribution phase, which some respondents felt that the survey, and by implication the overlay model, appeared only to address. However, the costs associated with the work of scientific editors, with the integrity and long-term archiving of journal content, and with the transparency of peer review were highlighted as worthwhile (Table 1, scale 1 (little) – 5 (most of the amount) ). An indicative comment is reproduced below:

**Table 1.** Suggested expenditure/priorities

Suggested expenditure/priority	None	1	2	3	4	5	Not sure
Paying scientific editors	23	23	60	240	141	15	21
Paying copy editors	8	28	73	256	134	6	15
Maintenance of journal software	4	20	73	238	147	9	30
Journal website	5	28	79	225	149	20	15
Online archive of journal's own back issues	9	27	52	202	189	18	19
Production of paper version	138	101	125	107	29	4	14
Extra features such as storage of associated data	30	63	105	182	100	6	26
Publisher profits	142	122	138	91	9	0	19
Paying referees	249	70	70	85	22	8	18
Other	3	1	1	1	3	2	3



“... Very-little of a high-cost journal may be more than a considerable amou[n]t of a low-cost one. Perhaps it would be better posed in terms of one’s priorities in paying for the journal. I think that in this day paying those such as the editors and referees, and ensuring the integrity of the archive, ought to be a higher priority than producing a paper version of the journal. Especially for an overlay journal such as you propose”.

### 4.5 Copy Editing

Copy editing, the level of author involvement in it, and who should be responsible for any costs associated with it, were also issues that were commented upon. Some respondents favoured the idea of charging extra for papers that require extensive copy editing. Almost half of the respondents favoured the suggestion that the cost of copy editing should be borne by the author, and that it should also be variable based on the amount of copy editing required. Furthermore, almost half of the respondents (47%) appear to be in agreement that those changes should be carried out by the author (Table 2). The appearance and layout of the published papers were considered important.

*“The idea of charging authors for papers that require excessive copyediting is a great one!”*

*“Copy editing is a difficult issue: it should be the [responsibility] of authors to improve their writing, on the other hand the journal should take [responsibility] for what it published. Perhaps an author could have say three chances and after that should pay for copy editing?”*

*“...my position is that a basic copy editing should be provided by the journal, but that extremely messy papers should be penalized, perhaps by introducing extra costs”*

**Table 2.** Copy editing

Statement	Rating	% agree	95% confidence limit
The cost of copy editing should be borne by the author and vary from paper to paper, depending on the amount of copy editing required		48.2	± 3.8
Copy editing should be carried out by the author		47.3	± 3.8
A referee should be prepared to assess whether or not copy editing is required		18.1	± 2.9
The cost of copy editing should be borne by the journal		11.1	± 2.4
When a journal makes copy edits, the corrected LaTeX should be returned to the author (after his/her approval)		4.7	± 1.6

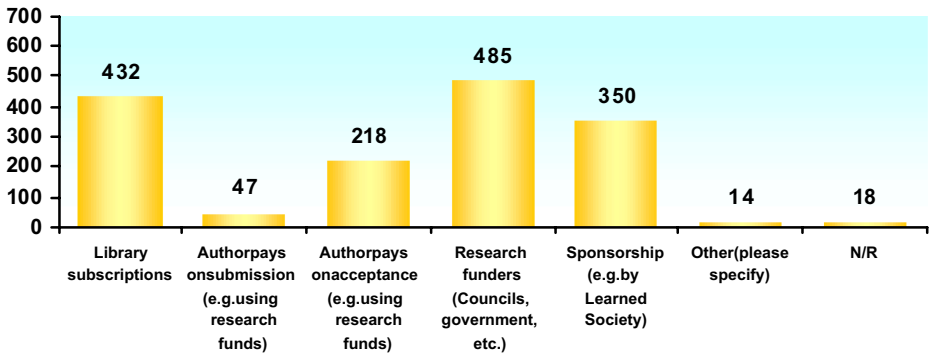
Key: Strongly disagree   Slightly disagree   Neither   Slightly agree   Strongly agree

*“I do believe money [is] being wasted on the copy-editing of already copy-edited articles, on paper copies of journals, on library subscriptions, etc. The publications process needs to be streamlined and a new type of open-access peer-reviewed journal might just be the right thing”.*

### 4.6 Funding

When asked where the funding to meet those costs should come from, the respondents preferred to select research funders (485 people, 71% of base=683), library subscriptions (432 people, 63%) and sponsorship, for example by a Learned Society (350 people, 51%). Models requiring an author to pay from research funds either on acceptance (218 people) or on submission (47 people) of a paper were not endorsed. Other possible funding sources mentioned in comments included: personal donations, professional association contributions, commercial and/or not-for-profit organisations, advertisements, subscriptions and even models in which authors pay partially on submission and partially on acceptance.

**Sources for covering journals' costs**



**Fig. 1.** Sources for covering journals' costs

## 5 Summary and Conclusions

The RIOJA community surveys found some encouragement for the overlay journal model in the fields of Astrophysics and Cosmology. However, they raised several implementation issues that they would consider important, primarily relating to the quality of the editorial board and of the published papers, and to the long-term archiving of the accepted research material. The traditional copy-editing function remains important to researchers in these disciplines, as is visibility in indexing services. The traditional printed volume is of little interest. These are generic concerns, for which repository overlay is not necessarily the complete answer.

Although the interviews with publishers and editors did not reveal any substantial information about costings that have not already been reported in the literature (King,

2007, SQW Limited, 2004; Waltham, 2004) or are available on some publishers' websites, some of the cost projections and business analysis for the development and maintenance of a journal founded on overlay certification in this field could help to inform future undertakings of this nature in different disciplines and with different repositories. Case studies or scenarios which involve setting up a new journal or converting an existing one to an overlay model will allow more precise definition of costing projections.

## References

- Cockerill, M.: Business models in open access publishing. In: Jacobs, N. (ed.) *Open Access: Key Strategic, Technical and Economic Aspect*, pp. 89–95. Chandos Publishing, Oxford (2006) (Last accessed 13/08/2008),  
<http://demo.openrepository.com/demo/handle/2384/2367>
- Crow, R., Goldstein, H.: *Model Business Plan: A Supplemental Guide for Open Access Journal Developers & Publishers*, Open Society Initiative (Last accessed 13/08/2008) (2003a),  
[http://www.soros.org/openaccess/oajguides/oaj\\_supplement\\_0703.pdf](http://www.soros.org/openaccess/oajguides/oaj_supplement_0703.pdf)
- Crow, R., Goldstein, H.: *Guide to Business Planning for Launching a New Open Access Journal*, Open Society Institute. 2nd edn. (Last accessed 13/08/2008) (2003b),  
[http://www.soros.org/openaccess/oajguides/business\\_planning.pdf](http://www.soros.org/openaccess/oajguides/business_planning.pdf)
- Crow, R., Goldstein, H.: *Guide to Business Planning for Converting a Subscription-based Journal to Open Access*, Open Society Institute (Last accessed 13/08/2008) (2003c),  
[http://www.soros.org/openaccess/oajguides/business\\_converting.pdf](http://www.soros.org/openaccess/oajguides/business_converting.pdf)
- Ginsparg, P.: *Winners and Losers in the Global Research Village*. Invited contribution, UNESCO Conference HQ, 19–23 (1996) (Last accessed 13/08/2008),  
<http://xxx.lanl.gov/blurb/pg96unesco.html>
- Hagemann, M.: *SPARC Innovator* (Last accessed 13/08/2008) (December 2006),  
<http://www.arl.org/sparc/innovator/hagemann.html>
- Halliday, L., Oppenheim, C.: *Economic models of the Digital Library* (Last accessed 13/08/2008) (1999),  
<http://www.ukoln.ac.uk/services/elib/papers/ukoln/emod-diglib/final-report.pdf>
- King, D.: *The cost of journal publishing: a literature review and commentary* 20(2), 85–106 (2007)
- Lewis, A.: *RIOJA Journal-Repository APIs* (Last accessed 13/08/2008) (2007),  
<http://cosmologist.info/xml/APIs.html>
- Prosser, D.C.: *Fulfilling the promise of scholarly communication - a comparison between old and new access models*. In: Nielsen, E.K., Saur, K.G., Ceynowa, K. (eds.) *Die innovative Bibliothek: Elmar Mittler zum 65.Geburtstag*, pp. 95–106 (2005) (Last accessed 13/08/2008),  
<http://eprints.rclis.org/archive/00003918>
- Rightcom Ltd. *Business model for journal content: final report*, JISC, (2005) (Last accessed 13/08/2008),  
[http://www.nesli2.ac.uk/JBM\\_o\\_20050401Final\\_report\\_redacted\\_for\\_publication.pdf](http://www.nesli2.ac.uk/JBM_o_20050401Final_report_redacted_for_publication.pdf)

Smith, J.W.T.: The deconstructed journal: a new model for academic publishing. *Learned Publishing* 12(2), 79–91 (1999)

SQW Limited: Costs and business models in scientific research publishing: a report commissioned by the Wellcome Trust, Wellcome Trust (2004)

Waltham, M.: Learned Society Open Access Business Models, JISC, (2005) (Last accessed 13/08/2008),

[http://www.jisc.ac.uk/uploaded\\_documents/Learned%20Society%20Open%20Access%20Business%20Models.doc](http://www.jisc.ac.uk/uploaded_documents/Learned%20Society%20Open%20Access%20Business%20Models.doc)

# A Collaborative Filtering Algorithm Based on Global and Domain Authorities

Li Zhou, Yong Zhang, and Chun-Xiao Xing

Tsinghua National Laboratory for Information Science and Technology  
Research Institute of Information Technology, Tsinghua University, Beijing 100084, China  
zhouchi03@mails.tsinghua.edu.cn,  
{zhangyong05, xingcx}@tsinghua.edu.cn

**Abstract.** Collaborative filtering has been very successful in both applications and researches. In real situation, different users may have different influences on other users' decisions. Those authoritative users usually play more important roles. But few existing collaborative filtering algorithms consider the authorities of users. In this paper, we present the concepts of global and domain authorities of users, and apply them in collaborative filtering algorithms. This paper designs the experiments and discusses the effects of global and domain authorities. The initial results show our method can improve the performance of collaborative filtering algorithm.

**Keywords:** Collaboration Filtering, Global Authority, Domain Authority.

## 1 Introduction

With the rapid growth of the web, people have to spend more and more time to find what they are interested in. Personalized recommendation technology can record users' interests while users are browsing or using items which are resources. It can analyze interests of different users and recommend items to users according to their interests. Therefore, it can solve the conflict between users' needs and the massive information [1].

Collaborative filtering is one of the most successful personalized recommendation technologies which have been used in both applications and researches. The key task of collaborative filtering algorithm is to predict rating of an active user to a target item based on existing users' ratings. With the blooming of Web2.0, more and more people publish dairies on blogs, collaborate by writing on wikis, or communicate by SNS(Social Networking Services), etc. These Web 2.0 applications have generated a large amount of rating data, which causes appealing demand of collaborative filtering. Different users have different backgrounds, professional skills, occupations, and so on. If a user does not have so much experience or knowledge about one item, his ratings may not be so relevant. Therefore, it is necessary to consider the authorities of users. Authoritative users include users who have the related experiences or influences, domain experts or users who come from authoritative organizations. Authoritative users can rate items more accurately. If we could consider the influences of authoritative users in collaborative filtering algorithms, we can make better predictions. However,

few existing collaboration filtering algorithms consider authority and there lacks the methods to calculate authority.

In this paper, we present collaborative filtering algorithms based on authorities of users to improve the performance of recommendation. The rest of the paper is organized as follows. The next section presents a brief background of collaborative filtering algorithms and authorities of users. In section 3, we present global and domain authorities, and then design a method to calculate global and domain authorities that can improve the similarity calculation in collaborative filtering. Section 4 describes our experimental work including data sets, evaluation metrics, results and discussion. The final section provides conclusion and directions of future research.

## 2 Related Work

### 2.1 Collaborative Filtering Algorithm

Collaborative filtering has been the most successful personalized recommendation technology. Generally collaborative filtering algorithms recommend items according to the similarities between users. Instead of analyzing the content of items, collaborative filtering algorithms utilize the users' ratings of items to find out the similarities between users, and then predict ratings of some user by using the ratings from similar users. There the most important problems in collaborative filtering algorithms are how to compute the similarity between users and how to predict rating using the ratings from similar users.

#### 2.1.1 Memory-Based and Model-Based Algorithm

According to [2], there are generally two kinds of algorithms: memory-based methods and model-based methods.

The memory-based algorithms read all data into memory, calculate the similarities between users or items, and recommend items which the active user may be interested in using similarities. The classical methods are Correlation-based or Cosine-based.

The model-based collaborative filtering algorithms build a model using rating matrix and calculate similarities between users or items with this model. There are many methods to build the models, such as statistics method and machine learning method. For example, Breese uses probability to build model. Chien et al [3] uses Bayesian model. Hofmann [4, 5] uses latent semantic models. Sarwar et al [6] uses linear regression. Luo Si et al [17] extends existing portioning/clustering algorithms by clustering both users and items together simultaneously without assuming that each user and item should only belong to a single cluster. David et al [19] propose and evaluate a CF method called personality diagnosis that can be seen as a hybrid between memory- and model-based approaches.

#### 2.1.2 User-Based Algorithm and Item-Based Algorithm

The original method of collaborative filtering calculates similarities between users and recommends same items to similar users. Sarwar et al [6] give a method to recommend items which are similar to the items that the active user likes. The former algorithm is a user-based algorithm, while the latter one is an item-based algorithm.

Pearson Correlation Coefficient (PCC) algorithm is the most popular user-based collaborative filtering algorithm, and is a memory-based algorithm. The basic steps of PCC algorithm is as follows.

Assume there is a user  $u_a$ , we are going to predict  $u_a$ 's rating on item  $t_b$ .

**Step 1.** Calculate similarity between user  $u_a$  and  $u_i$   $sim(u_a, u_i)$  using PCC, please refer to Zeng's paper [1]. However, PCC only measures overlapping items between users. In fact the number of overlapping items is very small between some users. To decrease the error brought by few co-rated items Sarwar et al add a significance-weighting factor [6].

$$sim(u_a, u_i)' = sim(u_a, u_i) * \frac{|T_{u_i} \cap T_{u_a}|}{50} \quad (1)$$

where  $|T_{u_i} \cap T_{u_a}|$  is the number of items that rated by both user  $u_a$  and user  $u_i$ . If  $|T_{u_i} \cap T_{u_a}|$  is more than 50, we let  $sim(u_a, u_i)' = sim(u_a, u_i)$ .

**Step 2.** Select K nearest neighbors of target user  $u_a$  according to  $sim(u_a, u_i)$ .

**Step 3.** Predict the rating of user  $u_a$  on item  $t_b$  by utilizing weighted mean of its K nearest neighbors.

$$p_{u_a, t_b} = \bar{r}_{u_a} + \frac{\sum_{u \in S_{u_a}} sim(u_a, u_i)' * (r_{u, t_b} - \bar{r}_u)}{\sum_{u \in S_{u_a}} sim(u_a, u_i)'} \quad (2)$$

where  $S_{u_a}$  is the neighbor set of user  $u_a$ .

## 2.2 Authority

PageRank algorithm [7] and HITS algorithm [8] are the most famous page rank algorithms. Page Rank algorithm takes the whole Internet as a web graph. The more web pages that a page is pointed to, the more authoritative the page is. Each web page has a PageRank score by iterative calculation. Then we can rank the importance of pages by PageRank score. HITS algorithm calculates two scores for each page, hub score and authority score. These two classical algorithms have mentioned the authorities of pages which is similar to the authorities of users in this paper. Anselm [9] proposed the concept of authority effect. Documents found by multiple systems are more likely to be relevant, which is referred to as the authority effect. It is also similar to the authority in this paper. Domingos et al [10] take user as a node of social network and the influence of user to others as a Markov random field. Then they find users who are important to enterprises by mining the influence of users on data set EachMovie. Rashid et al [11] introduced the concept of user influence, and use seven metrics to measure user influence. They bring forward an algorithm to evaluate user influence, which measures user influence by removing some user's ratings and comparing the

difference between the recommendation results, that is, if the difference is big, the user is more influential.

Many papers have discussed the authority or concepts which are similar to authority, but have not utilized authority to improve collaboration filtering algorithm. What's more, they did not consider the problem that some users may be authoritative just in some domains. Therefore, few collaboration filtering algorithms consider the authorities of users, especially authorities in specific domains. In this paper we improve collaboration filtering algorithm using authority.

### 3 Collaborative Filtering Algorithm Based on Authority

We divide authorities into global authority and domain authority. Global authority is the authority on the whole items that some user has. Domain authority is the authority on the items of some domain that some user has. This section is organized as follows. First we present Global Authority Pearson Correlation Coefficient (GAPCC) algorithm and then we talk about Domain Authority Pearson Correlation Coefficient (DAPCC) algorithm. Finally we present a simple Hybrid Authority Pearson Correlation Coefficient (HAPCC) algorithm which combines GAPCC and DAPCC.

#### 3.1 GAPCC Algorithm

We calculate global authorities by statistic user-item matrix. We also consider about users' activeness, justness of rating, attention on cold items and so on. Rashid et al [11] have compiled some qualitative factors that seem to affect influence levels of user. Since influence of users is similar to global authority here. We choose some factors to measure global authority in our experiments.

We define  $AU(u)$  as global authority and  $T(u)$  as the item set that user  $u$  has rated.

The more items some user has rated, the more possibility that he or she will be similar to other users. If a user has rated a lot of items, he or she may have a lot of experience. Therefore, the simplest way to measure global authority is to calculate rating number. The more items that a user rates, the more authoritative the user is. The first feature of global authority is number of rating:

$$AU(u) = T(u) \quad (3)$$

If rating of a user is similar to most users, the user's rating can be used to represent most people's opinion. So the second feature of global authority is degree of agreement with others:

$$AU(u) = \frac{\sum_{t_j \in T(u)} |r_{u,t_j} - \bar{r}_{t_j}|}{|T(u)|} \quad (4)$$

If a user rates many cold items, we can think that the user is very serious. So he or she has higher authority. Cold items are those items which are rarely rated after a long period of time in the system. This measurement method is similar to inverse document frequency. So the third feature of global authority is rarity of rated items.



$$AU(u) = \sum_{t_j \in T(u)} \frac{1}{freq(t_j)} \tag{5}$$

$freq(t_j)$  is the rating number of item  $t_j$ .

We can consider the mean variance of rating set. If the mean variance is very small, that means the user always give same rating to items and is not serious. Otherwise, the user can be taken as serious and thus the authority of him is higher. So the forth feature of global authority is the involvement degree of user:

$$AU(u) = \sum_{t_j \in T(u)} \frac{(r_{u,t_j} - \bar{r}_u)^2}{|T(u)|} \tag{6}$$

If the similarities of a user to his nearest neighbors is high, that means the user can represent others' opinions. So the fifth feature of global authority is the average similarity of a user to K nearest neighbors:

$$AU(u) = \sum_{u_i \in Neighbor(u)} \frac{sim(u, u_i)}{K} \tag{7}$$

Because the ranges of authorities calculated by different ways are different, the authorities need to be normalized. We can do linear transformation on initial data utilizing min-max standardization. We assume that  $Max(AU)$  is the maximum value of global authority  $AU(u)$  and  $Min(AU)$  is the minimum value of  $AU(u)$ . Here we let  $a=0$  and  $b=1$ . Authority  $v$  can be defined like:

$$v = \frac{AU(u) - Min(AU)}{Max(AU) - Min(AU)}(b - a) + a \tag{8}$$

Finally, we can integrate global authority into collaborative filtering algorithm to predict rating given to a target item by an active user. GAPCC algorithm is as follows:

**Input:** (1)User set  $U=\{u_1, u_2, u_3, \dots, u_m\}$  (2)Item set  $T=\{t_1, t_2, t_3, \dots, t_n\}$  (3)Rating matrix  $R_{m \times n}=(r_{u_i, t_j})$ , the rating given to item  $t_j$  by user  $u_i$  is  $r_{u_i, t_j}$  (4)User  $u_a$  (5)Item  $t_b$

**Output:** the predicted value  $P_{u_a, t_b}$  of user  $u_a$  on item  $t_b$

**Step 1.** For each user  $u_i \in U$ , calculate global authority  $w(u_i)$  according to formula 3-8.

**Step 2.** Calculate similarities between users.

**Step 3.** Calculate the predicted value given to item  $t_b$  by user  $u_a$ .

$$P_{u_a, t_b} = \bar{r}_{u_a} + \frac{\sum_{u \in S_{u_a}} w(u_i) * sim(u_a, u_i) * (r_{u, t_b} - \bar{r}_u)}{\sum_{u \in S_{u_a}} w(u_i) * sim(u_a, u_i)} \tag{9}$$

### 3.2 DAPCC Algorithm

It is impossible for every user to be authoritative in all domains. Usually a user only has authority in some domain, and we think the user has some domain authority.

We divide items into  $K$  clusters,  $\{C_1, C_2, C_3, \dots, C_K\}$ ,  $T=C_1 \cup C_2 \cup C_3 \dots \cup C_K$ . If the items are texts, for example papers, we can use Bayesian. For films we can use fuzzy K-means clustering algorithm. In traditional K-means clustering algorithms, each item can only belong to one cluster, but in fact an item can belong to several clusters. Therefore we use fuzzy K-means algorithm to put an item into several clusters with different probabilities, please refer to Bezdek's paper [18].

$w(u_i, C_k)$  is the authority of user  $u_i$  in domain  $C_k$ . It can be calculated by the following ways:

$$w(u_i, C_k) = \sum_{t_j \in T(u_i)} P(t_j \in C_k) \quad (10)$$

$$w(u_i, C_k) = \left\{ \sum_{t_j \in T(u_i)} P(t_j \in C_k) \right\} * \frac{\sum_{t_j \in T(u_i)} (r_{u_i, t_j} - \bar{r}_{t_j})}{|T(u_i)|} \quad (11)$$

After we integrate domain authority into similarity between users, the similarity between users contains two parts, one is Pearson correlation coefficient between users' rating vector, the other part is Pearson correlation coefficient between users' domain authority vector.

$$sim(u_1, u_2)_{authority} = (1-\alpha) * correlation(u_1, u_2)_{authority} + \alpha * correlation(u_1, u_2)_{rating} \quad (12)$$

where  $correlation(u_1, u_2)_{authority}$  is the correlation between authority vectors of user  $u_1$  and  $u_2$ ,  $correlation(u_1, u_2)_{rating}$  is the correlation between rating vectors of user  $u_1$  and  $u_2$ ,  $\alpha$  is a weighting factor, the minimum number of  $\alpha$  is 0 which means the similarities are all from the authorities of user clustering matrix, the maximum number of  $\alpha$  is 1 which means the similarities are all from traditional collaborative filtering algorithm. The collaborative filtering algorithm based on domain authority is follows (input and output are the same as GAPCC):

**Step 1.** Divide item set into  $K$  clusters using Fuzzy K-means algorithm and get item-cluster matrix.

**Step 2.** Calculate user-cluster matrix and each value in the matrix is the authority of some user in some cluster. Then normalize all the values on each column, whose sum is 1 (that is the sum of all users' authorities on a cluster is 1).

**Step 3.** Calculate similarities between users.

**Step 4.** Calculate the predicted rating which given to item  $t_b$  by user  $u_a$ .

### 3.3 HAPCC Algorithm

We combine GAPCC and DAPCC to get hybrid authority Pearson correlation coefficient (HAPCC) algorithm.

$$p_{u_a, t_b \text{HAPCC}} = (1 - \beta) * p_{u_a, t_b \text{GAPCC}} + \beta * p_{u_a, t_b \text{DAPCC}} \quad (13)$$

where  $\beta$  is a weighting factor,  $p_{u_a, t_b \text{GAPCC}}$  is the predicted rating calculated by GAPCC algorithm,  $p_{u_a, t_b \text{DAPCC}}$  is the predicted rating calculated by DAPCC algorithm.

## 4 Experiment and Result Analysis

### 4.1 Experiment Design

This paper uses data set MovieLens [12] which contains 100,000 ratings given to 1682 films by 943 users. We process data set in two methods. The first one is to divide data set into five groups and then divide each group into two sets: training set (80%) and test set (20%). The other one is that we first divide data set into two groups, and then in each group we just select 10 rating items as test items while the other as training items. We call this method All but 10.

The common evaluation metrics are accuracy, coverage, novelty, confidence and so on [13]. In this paper we use Mean Absolute Error (MAE), accuracy, coverage and run time as evaluation metrics. MAE is the most popular metric to evaluate prediction experiments and has been used in many papers [2,14,15,16]. It calculates absolute error between the real rating and predicted value in the test set, and then gets average value of these errors. The calculation of MAE is as follows:

$$\text{MAE} = \frac{\sum_{u \in U_{\text{test}}} |r_{u, t_j} - p_{u, t_j}|}{|U_{\text{test}}|} \quad (14)$$

where  $r_{u, t_j}$  is the rating given to item  $t_j$  by user  $u$ ,  $p_{u, t_j}$  is the predicted value of user  $u$  on item  $t_j$ ,  $U_{\text{test}}$  is the test set,  $|U_{\text{test}}|$  is the size of test set.

Accuracy is the percentage of the predicted values which hit the real rating values in the whole test set. Coverage is the percentage of the instance that can be predicted by algorithm in the whole test set. Run time is the whole run time when the size of test set is 20,000. Our experiments were conducted on AM300+ CPU with 1G memory.

There are three groups of experiments. We consider MAE of PCC, GAPCC and DAPCC with maximum neighbor number. We also consider the influence of weighting factor  $\alpha$  to accuracy. At last, we compare run time, coverage, accuracy of the three algorithms.

### 4.2 Results Analysis

Figure 1 and 2 compare PCC, GAPCC and DAPCC. Figure 1 uses the five data groups which are divided into 80% training set and 20% test set. Figure 2 uses data set which is divided using All but 10. The average values are used. We compare the MAEs of the three algorithms when maximum neighbor numbers are 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100.

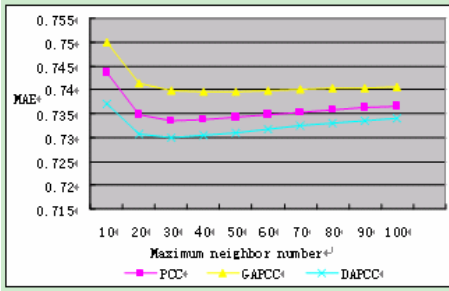


Fig. 1. MAE with maximum neighbor number (80%/20%)

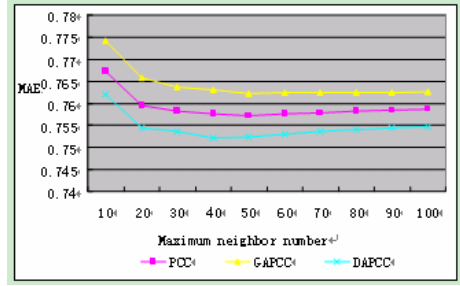


Fig. 2. MAE with maximum neighbor number (All but 10)

From figure 1 and figure 2, we can find that GAPCC is a little worse than PCC. When we bring domain authority into PCC, the accuracy is improved. For example, when maximum neighbor number is 30, the MAE of PCC is 0.731 while DAPCC is 0.728. Therefore, domain authority can express authorities of users on items more accurately and can improve the accuracy of recommendation.

Figure 3 shows the MAE with different weighting factor  $\alpha$ . We fix maximum neighbor number as 50. The result shows that the accuracy of HAPCC is correlative to weighting factor  $\alpha$ . At most time, the performance of HAPCC is not as good as DAPCC. But when  $\alpha=0.3$ , the performance of HAPCC is better than DAPCC.

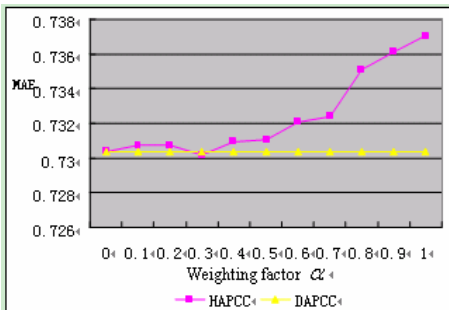


Fig. 3. MAE with weighting factor  $\alpha$

Table 1. Comparison of the four algorithms

	MAE	Accuracy	Coverage
PCC	0.733	0.423	0.9969
GAPCC	0.738	0.422	0.9968
DAPCC	0.731	0.425	0.9972
HAPCC	0.731	0.423	0.9971

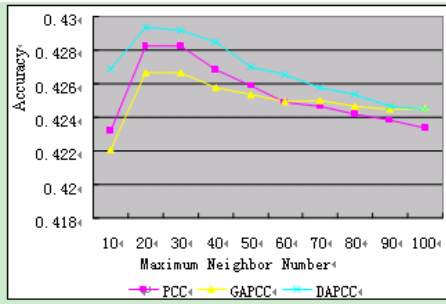


Fig. 4. Accuracy with maximum neighbor number

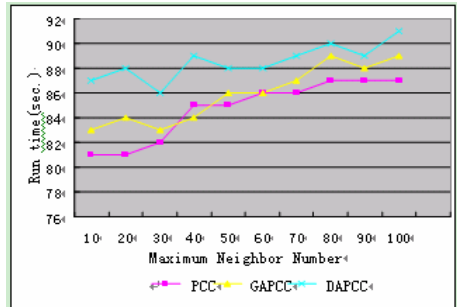


Fig. 5. Run time with maximum neighbor number

At last we test run time, coverage and accuracy of the four algorithms PCC, GAPCC, DAPCC and HAPCC. In previous experiments, we see the accuracy of HAPCC algorithm is highest when  $K=30$  and  $\alpha=0.3$ . So we consider run time, coverage and accuracy of the four algorithms under the condition of  $K=30$  and  $\alpha=0.3$ .

From table 1, figure 4 and figure 5, we can find that DAPCC improves on MAE comparing with PCC. They have similar accuracy, coverage and run time. Since the performance of HAPCC algorithm is not satisfactory. We did not include HAPCC in figure 4 or figure 5.

## 5 Conclusion

Collaborative filtering has been very successful in both applications and researches. In real situations, different users have different influences on other users' decisions. Authoritative users play more important roles to represent others. However, few existing collaborative filtering algorithms consider the authorities of users. In this paper, we propose concepts of global and domain authorities, and apply them in collaborative filtering algorithms PCC. The key points we focus on are how to calculate authority and then use it to calculate similarities between users. The initial results have shown some improvements to the performance of collaborative filtering algorithm. In the future, we will continue to investigate on the problems such as how to estimate and measure authorities of users to improve performance further.

**Acknowledgments.** This paper is supported by the Key Technologies R&D Program of China under Grant No. 2006BAH02A12, and the National High-tech R&D Program of China under Grant No.2006AA010101

## References

1. Xing, Z.C.-X., Zhou, L.-Z.: Similarity Measure and Instance Selection for Collaborative Filtering. In: Proceedings of 12th International World Wide Web conference (2003)
2. Breese, J.S., Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp. 43–52 (1998)

3. Chien, Y.-H., George, E.I.: A Bayesian Model for Collaborative Filtering. In: Proc. Seventh Int'l Workshop Artificial Intelligence and Statistics (1999)
4. Hofmann, T.: Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis. In: Proc. 26th Ann. Int'l ACM SIGIR Conf. (2003)
5. Hofmann, T.: Latent Semantic Models for Collaborative Filtering. *ACM Trans. Information Systems* 22(1), 89–115 (2004)
6. Sarwar, B., Karypis, G., Konstan, J., et al.: Item based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference, pp. 285–295 (2001)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the Seventh International World Wide Web Conference (1998)
8. Kleinberg, J.: Authoritative sources in a hyperlinked environment. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (1998)
9. Spoerri, A.: Authority and Ranking Effects in Data Fusion. *Journal of the American society for information science and technology* 59(3), 450–460 (2008)
10. Domingos, P., Richardson, M.: Mining the Network Value of Customers. In: Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, pp. 57–66. ACM Press, New York (2001)
11. Rashid, A., Karypis, G., Riedl, J.: Influence in Ratings-Based Recommender Systems: An Algorithm-Independent Approach. In: *SDM* (2005)
12. MovieLens: <http://movielens.umn.edu>
13. Herlocker, J., Konstan, J., Terveen, L., Riedl, J.: Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22, 5–53 (2004)
14. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Analysis of Recommendation Algorithms for E-Commerce. In: Proceedings of the ACM EC 2000 Conference Minneapolis, MN, pp. 158–167 (2000)
15. Herlocker, J., Konstan, J., Borchers, A., Riedl, J.: An Algorithmic Framework for Performing Collaborative Filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 1999, pp. 230–237 (1999)
16. Lemire, D., Maclachlan, A.: Slope One Predictors for Online Rating-Based Collaborative Filtering. In: *SIAM Data Mining (SDM 2005)*, Newport Beach, California, April 21–23 (2005)
17. Si, L., Jin, R.: Flexible mixture model for collaborative filtering. In: Proceedings of the Twentieth International Conference on Machine Learning (2003)
18. Bezdek, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York (1981)
19. Pennock, D.M., Horvitz, E., Lawrence, S., Giles, C.L.: Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In: The Proceeding of the Sixteenth Conference on Uncertainty in Artificial Intelligence (2000)

# Complex Data Transformations in Digital Libraries with Spatio-Temporal Information

Bruno Martins, Nuno Freire, and José Borbinha

Instituto Superior Técnico, Technical University of Lisbon,  
Av. Rovisco Pais, 1049-001 Lisboa, Portugal  
{bruno.g.martins,nuno.freire,jlb}@ist.utl.pt

**Abstract.** The DIGMAP project researched automated methods for enriching metadata records with structured geo-temporal information. This paper presents our findings regarding the use of XML technology for expressing transformations between the different XML schemas used in DIGMAP metadata records and service interfaces. Both XSLT and XQuery are functional, declarative languages that effectively support XML data integration. They are also extensible, in the sense that new functions can be specified in Java and then combined with general XPath expressions. We extended an XSLT/Xquery engine with additional functions for processing spatio-temporal information and for dealing with incompleteness and inconsistencies in the data. The paper discusses the application over different XML formats and metadata standards.

**Keywords:** XML, data integration, spatio-temporal reasoning, interoperability.

## 1 Introduction

The DIGMAP project<sup>1</sup> addressed the development of an architecture of services for virtual digital libraries of old maps, capable of collecting metadata from different providers and offering rich searching and browsing interfaces that combine thematic and spatio-temporal aspects. Some of its components, e.g. the DIGMAP gazetteer [10] and geoparser [9] services, specifically deal with automated methods for enriching the metadata with structured spatio-temporal information. This enriched metadata is afterwards used to support the searching and browsing interfaces.

DIGMAP collects resources through warehousing, leveraging on the widespread availability of protocols like OAI-PMH<sup>2</sup>. We store and redistribute the original metadata records, and we also map them into a common model based on the schema of The European Library (TEL)<sup>3</sup>. A particular challenge was to find effective ways for expressing the transformations between the original source metadata and the DIGMAP format. Because these transformations are mostly declarative (for the most part, they are as trivial as *element A in schema S<sub>1</sub> maps to element B in schema S<sub>2</sub>*), we want to express them using a correspondingly concise and declarative language.

---

<sup>1</sup> <http://www.digmap.eu>

<sup>2</sup> <http://www.openarchives.org/pmh/>

<sup>3</sup> <http://www.theeuropeanlibrary.org/>

However, not all the required transformations are declarative. Some involve the use of procedural or functional logic in order to express syntactic transformations (e.g. converting geographic coordinates and dates) and to deal with the pervasive problems of metadata inconsistency, non-uniformity, incorrectness and incompleteness.

Another particular challenge concerned the service interfaces associated with some of the components in the DIGMAP architecture. For instance both the geoparser and the gazetteer services support the enriching of the metadata records by finding geo-temporal references given over text (i.e. extracting and disambiguating the names for places and historical periods). In order to increase the potential for reuse, these services support many common standards and interfaces (e.g. geoRSS<sup>4</sup>, KML<sup>5</sup>, SpatialML<sup>6</sup> or the Alexandria Digital Library gazetteer standard<sup>7</sup>). Special attention was therefore given to finding XML processing technology capable of dealing with the specificities of spatio-temporal information, in order to express transformations between the required formats and to support the combination of DIGMAP service outputs with the metadata information, this way producing the enriched metadata.

This paper describes our findings regarding the use of XML processing languages, i.e. XQuery and XSLT, for expressing mappings between different XML metadata schemas and service interfaces. XML technology is already widely used to extract data, transform it, and create new structures in easily understandable ways. However, XML languages lack predicates for processing data according to spatio-temporal criteria, or for appropriately dealing with the inconsistencies that occur in real-world metadata. We extended an open-source XSLT/XQuery engine with new functions for dealing with these problems. The extensions enable the use of XSLT/XQuery in many advanced spatio-temporal data integration tasks, involving the querying and transformation of distributed and/or heterogeneous sources.

The rest of this paper is organized as follows: Section 2 presents the main concepts and related research. Section 3 describes our extensions to the XSLT/XQuery engine, discussing their implementation with basis on existing open-source APIs. Section 4 presents test cases, demonstrating the applicability of the proposed approach. Finally, Section 5 gives our conclusions and presents directions for future work.

## 2 Concepts and Related Work

The eXtensible Markup Language (XML) is currently the lingua franca for exchanging information between heterogeneous systems. Application areas like Digital Libraries increasingly rely on XML for system interoperability.

XML documents can be queried through expression languages like XPath and XQuery, or through specialized software using standard programming APIs like SAX or DOM. Standards for specifying transformations between XML documents have also been developed, e.g. Extensible Stylesheet Language Transformations (XSLT).

The XSLT 2.0 and XQuery 1.0 languages were developed by separate working groups within the World Wide Web Consortium (W3C). Still, these groups worked

---

<sup>4</sup> <http://georss.org/>

<sup>5</sup> <http://code.google.com/apis/kml/>

<sup>6</sup> <http://mirror.optus.net/sourceforge/s/sp/spatialml/SpatialML-1.0-March30-2007.pdf>

<sup>7</sup> <http://www.alexandria.ucsb.edu/gazetteer/ContentStandard/version3.2/GCS3.2-guide.htm>



together to ensure a common approach. Both languages share the same data model, type system and function library, and both include XPath 2.0 as a sublanguage. XPath is essentially a description language for defining navigational paths over XML documents, with conditions on the labels and contents of elements and attributes.

XQuery and XSLT are both declarative, functional languages that have been shown to be Turing-complete [5,14]. However, they are rooted in different traditions and serve the needs of different communities. XSLT is stronger in its handling of narrative documents with a more flexible structure, while XQuery is stronger in its data handling capabilities, e.g. simplicity in performing relational joins using FLWOR expressions (e.g. series of FOR, LET, WHERE, ORDER-BY and RETURN clauses).

The Digital Libraries community has dealt with the problem of assimilating heterogeneous metadata into common frameworks, in order to produce enhanced metadata that can latter be re-used [4]. Crosswalks, i.e. tables mapping schema relationships and equivalencies, have been created between all the major metadata standards and are being gathered into repositories [2]. Given that the task is often presented as the mapping from one XML-encoded format to another, XSLT suggests itself as a good candidate and has indeed been the most commonly employed tool for this purpose [15]. However, it has been noted that XSLT can be too verbose and lacking on expressiveness [3,6]. It is difficult to write transformations beyond the simple matching between XML structures, rendering things such as the transformation of particular metadata elements (e.g. processing geospatial coordinates) from very difficult to practically impossible. Past works have attempted to overcome XSLT's deficiencies by integrating it into a procedural language like Java [6]. However, this approach creates a large division between the declarative and procedural aspects of the transformations. Java is also generally too rigid and low-level for the purposes of metadata transformations. Other works have addressed the problem by using high-level scripting languages like Python to express mappings [3]. However, reusing the existing mappings expressed in XSLT in an important concern.

The Geographic Information Systems (GIS) community, by way of the OpenGIS Consortium (OGC)<sup>8</sup>, has also embraced XML for spatial data interoperability. OGC is currently leading the development of standards for Web-based GIS, promoting XML service interfaces for accessing and manipulating geographic information. OGC service interfaces, in turn, build upon other XML standards, such as the Geography Markup Language (GML) to encode spatial information or the Filter Encoding Specification to allow the formation of query filters. Despite these efforts, geospatial data integration remains a complex problem. Different sources may vary on the employed projections, accuracy levels and formats, presenting many challenges to interoperability. Previous studies have addressed the usage of XML technology to integrate geospatial datasets [12], but further research is still needed. Previous works have also addressed XQuery extensions to better support specific problem domains, e.g. spatio-temporal data and moving objects databases [11]. GQL is an example of a query language supporting spatial queries over GML documents, extending XQuery in what regards the data model, algebra and formal semantics, as well as adding various spatial functions and operations [8]. This paper presents similar ideas, as we describe the extension of XPath's function library with novel spatio-temporal and data integration capabilities related to the context of digital libraries.

---

<sup>8</sup> <http://www.opengeospatial.org/>

### 3 The Proposed Extensions to the XPath 2.0 Function Library

XML processing languages were designed to be extendable by both users and implementers. The XPath 2.0 function library<sup>9</sup> already defines a basic set of functions suitable for a broad range of purposes. However, advanced computations related to data integration require new functions. This work describes extensions to the XPath 2.0 function library related to data integration and spatio-temporal computations (e.g. measuring distance, containment and similarity). The specific computations can be made by the extension functions without affecting the XSLT/XQuery syntax. This way, existing mappings in XSLT or XQuery can be used without any modifications, or they can be enhanced by adding new transformation rules or simplifying existing ones. Typical Java implementations for XQuery and XSLT support the calling of regular Java methods similarly to other XPath 2.0 functions. We therefore only have to provide appropriate Java implementations for the required extension functions. The extensions that we used within DIGMAP can be grouped into four general categories:

1. **Spatial reasoning:** Functions for processing geospatial data (e.g. format conversions or computing distance/containment).
2. **Temporal reasoning:** Functions for processing temporal data.
3. **Text mining:** Functions for processing textual data (e.g. tokenization, keyword-based search, string similarity and thesaurus mapping with basis on names).
4. **Miscellaneous:** General functions related to data integration (e.g. calling XSLT from XQuery, calling external Web services, importing CSV data, etc.).

Due to space constraints, we only give brief descriptions for some of the considered extension functions<sup>10</sup>. The rest of this Section focuses on the functions for spatial and temporal reasoning, ending with a discussion on implementation issues.

#### 3.1 Extension Functions for Spatial and Temporal Reasoning

Some of the considered extension functions support complex selections and joins over spatio-temporal data (e.g. combining two or more sets of spatial elements according to a spatial predicate such as distance or intersection). They were implemented as Java functions that i) take as input a string or XML representation for the geometry(ies) or temporal extent(s), ii) perform the actual computations, and iii) return a string or XML representation for the result. The first and last steps correspond to data marshalling and marshalling operations (e.g. a GML tree is flattened into a Java object, the operation is performed over this object and the result is converted back into a GML tree).

The input to the spatial predicates (i.e. the geometries) can be provided as either:

- An XML element encoding a spatial geometry in GML or KML (e.g. the input is an XML Node corresponding to the root element of a GML or KML tree).
- A string encoding a spatial geometry in OGC's well-known text (WKT) format.

---

<sup>9</sup> <http://www.w3.org/TR/xpath-functions/>

<sup>10</sup> <http://transform.digmap.eu/functions.jsp>

- A string encoding a point through C-Squares<sup>11</sup> or Geohash<sup>12</sup> codes.
- A string encoding a point through a pair of numeric values in one of several possible notations (e.g. decimal degrees or degrees, minutes, and seconds).
- A sequence of XML elements, where each element encodes a spatial geometry.

Coordinates for the geometries can be provided in any given datum, geoid, coordinate system and map projection. When an OGC Spatial Reference Identifier (SRID) is not provided as an attribute, we assume the value `EPSG:4326`, which corresponds to the WGS84 encoding that is commonly used in the GPS system.

The considered operators for performing geospatial analysis are based on the functions described in the OGC Simple Features and Filter Encoding specifications. They include the testing of spatial relationships, and the computation of distances and spatial operators. Some of the functions that we considered are listed next:

- Return the distance, union, intersection or difference between two geometries.
- Check the validity of a given spatial filter, or check if two given geometries are spatially related (i.e. contained, overlap, equals, touches).
- Check if two geometries fall below a given distance threshold.
- Return the area, length, buffer, centroid, boundary or envelope of a given geometry.
- Perform geometric computations (e.g. translation or scaling) over a given geometry.
- Return the GML, KML, C-Square, Geohash or WKT encoding of a given geometry.
- Perform transformations on the coordinate systems used to represent geometries.

On what concerns temporal data, the considered extension functions mostly deal with the spatial reasoning operators of Allen's interval algebra [13], together with operations related to metric temporal information. The input to these functions can be given as an XML node encoding a GML time instant or time period, or as a string encoding a date in one of several supported formats (e.g. the ISO 8601 formats).

### 3.2 Implementation Issues

We used the Saxon<sup>13</sup> Java API for manipulating XML documents. Saxon supports XSLT 2.0 and XQuery 1.0. It also supports external functions by binding any Java class to a namespace prefix. With this binding in scope, users can invoke any static method in the referenced class. Saxon will invoke the Java method with the right name and number of arguments, converting the types of arguments and the return values as necessary to make a function fit (e.g. lists are converted to XPath node sets and Java number types are converted into their XML equivalents). We explicitly take care of converting the complex types that store results from the specialized functions (e.g. the marshaling and unmarshaling of the *Geometry* objects storing GML data).

The proposed extensions were mostly implemented through wrapper functions around the methods provided by other open-source Java APIs. For the functionalities not provided such APIs (e.g. thesaurus mappings or temporal reasoning), we provided our own implementations. Some of the Java APIs that we used were:

---

<sup>11</sup> <http://www.cmar.csiro.au/csquares/>

<sup>12</sup> <http://geohash.org/site/tips.html>

<sup>13</sup> <http://saxon.sourceforge.net>

- Nux<sup>14</sup>, which provides extension functions to Saxon related to keyword search and text processing (e.g. tokenization).
- NGramJ<sup>15</sup>, which provides an n-gram based language identification method.
- GeoTools<sup>16</sup> and Java Topology Suite (JTS<sup>17</sup>), which provide spatial object models and geometric functions, transformations between spatial referencing systems, methods for handling GML 2.0 documents and general support for OGC standards.
- SimPack<sup>18</sup>, which supports many different similarity functions for strings and tree structures, supporting operations like duplicate detection or similarity joins.
- MARC4J<sup>19</sup>, which provides methods for reading bibliographic information in the MARC and MODS standards.

The extended Saxon API can either be called from a command line or from a Java servlet that implements a simple REST interface for deploying XQueries or XSLTs. This service takes as parameters the actual XQuery/XSLT code or URLs pointing to documents with the code, together with `name=value` pairs for external variables. To execute an XQuery or an XSLT with the REST service, we require nothing more than an HTTP GET or POST request submitted to the service interface<sup>20</sup>.

## 4 Test Cases within the DIGMAP Project

The proposed approach has been successfully used within the DIGMAP project, supporting transformations between several well-known metadata standards as well as the definition of wrappers around some of DIGMAP's service interfaces. The transformations were exercised on real-world data from the project, exhibiting the classical problems of incorrectness, inconsistency, non-uniformity or incompleteness.

It should be noted that the mere ability to perform a transformation is not an evaluation criterion. Transformations can be created by any full-featured programming language or, in simple cases, by XSLT without considering any extensions. Evaluation should be based on the qualitative considerations that apply to programming language design [23, 8]: writability and readability; expressiveness; brevity; appropriateness of the level of abstraction to the domain; and orthogonality.

The next subsections describe three example transformations, following a discussion on the computational performance of the proposed approach.

### 4.1 Handling Bibliographic Records

Among others, the Royal Library of Belgium (KBR) provided us with a set of 5,371 UNIMARC bibliographic records encoded in ISO 2709. We used extension functions

---

<sup>14</sup> <http://acs.lbl.gov/nux/>

<sup>15</sup> <http://ngramj.sourceforge.net/>

<sup>16</sup> <http://geotools.codehaus.org/>

<sup>17</sup> <http://www.vividsolutions.com/jts/>

<sup>18</sup> <http://www.ifi.uzh.ch/ddis/simpack.html>

<sup>19</sup> <http://marc4j.tigris.org>

<sup>20</sup> <http://transform.digmap.eu>

related to text processing, together with wrappers around the MARC4J library, for converting the original metadata into MarcXChange as an intermediate format, and latter on into the DIGMAP format (essentially an extended Dublin Core profile).

Like in most real-world collections, the KBR records exhibit a number of deficiencies that require correction. Moreover, although most of the mappings were particularly trivial to declare once the data is in MarcXChange, some metadata fields deserved special attention. Since UNIMARC does not specify a structured field for holding coordinates or bounding boxes, the records have geo-referencing information in a general purpose field that contains notes for cartographic materials in free text. We used string processing functions together with a function for reading and creating spatial geometries, in order to convert this information into a GML encoding. The MarcXChange representation gives the geo-referencing information as shown below:

```
<mx:datafield tag="206" ind1=" " ind2=" "><mx:subfield code="a">
Germa[nia] commu[nia], [1]=[37 mm] (E 4°47'-E 6°01'/N 51°08'-N 50°27')
</mx:subfield></mx:datafield>
```

The user-defined XSLT function given next shows how to convert the data into GML:

```
<function xmlns="http://www.w3.org/1999/XSL/Transform" name="foo:exf">
<param name="n1"/>
<variable name="s1" value="$n1//mx:datafield[@tag=206]/mx:subfield[@code=a]/text()"/>
<value-of select="gis:toGML(string-before(string-after($s1, " (", ")"))"/>
</function>
```

Currently, DIGMAP contains over 40,000 bibliographic records from many different sources. We used XML processing with the help of the extension functions for converting records into the DIGMAP format and for performing several validation and data enrichment tasks (e.g. language identification, consolidation of author names [17] or calling the geoparser service to geo-reference the textual contents when no explicit geo-references are given in the metadata records). The processing of the bibliographic records would be very hard to express using standard XSLTs alone.

## 4.2 Integrating Gazetteer Data

A particular challenge within DIGMAP was related to integrating information from several existing gazetteers (e.g. *geonames.org* and many other online sources) into the common repository of our Web Gazetteer Service [10], which is based on the data model proposed for the Alexandria Digital Library Gazetteer [16]. The XQuery given next illustrates a screen-scraping procedure for extracting data from an HTML page. The data from the Bulgarian Antarctic Gazetteer<sup>21</sup> is extracted from the Web page and converted into the XML format used in the DIGMAP gazetteer service.

The HTML data is first converted to XHTML through an extension function. XPath expressions are then used to retrieve the textual data inside specific formatting elements. Extension functions for spatial data processing and finally used to convert the textual representation for the geographic footprints of the gazetteer features (e.g. *62°36'25" S; 60°12'50" W*) into GML centroid points and bounding boxes.

<sup>21</sup> <http://www.geocities.com/apcbg/toponyms.htm>

```

declare namespace x = "http://www.w3.org/1999/xhtml";
declare namespace gp = "http://www.alexandria.ucsb.edu/gazetteer";

let $url := "http://www.geocities.com/apcbg/toponyms.htm"
let $data := misc:htmltoxml($url)
let $name := $data//x:tr[./x:a[@href]]/x:td[1]**/replace(text(),'','\n',' ')
let $coord := $data//x:tr[./x:a[@href]]/x:td[3]**/replace(text(),'','\n',' ')
let $id := $data//x:tr/x:td[1]/x:a[@href][starts-with(.,'#')]
let $altn := for $k in (1 to count($id))
let $aux := $data//x:p//x:a[@id=substring($id[$k],2)]//x:strong
return $aux//child::text()[2][string-length(.)>1]/replace(.,'\n','')

let $result := for $k in (1 to count($id)) return
<gp:gazetteer-standard-report>
<gp:identifier>{concat($url,$id[$k])}</gp:identifier>
<gp:place-status>current</gp:place-status>
<gp:display-name>{$name[$k]}</gp:display-name>
<gp:names><gp:name primary="false" status="current">{$altn[$k]}</gp:name></gp:names>
<gp:bounding-box>{gis:envelope($coord[$k])}</gp:bounding-box>
<gp:footprints><gp:footprint>{gis:centroid($coord[$k])}</gp:footprint></gp:footprints>
<gp:classes><gp:class primary="true">glacier features</gp:class></gp:classes>
</gp:gazetteer-standard-report>

return
<gp:gazetteer-service><gp:query-response>
<gp:standard-reports>{$result}</gp:standard-reports>
</gp:query-response></gp:gazetteer-service>

```

XML processing is used extensively in our gazetteer, for extracting and integrating data from external sources (e.g. KML files) and for consolidating gazetteer entries.

### 4.3 Mapping Results from the Geoparser into SpatialML, geoRSS and KML

The DIGMAP geoparser [9] is a service that takes text as input, extracts the names for places and time periods, associates these names to identifiers in the DIGMAP gazetteer, and outputs the results in an XML format based on OGC's draft specification for a Web GeoParser Service<sup>22</sup>. Through XSLTs using the proposed extension functions, we support the conversion of the main geoparser format into other well-know formats, namely SpatialML, geoRSS and KML. These are relatively simple mappings that mostly deal with converting the XML tree structure. However, the extension functions facilitate the converting of the spatial information (e.g. the original format uses GML to encode general geometries, whereas SpatialML uses a string representation and geoRSS uses centroid points), as well as the mapping of thesaurus terms (e.g. the original format associates placenames to types from the ADL Feature Type Thesaurus<sup>23</sup>, whereas SpatialML uses a different type hierarchy).

### 4.4 Computational Performance

The time required to perform a transformation is dependent on a number of variables, chief among them the HTTP calls to get data from external services, the size and complexity of the input XML documents and the amount of processing specified by the transformation operations. To give a rough order of magnitude, mapping an XML file of 22 KBytes with 93 elements containing output from the DIGMAP geoparser into the SpatialML format requires approximately 1 second on a 2GHz Intel Core 2 Duo MacBook. Converting the 144 KBytes HTML file for the Bulgarian Antarctic Gazetteer takes around 3 seconds on the same hardware.

<sup>22</sup> [http://portal.opengeospatial.org/files/?artifact\\_id=1040](http://portal.opengeospatial.org/files/?artifact_id=1040)

<sup>23</sup> <http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/FTT2HTML/>

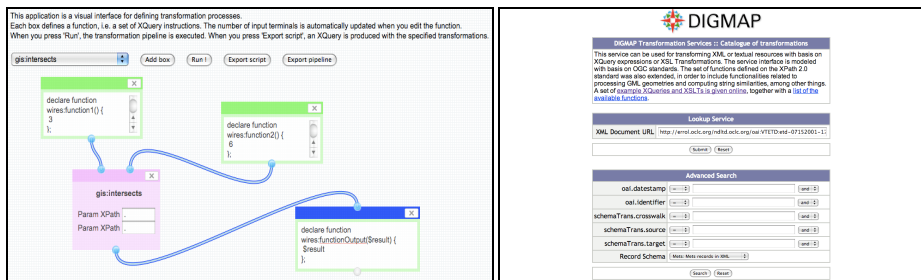
Instrumentation revealed that a significant portion of the time taken up in the transformations is related to the marshalling and unmarshalling operations of GML and KML elements. Saxon is reasonably fast and it can reliably process whatever data fits into main memory. However, it is not an XML database system, and neither does it attempt to be one. If similar extension mechanisms are to be provided over a full-featured XML database, there are many interesting technical challenges, as well as significant room for optimization, in what regards data indexing, processing join operations and streamlining the processing of the XML documents.

## 5 Conclusions and Future Work

In automating DIGMAP's functionalities for collection building and service composition, a need arose for defining XML schema transformations. Specifying these mappings is a complex problem that is largely, but not entirely, declarative in nature. Complex computations are often required, for instance when reasoning with spatio-temporal data or dealing with problems of inconsistency and incompleteness.

To address this need, we studied the use of extension functions together with the XSLT and XQuery languages, adding new predicates to the XPath 2.0 function library for processing spatio-temporal information and for data integration in general. The syntax of XSLT and XQuery is not affected and the XML tree manipulations can still be made through XPath, while the extension functions provide the required functionality that is not directly supported in the standard XML processing APIs.

Our experience with this approach has shown that it is a good match for the problem at hand, enabling users to combine and match XML data in natural, straightforward, seamless, effective and standard-compliant manners. We believe that the proposed extensions can indeed have many applications in spatio-temporal data integration.



**Fig. 1.** The visual interface for specifying mappings that is currently under development

Currently ongoing work addresses the development of a Web-based visual interface for specifying data transformations, similar in style to that of the Web data integration tool Yahoo! Pipes<sup>24</sup> – see Figure 1. Developers of transformations must currently be fluent in the syntax of XSLT or XQuery. The idea is to hide the complexity of these languages from the users who are less familiar with them, by providing

<sup>24</sup> <http://pipes.yahoo.com/>

visual mechanisms for combining simple commands together in order to create a pipeline of transformations. The visual compositions can then be saved, exported as an XQuery or XSLT, or executed directly from the Web-based interface.

Other challenges for future work concern the complete support for GML 3.0 (e.g. handling geographic features that evolve through time) and the development of better functionalities for mapping thesauri (e.g. GML's dictionary components) [1,7].

## References

1. Doerr, M.: Semantic Problems of Thesaurus Mapping. *Digital Information* 1(8) (2001)
2. Godby, C.J., Young, J.A., Childress, E.: A Repository of Metadata Crosswalks. *D-Lib Magazine* 10(12) (2004)
3. Janée, G., Frew, J.: A hybrid declarative/Procedural metadata mapping language based on python. In: Rauber, A., Christodoulakis, S., Tjoa, A.M. (eds.) *ECDL 2005. LNCS*, vol. 3652, pp. 302–313. Springer, Heidelberg (2005)
4. Hillmann, D., Dushay, N., Phipps, J.: Improving Metadata Quality: Augmentation and Re-combination. In: *International Conference on Dublin Core and Metadata Applications* (2004)
5. Kepsner, S.: A Simple Proof for the Turing-Completeness of XSLT and XQuery. In: *The 2004 Conference on Extreme Markup Languages* (2004)
6. Manghi, P., Simeoni, F., Lievens, D., Connor, R.: Hybrid Applications over XML: Integrating the Procedural and Declarative Approaches. In: *The 4th ACM CIKM Workshop on Web Information and Data Management* (2002)
7. Miles, A., Matthews, B.: Inter-Thesaurus Mapping. *SWAD-E Deliverable 8.4* (2001)
8. Guan, J., Zhu, F., Zhou, J., Niu, L.: GQL Extending XQ to query GML documents. *Geo-Spatial Information Science* 9(2) (2006)
9. Martins, B., Manguinhas, H., Borbinha, J.: Extracting and Exploring Semantic Geographical Information from Textual Resources. In: *The 2nd IEEE International Conference on Semantic Computing, ICSC* (2008)
10. Manguinhas, H., Martins, B., Borbinha, J., Siabato, W.: A Geo-Temporal Web Gazetteer Service Integrating Data From Multiple Sources. In: *The 3rd IEEE International Conference on Digital Information Management, ICDIM* (2008)
11. Chung, W., Park, S.-Y., Bae, H.-Y.: An Extension of XQuery for Moving Objects over GML. In: *The IEEE International Conference on Information Technology, ITCC* (2004)
12. Lehto, L.: Real-Time Content Transformations in a Web Service-Based Delivery Architecture for Geographic Information. PhD, Helsinki University of Technology (2007)
13. Allen, J.F.: Time and time again: The many ways to represent time. *International Journal of Intelligent Systems* 6(4) (1991)
14. Novatchev, D.: Higher-Order Functional Programming with XSLT 2.0 and FXSL. In: *The 2006 Conference on Extreme Markup Languages* (2006)
15. Keith, C.: Using XSLT to manipulate metadata. *Library Hi Tech* 22(2) (2004)
16. Hill, L.L., Frew, J., Zheng, Q.: Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib* (1999)
17. Freire, N., Borbinha, J., Martins, B.: Consolidation of References to Persons in Bibliographic Databases. In: Buchanan, G., Masoodian, M., Cunningham, S.J. (eds.) *ICADL 2008. LNCS*, vol. 5362, pp. 256–265. Springer, Heidelberg (2008)



# Sentiment Classification of Movie Reviews Using Multiple Perspectives

Tun Thura Thet, Jin-Cheon Na, and Christopher S.G. Khoo

Wee Kim Wee School of Communication and Information  
Nanyang Technological University  
31 Nanyang Link, Singapore 637718  
{ut0001et,tjcna,assgkhoo}@ntu.edu.sg

**Abstract.** This study develops an automatic method for in-depth sentiment analysis of movie review documents using information extraction techniques and a machine learning approach. The analysis results provide sentiment orientations in multiple perspectives, each focusing on a specific aspect of the reviewed entity. Sentiment classification in multiple perspectives can provide more comprehensive sentiment analysis for applications like sentiment ranking and rating. By utilizing information extraction techniques such as entity extraction, co-referencing and pronoun resolution, the review texts are segmented into sections where each section discusses particular aspect of the reviewed entity. For each section of sentences, Support Vector Machine (SVM) using vectors of terms is applied to determine sentiment orientation toward the target aspect. In our exploratory study, we focus on the sentiment orientations toward overall movie, movie directors and casts in the movie. The experimental results prove the effectiveness of the proposed approach for sentiment classification of movie reviews.

**Keywords:** Sentiment Classification, Movie Review Documents, Information Extraction.

## 1 Introduction

The emergence and wide use of digital libraries and web portals impose many new challenges. It is increasingly important to innovate new ways of finding information and locating online documents. Digital libraries are about new ways of dealing with knowledge and they essentially change the way information is used in the world [13]. Researchers are also considering the problem of organizing and searching digital objects, not just by standard metadata fields but by sentiment analysis.

Recently, researchers have been working beyond binary classification of positive and negative sentiments which predicts an overall sentiment of a review document. Sentiment ranking, which could be a standard feature in future digital libraries, is far more challenging than binary classification. It requires more advanced processing and in-depth analysis of review texts. For instance, when looking at reviews of music for their sentiments, the opinions in review texts cover not only the overall sentiment but also many specific aspects such as vocals, lyric, recording quality, catchiness, creativity and

so on. Sentiment analysis of review items such as music and movie cannot be achieved effectively by classifying overall sentiment orientations. More in-depth sentiment analysis is essential and various aspects in a review document should be considered. For example, a reviewer may like some aspects of a movie but not all. Therefore, applications like comprehensive ranking of movie reviews by sentiment do require in-depth analysis of opinions toward all important aspects which may or may not be interdependent.

We investigate movie review documents in this study mainly due to a large amount of review data available on the Internet and the challenging nature of such reviews. Movie reviews are believed to be more challenging than other reviews such as product reviews [12]. Separate consideration for different aspects of movie reviews such as opinions on cast, director, storyline, or animation provides better sentiment analysis of a movie. In order to process multiple perspectives for sentiment classification, review texts are segmented into different sections so that they can be analyzed and processed independently. For this purpose, information extraction techniques such as entity annotation, co-referencing and pronoun resolution are employed. Then the automatically separated sections are also reviewed by two manual coders in order to verify the effectiveness of our automatic segmentation approach. The independent coders also read sentences in each movie reviews, and manually code a sentiment orientation toward specific aspects. Intercoder reliability is verified since the manually coded data is used for the supervised machine learning to determine sentiment orientations toward each aspect of the movies.

In the following sections, section 2 discusses related works, sections 3 and 4 present our approaches for sentiment classification and error analysis, and finally section 5 discusses future work and conclusion.

## 2 Related Works

In recent years, many researchers have been focusing their attentions on sentiment classification and developing new methods for sentiment and subjectivity analysis. Most research in automatic sentiment classification seeks to develop supervised machine learning approaches for classifying new documents or document segments by sentiments based on a set of training documents that have been classified by domain experts. Some basic concepts and methods for automated text classification have been discussed by Sebastiani [9]. A challenging aspect of sentiment classification that distinguishes it from traditional topic-based classification is that while topics are often identifiable by keywords alone, sentiment can be expressed in a more delicate expression. For example, the sentence “*who would vote for this presidential candidate?*” contains no single word that is obviously negative. Sentiment classification requires more understanding than the usual topic-based classification.

Turney [12] proposed an unsupervised learning algorithm for classifying review documents as recommended (thumbs up) or not recommended (thumbs down) by calculating the mutual information between the given phrase and the word “*excellent*” minus the mutual information between the given phrase and the word “*poor*”. This approach calculates the semantic orientation of a given phrase by examining its similarity to a positive reference word with its similarity to a negative reference word. The

method was experimented with automobile reviews and movie reviews, and resulted accuracies of 84% and 66% respectively.

Pang et al. [7] used standard machine learning techniques for classification of documents by overall sentiment. They chose movie reviews as dataset for the experiment and employed three machine learning methods: Naïve Bayes, maximum entropy classifications and support vector machines (SVM). The results produced by machine learning methods are better in comparison to the human generated baselines. In terms of relative performance, SVM tends to do the best and Naive Bayes tends to do the worst, although the differences are not significant. Yi et al. [14] proposed another method of sentiment analysis using natural language processing techniques to extract positive and negative sentiments for specific subjects from a document, instead of classifying the whole document into positive or negative. They used semantic analysis with a syntactic parser and sentiment lexicon. The prototype system achieved good results for Web pages and news articles. Another work by Pang and Lee [8] proposed a novel machine-learning method by applying text categorization techniques only to the subjective portions of the document. They examined the relation between subjectivity detection and polarity classification, showing that subjectivity detection can compress reviews into much shorter extracts that still retain polarity information at a level comparable to that of the full review. They proved that employing the minimum-cut framework results in the development of efficient algorithms for sentiment analysis. Utilizing contextual information via this framework can lead to statistically significant improvement in polarity-classification accuracy.

Hu and Liu [3] worked on summarization of customer reviews using data mining and natural language processing methods. However, their approach does not deal with pronoun resolution which is a rather complex and computationally expensive problem in natural language processing. Li et al. [6] proposed a new method for mining and summarization of movie reviews. The method involves in finding concrete feature-opinion pairs. For example in the sentence “*the sound effects are excellent*” the feature word is “*sound effects*” and the opinion word is “*excellent*”. The feature classes for summarization include overall, screenplay, character design, vision effects, music and sound effects, special effects, producer, director, screenwriter, actor and actress, people in charge of music and sounds, and people in charge of techniques of movie-making. The goal of this approach is to find feature-opinion pairs and to identify their polarity and the feature classes of the opinions in order to produce a structured sentence list as the summary. A study by Snyder and Barzilay [11] proposed another approach to address the problem of analyzing multiple related opinions in a text by providing a set of numerical scores, one for each aspect. It analyses meta-relations between opinions such as agreement and contrast in order to guide the prediction of individual rankers. The result of the experiments proves empirical improvements over both individual rankers and a state-of-the-art joint ranking model.

### 3 Sentiment Classification

We conducted this study with a dataset of 876 movie review documents: 438 positive and 438 negative. The movie review documents are harvested from a movie review site ([www.reelviews.net](http://www.reelviews.net)) using a web crawler. They are movie reviews by a movie

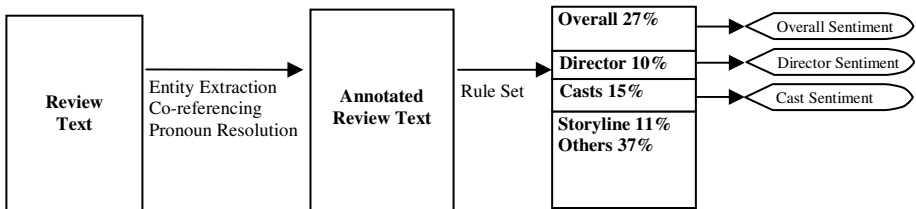
critic and most of them have about five to seven paragraphs of texts. The metadata of review documents such as movie title, reviewer, director, and cast names are also extracted by parsing the HTML codes of review documents. The ratings are extracted and normalized so that different rating scales from various Web sites can be combined and processed together later on. For example, 3 out of 10 stars and 3 out of 5 stars are normalized and stored as 0.3 and 0.6 respectively.

An average length of review documents in our dataset is around 30 sentences and they discuss various aspects of the movie. In general, these sentences contain overall opinion as well as specific opinions about directors, casts, or other aspects.

### 3.1 Sentence Tagging

We have applied information extraction techniques to tag sentences in the movie reviews. The title, cast, and director names of the movie are first extracted since they are important indicators in determining specific aspects of the movie in current sentences. Co-referencing [1] is a technique used for determining which named entities have the same referent. For example, it is to determine if “*Jim Carrey*” and “*Carrey*” refer to the same entity. Another technique used is known as Pronoun or Anaphora Resolution which addresses the problem of resolving what a pronoun, or a noun phrase refers to in the text. For example, it is to determine which entity the pronoun “*he*” in the sentence “*he has been quite disappointing*” is referring to. It could be referring to a director or an actor.

In this study, a rule-based approach is used for automatic tagging of sentences into different sections: overall, cast, director, storyline or others. We have developed an application in Java to perform text annotation with the extracted metadata and sentence tagging as illustrated in Figure 1, which also shows the distribution of automatically tagged sentences. We have studied General Architecture for Text Engineering (GATE) [10] for sentence tagging, which has an information extraction system called a Nearly-New Information Extraction System (ANNIE). For the experiments, we used our own sentence tagging tool which implements some of GATE features.



**Fig. 1.** The overall process for sentence tagging

The rule set used for automatic sentence tagging is shown in Figure 2. Each sentence is compared with conditions in the rules, and if there is a match, the corresponding tag is assigned to the sentence. Priorities of the rules are set by arranging them in a specific order. For instance, the rule in line 3-4 fires if a sentence matches a specific pattern for sentence tagging. These patterns tend to be domain specific and depend on reviewers’ writing styles. In our movie reviews, sentences having actor/actress names

in brackets before a character name represent storylines. For example, “*Katie Burke (Katie Holmes) is bringing to a close a highly satisfactory run at a prestigious college*” is tagged as storyline. The rules in Figure 2 are self explanatory and they make use of gazetteers (lists of terms), co-referencing, and pronoun resolutions for the entities.

```

1   FOR each review document
2     FOR each sentence in a review document
3       IF specific patterns recognized // e.g. a pattern for storyline is found
4         sentence tag = corresponding tag // e.g. assign storyline-tag
5       ELSE IF sentence contains director names OR co-references
6         sentence tag = director
7       ELSE IF sentence contains cast names OR co-references
8         sentence tag = cast
9       ELSE IF sentence contains director pronouns
10        sentence tag = director
11      ELSE IF sentence contains cast pronouns
12        IF sentence contains this movie, reviewer-audience OR casting terms
13          sentence tag = cast
14        ELSE
15          sentence tag= storyline
16        ELSE IF sentence contains this movie OR reviewer-audience terms
17          sentence tag = overall
18        ELSE
19          sentence tag = others
20      END FOR
21    END FOR

```

**Fig. 2.** Rules for automatic sentence tagging

Table 1 shows some sample entries of gazetteers which are used by the rule engine for the condition in line 12. The terms for “*this movie*” and “*reviewer-audience*” are prepared manually by scanning through the movie reviews. The terms for “*casting*” are prepared automatically by selecting n-grams which appear more in cast sentences in the training data and have high mutual information (MI) in order to separate casting sentences from the storylines. The terms appearing in the storyline sentences usually have low MI because the storylines cover a variety of topics. The selection of terms can be more refined by using more training sentences and adjusting the threshold level to extract the terms.

**Table 1.** Sample entries of gazetteers

This Movie Terms	Reviewer-Audience Terms	Casting Terms
this movie	I	excel perform
this film	Me	transform
the movie	my	tribute
the film	You	has/have created
movies of this kind	Your	subtle
(9 entries)	(9 entries)	(45 entries)

To verify the accuracy of automatic tagging, two manual coders were asked to highlight sentences which were wrongly tagged as casts or directors. The manual and automatic sentence tagging had about 90% agreement and automatically tagged sentences were used for the experiments. The two coders also read cast and director sentences in each movie review, and manually classified the sentiment orientation toward the subjects. Their decisions were compared to verify intercoder reliability.

We used Cohen’s kappa coefficient in order to measure the agreement between the two independent coders who classified the cast and director sentences in each movie review into one of the following sentiment classes: positive, negative, neutral, and not applicable.

The equation for Cohen’s kappa, CK, is:

$$CK = \frac{\text{Pr (agree)} - \text{Pr (chance)}}{1 - \text{Pr (chance)}} \quad (1)$$

where Pr(agree) is the relative observed agreement among coders and Pr(chance) is the probability that agreement is by chance.

In our experiment, the intercoder agreement using Cohen’s kappa coefficient was 0.74 which is considered as a good agreement [2]. The conflicting labels by the two coders were reviewed and manually re-classified by one of the authors and these manually classified sentiment labels for casts and director sentences were used as answer keys for the supervised machine learning approach which we will discuss in the next section. Table 2 shows intercoder agreement between two coders. The numbers in bold indicate agreement between two coders, and the other numbers indicate disagreement.

**Table 2.** Intercoder agreement between two coders

		Coder A			
		Positive	Negative	Neutral	NA
Coder B	Positive	<b>630</b>	40	42	22
	Negative	54	<b>600</b>	31	13
	Neutral	16	22	<b>27</b>	13
	NA	18	16	9	<b>199</b>

### 3.2 Overall Sentiment

The star ratings given by the reviewer are used as overall sentiment orientation for the movies. The reviews are rated with a scale of up to five stars. Those with two stars and below are considered as negative reviews while those with three stars and above are considered as positive reviews.

For the experiments, SVM [4] is used as a supervised machine learning algorithm. The review text is converted into bags of terms (called document vectors), which are stemmed using Porter’s stemming algorithm [5] after removing the stop words. Negation of sentiment words sometimes requires n-gram terms having more than three words. Although “not”, “n’t” and “never” are commonly used negation terms which are usually placed next to adjectives or verbs, some usages of negation are not very straight forward. For example, in the sentence “*the film ceases to be boring when she comes on screen*”, the adjective “boring” is negated by the phrase “ceases to”. Our approach uses a gazetteer of negation terms and phrases to represent negated terms as additional features for the machine learning.

According to automatic document tagging, about 27% of the sentences from the review texts are tagged as ‘overall’ and they are used for predicting the sentiment orientations of the movie reviews. We have conducted a 3-folds cross validation for this experiment and although the method uses 27% of the review texts, it achieves relatively high accuracies.

**Table 3.** Accuracies of overall sentiment classifications

ID	Term Weighting		Adjective Terms only	Stemming	Negation	Removal of Stop words	Accuracy
	Presence	Frequency					
1	Y						88.31%
2		Y					82.01%
3	Y		Y				76.38%
4	Y			Y			88.88%
5	Y			Y	Y		89.00%
6	Y			Y	Y	Y	<b>90.48%</b>

The table 3 shows that using stemmed and negated terms as features improves the accuracies of the overall sentiment classification. It is also observed that using a standard list of stop words for feature reduction is not ideal. In fact, the accuracies of sentiment classification drop noticeably when a standard list of stop words is used. Unlike Information Retrieval applications, words like negations and prepositions carry some meanings for sentiment classification. Therefore, we have used a list of selective stop words for our experiments. The columns in table 3 represent various document feature options for the experiments: using presence or frequency for term weighting, adjective terms only from the full text, word stemming, handling of negation terms, and removal of stop words. As shown in the table, using frequency as term weighting or only adjective terms as features reduces the accuracies by a few percentages.

### 3.3 Sentiment toward Directors

The manually tagged sentiment labels are used as answer keys for the director sentences in movie reviews. The experiment shows that a sentiment orientation toward a director can be different from overall sentiment orientation toward the movie. Sentiment toward the director could be positive or neutral while overall sentiment toward the movie is clearly negative.

For example, the movie “*bad girls*” is rated half star out of 5 stars ([http://www.reelviews.net/movies/b/bad\\_girls.html](http://www.reelviews.net/movies/b/bad_girls.html)), but the review contains positive sentiment about the director as it says “At least director Jonathan Kaplan had the good sense to employ a competent cinematographer”. In some reviews, rating shows positive sentiment toward the overall movie, but the sentiment toward the movie director is clearly negative or neutral.

According to automatic sentence tagging, only 10% of the sentences from the review texts are about the director and they are used for predicting the sentiment orientation toward the director in the experiment. Similarly, SVM is used as a supervised machine learning approach for the experiment.

We have conducted a 3-fold cross validation and the results show an accuracy of 75.54% which is well above 50% accuracy of random guessing. Director sentences cover only 10% of the review texts and it still produces a good result for sentiment classification for the movie director. Using stop words for feature reduction and using stemmed and negated terms as feature improves the accuracies of sentiment classification of director sentences but not when using frequency as term weighting and only adjective terms as features. In Tables 4 and 5, the column IDs (1 to 6) indicate the document feature options (IDs) used in Table 3.

**Table 4.** Accuracies of director sentiment classifications

ID	1	2	3	4	5	6
Accuracy	73.25%	69.10%	69.81%	74.82%	75.25%	<b>75.54%</b>

### 3.4 Sentiment toward Casts

Similarly, we used the manually tagged sentiment labels as answer keys for the cast sentences in movie reviews. In some cases, a sentence can contain different sentiment orientations for the movie and the cast. For example, the sentence “*Actress Julie Delpy is far too good for this movie*” carries positive sentiment toward the cast while there is some negativity toward the movie.

According to automatic sentence tagging, only 15% of the sentences from the review texts are about casts and they are used for predicting the sentiment orientation toward the casts in the experiment. Similarly, SVM is used as a supervised machine learning approach for the experiment. We have conducted a 3-fold cross validation and the results show an accuracy of 78.74%.

**Table 5.** Accuracies of cast sentiment classifications

ID	1	2	3	4	5	6
Accuracy	76.40%	70.38%	73.17%	77.71%	77.86%	<b>78.74%</b>

Similarly, the accuracies of sentiment classification of cast sentences are improved when using stop words for feature reduction and using stemmed and negated terms as features but reduced when using frequency as term weighting and only adjective terms as features.

## 4 Error Analysis

We have analyzed the errors encountered in both the automatic sentence tagging and the sentiment classification of our approach. There are about 10% discrepancies in the automatic tagging of directors and casts sentences and about 20% errors in the sentiment classification for overall, director and cast sections.

The following is an example sentence which is wrongly tagged as a director sentence by our automatic approach. The sentence does mention the director name, “Peter Howitt”; however, it is obviously discussing about the cast.

*Meanwhile, Douglas McFerran, who gave a brilliant supporting performance in Peter Howitt’s Sliding Doors, plays the tough head of NURV security.*

The following is an example sentence about the director which is wrongly classified as negative sentiment by the machine learning approach. The words ‘worst’ and ‘pedestrian nature’ may appear negative but they represent positive meaning with contextual words. For instance, the word “justice” changes the sentiment orientation of “pedestrian nature” to positive. This is one of the limitations of the machine learning approaches which use just bags of terms. Deep syntactic analysis and pattern learning instead of just using bags of terms will be helpful in addressing such problems.



*Todd Graffs script is television-quality writing at its worst, and the direction by Ken Kwapis, who made the better Dunston Checks In, does justice to the screenplays pedestrian nature.*

In the following example sentence, there is a mixture of positive and negative sentiment about two different casts; Tony Danza and Joseph Gordon-Levitt. This can be a difficult case not only for the machine learning approach but also for human coders to decide whether it is positive, negative or neutral.

*Suffice it to say that Tony Danza gives one of the most impressive performances, and young Joseph Gordon-Levitt has serious credibility problems.*

As observed, some of the cases encountered are rather complex for automatic approaches to tag or classify them very accurately unless more advanced natural language processing techniques are employed.

## 5 Discussion and Conclusion

Sentiment analysis of review documents should consider multiple perspectives when there are sentiments towards different aspects of the reviewed entity. Our proposed method segments the review texts into sections based on their target aspects before applying the supervised machine learning approach using SVM. Our proposed method makes use of information extraction techniques such as entity extraction, co-referencing and pronoun resolution. The experimental results show that using selective stop words for feature reduction and using stemmed and negated terms as features for the automatic machine learning approach improves the accuracies of sentiment classification but not when using frequency as term weighting and only adjective terms as features. The best accuracies of sentiment classification for the overall movie, directors and casts are 90.48%, 75.54% and 78.74% respectively.

One of the limitations of this study is that it is only conducted for the movie review documents harvested from a movie review site and seems to be relatively domain specific. More evaluations and experiments are to be carried out with larger datasets, a wider range of reviewers as well as different domains. Another limitation is that we did not employ advanced information extraction techniques which could produce higher accuracy in sentence tagging. For future work, deep syntactic pattern learning and more advanced information extraction tools (such as GATE-ANNIE) are to be employed to analyze multiple perspectives of the movie reviews and provide more comprehensive sentiment analysis.

## References

1. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving GATE to Meet New Challenges in Language Engineering. *Natural Language Engineering* 10(3-4), 349–373 (2004)
2. Byrt, T.: How Good Is That Agreement? *Epidemiology*, vol. 7, p. 561 (1996)
3. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177 (2004)

4. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of 10th European Conference on Machine-learning, Chemnitz, Germany, April 21–24, pp. 137–142 (1998)
5. Jones, K.S., Willet, P.: Readings in Information Retrieval. Morgan Kaufman, San Francisco (1997)
6. Zhuang, L., Jing, F., Zhu, X.-Y.: Movie review mining and summarization. In: Proceeding of the 15th ACM Conference on Information and Knowledge Management, pp. 43–50 (2006)
7. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine-learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 79–86 (2002)
8. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics ACL, Barcelona, Spain, 271-278 (2004)
9. Sebastiani, F.: Machine-learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (2002)
10. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (2002)
11. Snyder, B., Barzilay, R.: Multiple aspects ranking using the good grief algorithm. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL), pp. 300–307 (2007)
12. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics ACL, pp. 417–434. Philadelphia (2002)
13. Witten, I.H., Bainbridge, D.I.: How to build a digital library. Morgan Kaufmann, San Francisco (2003)
14. Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.: Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. In: Proceedings of the Third IEEE International Conference on Data Mining ICDM, pp. 427–434 (2003)

# Scholarly Publishing in Australian Digital Libraries: An Overview

Bhojaraju Gunjal<sup>1,2</sup>, Hao Shi<sup>1</sup>, and Shalini R. Urs<sup>2</sup>

<sup>1</sup> School of Computer Science & Mathematics, Victoria University, Melbourne, Australia  
Bhojaraju.G@gmail.com

<sup>2</sup> Executive Director, International School of Information Management, University of Mysore,  
Mysore, Karnataka, India

**Abstract.** The aim of this paper is to examine current trends in the development of scholarly publishing in digital libraries (DL), with particular reference to Australian Universities. This paper is the result of preliminary study based on the visits made to university libraries in Canberra, Sydney and Melbourne. The objective of this paper is to analyze the various aspects of Institutional Repositories (IR) in Digital Libraries (DL) and its development in Australia.

This article showcases the latest trends in DL which have adopted various technologies (i.e. open source or commercial or mixed) adopted in Australia. The government plays a prominent role in development of DL and supports scholarly publication through various digitisation projects. The paper concludes that there is a need for change in policy on mandatory submission of theses to reduce the cost, time and manpower efforts towards conversion.

**Keywords:** Australia, Digital Libraries, Scholarly Publishing, Institutional Repositories, Open Source, Knowledge Organisation System, Search Interface.

## 1 Introduction

Scholarly publishing (SP) has got a great value now-a-days through open access initiatives (OAI) and Institutional repositories (IR) which are being used to capture original research and other intellectual property generated by university members. The Australian government plays an important role in supporting various digitisation projects in universities. A review of recent developments in SP, with a focus on OAI and IR in Australia is provided in this paper. This study involved visiting and interacting with concerned professionals with a view to understand the issues and challenges faced by the universities.

Though universities initiated the Electronic Theses and Dissertations (ETD) programs to begin with, now many seem to be transiting towards a broader program of IR.

## 2 Scope

The scope of this paper is limited to Canberra, Sydney and Melbourne university libraries only. Universities are randomly selected, visited and analysed. These case

studies do not depict the same aspects for other parts of Australian libraries. List of libraries visited are mentioned at the acknowledgements section.

### 3 Institutional Repositories in Australian Universities

The development of institutional repositories (IR) emerged as a new strategy that allows universities to preserve their scholarly publications along with other digital materials. IRs have adopted various DL technologies (both open source or commercial or mixed). Our study provides an overview of the same [1].

The details on IR in Australian Universities can be found at *Appendices A1 and A2*. Following are the highlights of few cases to show case the IR scenario in Australian universities.

- Most of the universities use Open Source (OS) tools like EPrints, DSpace, Fedora where as few universities use commercial tools such Digitool along with OS or commercial user interfaces (UI).
  - OS/OS >> Fedora/Fez (e.g. Deakin University)
  - OS/Commercial >> Fedora/Vital (e.g. La Trobe, Monash)
- All universities are OAI-PMH compliant and support harvesting to make their IR visible and accessible through a global network of services such as OAIster, ARROW, Google Scholar and Internet search engines.
- A few universities have separate team to handle Copyright process (e.g. University of Melbourne, UNSW).

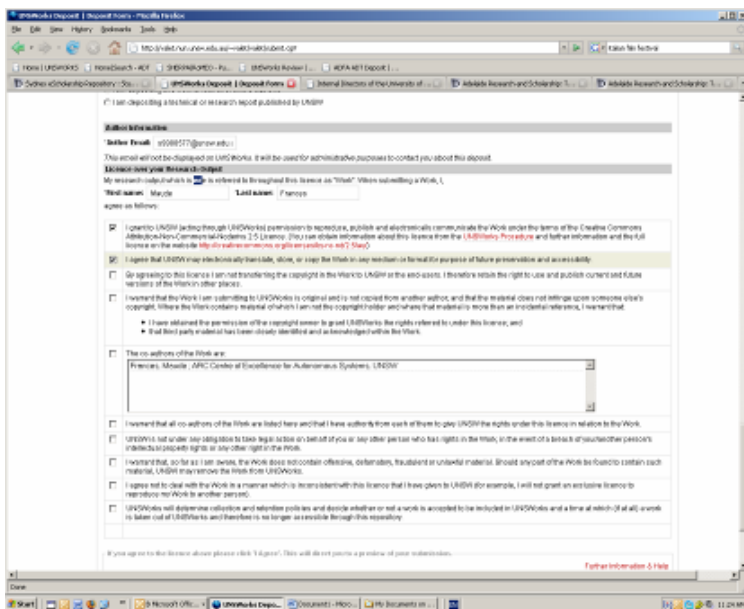


Fig. 1. Copyright agreement Page

## 4 Related Digital Library Projects

### 4.1 ADT Program

The aim of the Australasian Digital Theses (ADT) program is to establish a distributed database of digital versions of theses produced by the postgraduate research students at Australian universities. The theses will be available worldwide via the web. The idea behind the program is to provide access to, and promote Australian research to the international community. The ADT concept was an initiative of 7 Australian universities in association with the Council of Australian University Librarians (CAUL) [2]. Presently, 40 CAUL members have joined this ADT program.

### 4.2 ARROW Project

The Australian Research Repositories Online to the World (ARROW) project has been very successful in providing tools to enable accessibility and discoverability of research from institutional repositories. The NLA announced the new ARROW Discovery Service website and is operational at <http://search.arrow.edu.au>. Its new interface aggregates the institutional repositories of all Australian universities. Currently it contains more than 216,663 records harvested from 23 university repositories and 12 other research collections, including the Australasian Digital Theses program, and several e-journals [3].

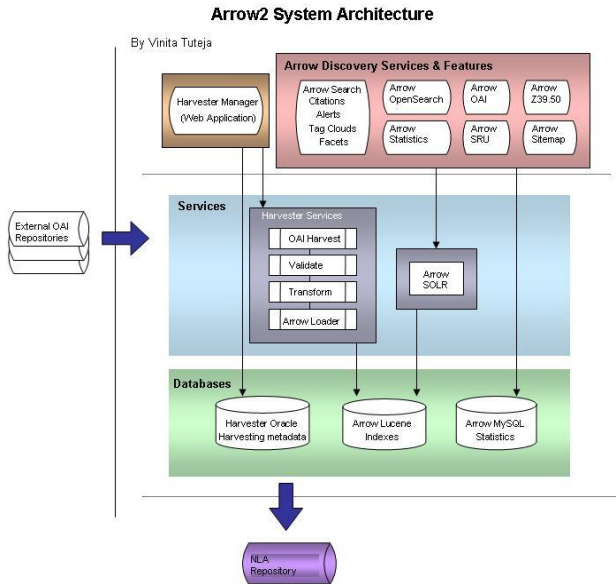


Fig. 2. ARROW System Architecture (Source: National Library of Australia)

The Discovery service provides access to the research outputs of Australian universities: many are unpublished theses, preprints or postprints, as well as published journal articles, images, working papers and technical reports. The NLA has introduced several improvements to the Service including a more intuitive search interface, including faceted browsing & tag clouds and more access to statistics, including tables of the most popular authors and institutions.

Metadata harvesting in ARROW takes place as illustrated in Fig.3. It transforms MARCXML and ETD-MS metadata into qualified Dublin Core for OAI-PMH and internal purposes.

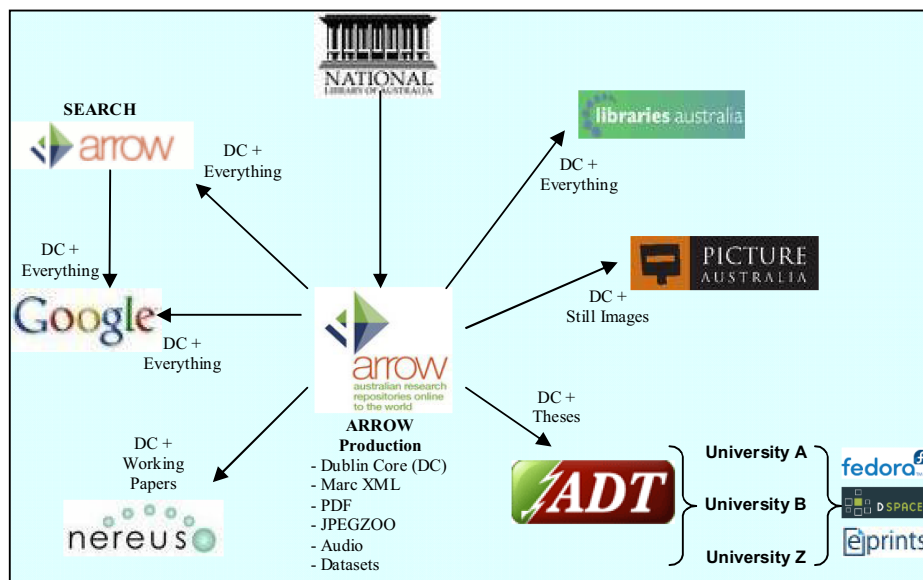


Fig. 3. Metadata Harvesting by ARROW

### 4.3 DART Project

The Dataset Acquisition Accessibility & Annotation e-Research Technologies (DART) project is funded by the Australian Commonwealth Government's Department of Education, Science and Training (DEST), through to the end of 2006. The partners are Monash University (lead institution) in Melbourne, The University of Queensland in Brisbane and James Cook University in Townsville.

The DART project is an ambitious proof-of-concept project to develop tools to support the new collaborative research infrastructure of the future. The project aims to enable researchers and reviewers to access original and analysed data, collaborate around the creation of research outputs, stored publications, plus add content, annotations and notes. It will also look at the collection of large datasets, including the remote control and automated data collection [4].

### 4.4 ARCHER Project

The Australian Research Enabling Environment (ARCHER) project is funded by the Australian Department of Education, Science and Training (DEST) via an SII Grant. The project partners are Monash University (lead institution) in Melbourne, The University of Queensland in Brisbane and James Cook University in Townsville.

The ARCHER has setup dedicated development teams within each partner University to: analyse eResearch data collection; analyse information management needs and requirements, and take special note of existing IT applications and services in each research area. The ARCHER development teams have built upon the prototype software developed by the DART and ARROW projects to produce a robust set of software tools [5].

### 4.5 APSR Project

The Australian Partnership for Sustainable Repositories (APSR) Project aims to establish a centre of excellence for the management of scholarly assets in digital format. APSR is a partnership that aims to promote excellence in building & managing these collections of digital research objects. The Partnership receives Federal Government funding to assist Australian researchers with research information management. To this end, APSR conducts outreach and educational programs and undertakes collaborative development of systems and tools. APSR works closely with research communities, information professionals, technical staff, and higher education policy makers on a series of development projects, surveys, publications, seminars, and training workshops [6].

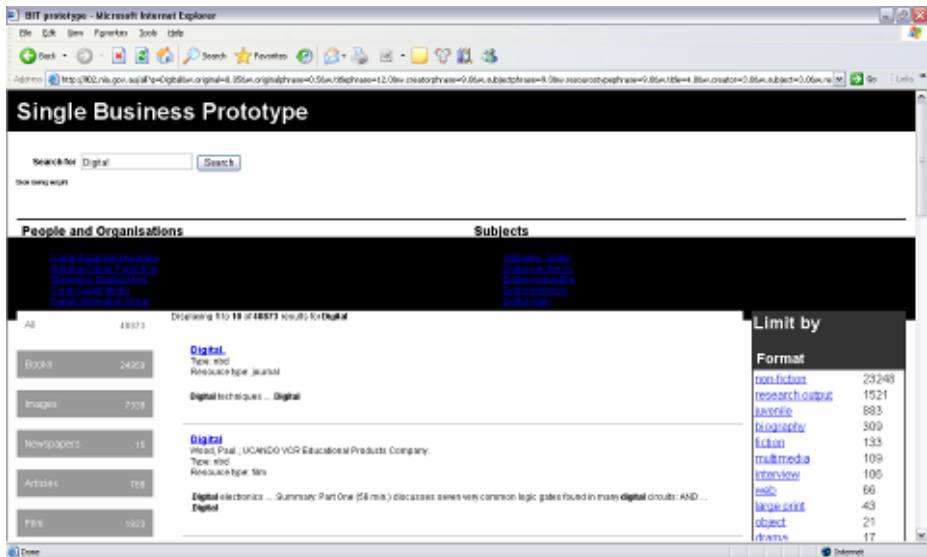


Fig. 4. Search Results in SBP

## 4.6 Single Business Prototype (SBP)

SBP is a search service from NLA. This project is in prototype stage and is expected to be operational by late 2008. In future, SBP may replace ARROW service. This is based on clustering search technology [7].

## 5 Issues and Challenges

The following are some issues and challenges faced by the universities/organisations:

- Copyright issue is one of the biggest challenges for IR development. Some universities have a separate team to handle activities related to this (e.g. University of Melbourne).
- Both *PDF* and *Word* formats are widely accepted. Few universities have collections in MS Excel with macros, images, music & other audio formats (e.g. University of Sydney). Copyright challenge for these materials is still high.
- Only PDF formats are used in IR because of user-friendly and ease-to-use aspects. SGML/XML formats are not used because of complexity, involves training and maintenance.
- Dublin Core Metadata Standards are used. All the records are harvested from the hosting repositories via the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH).
- Some universities are in the process of converting all theses from hardcopies to digital format.
- Most of the universities have made theses submission policy as *by voluntary*. Mandatory submission of theses in digital format is yet to be implemented (e.g. University of Technology Sydney).
- A few universities have mandated students to submit their work in digital copies along with hardcopy (e.g. UNSW).
- All universities are not mandated to collect theses in digital format from their students/staff. This may be one of the hurdles in developing digital libraries in e-scholarly publishing area. This may be due to various reasons viz.
  - due to *copyright* issue (permission perspective)
  - not to make *work public* (students' perspective)
  - information may contain *cultural sensitiveness*
  - may be *funded* by company (due to *commercial use/intellectual property* reasons)
  - *lack of support* from the university to students (Copyright, Technical and general aspects)
- Even though all member institutes are OAI-PMH compliant, some universities use different data structure which poses problems in metadata harvesting. This will affect the search results and show some anomalies in results.
- Some universities follow the policy on restriction access (whole or partial) to theses on web based on the authors' request (e.g. UNSW, La Trobe).



## 6 Conclusion

Most of the universities are migrating from ETD concept to IR because of desire and need to store wide range of resource types other than theses only. Due to management decision, some universities are moving their IR from *open source* to *commercial* tools.

Many universities have not made mandatory submission of theses since the beginning and are trying to convert the collected hardcopies of theses into digital format. For this purpose, library staff needs to search for the authors and seek their permissions for the same to abide copyright rules. This is again time dependent and involves huge cost and manpower effort in this process. Hence, there is a need for change in policy on mandatory submission of theses to reduce the cost, time and manpower effort on conversion.

**Acknowledgments.** The authors would like to thank the Department of Education, Science and Training, Australian Government for the 2008 Endeavour Research Fellowship, School of Computer Science and Mathematics, Victoria University for providing their support of the fellowship in conducting research study in Australia.

Bhojaraju also like to thank all university faculties mentioned below for their extended help and support during his visit to collect the information for his research study viz. University of Technology, Sydney, University of Sydney, University of New South Wales, National Library of Australia, Australian National University, Victoria University, University of Melbourne, LaTrobe University, RMIT University, Monash University and Deakin University.

## References

1. Gunjal, B., Urs, S., Shi, H.: Australian Digital Libraries: An Overview. In: WCECS 2008's International Conference on Education and Information Technology (ICEIT 2008), San Francisco, USA, October, pp. 22–24 (2008)
2. Australasian Digital Theses Program, <http://adt.caul.edu.au/>
3. ARROW Discovery Service Project, <http://www.arrow.edu.au>
4. DART Project, <http://dart.edu.au/>
5. ARCHER Project, <http://archer.edu.au>
6. APSR Project, <http://www.apsr.edu.au>
7. Single Business prototype, <http://1102.nla.gov.au/>
8. EPrints, <http://www.eprints.org/>
9. Encore - Innovative Interfaces, Inc,  
<http://www.encoreforlibraries.com/main.html>
10. A case study detailing the University of Melbourne DigiTool experience,  
<http://www.exlibrisgroup.com/files/CaseStudy/MelbourneDigiTool.pdf>
11. Fedora, <http://fedora.redhat.com/>
12. VITAL, <http://www.vtls.com/products/vital>
13. Fez Wiki,  
[http://dev-repo.library.uq.edu.au/wiki/index.php/Main\\_Page](http://dev-repo.library.uq.edu.au/wiki/index.php/Main_Page)
14. DSpace, <http://www.dspace.org/>

## Appendices

A.1. Summary of Scholarly Publishing tools and its features in Australian Universities

Repository/ Project Name	University/URL	Software Used	User /Search Interface	Search Interface	Comments
EPrints Repository	Victoria University, Melbourne <a href="http://eprints.vu.edu.au/">http://eprints.vu.edu.au/</a>	ePrints [8]	ePrints	Encore by Innovative Inc [9] DigiTool	Based on the management decision, moved from ePrints to the commercial software - DigiTool by Ex Libris [10]
UMER - University of Melbourne ePrints Repository	University of Melbourne, Melbourne <a href="http://www.lib.unimelb.edu.au/eprints/">http://www.lib.unimelb.edu.au/eprints/</a>	DigiTool (Commercial)	DigiTool	• Brief view • Table view • Full view	Site is on hold for public as waiting to seek permissions for copyright material from publishers.
	La Trobe University, Melbourne <a href="http://www.lib.latrobe.edu.au/thesis/index.php">http://www.lib.latrobe.edu.au/thesis/index .php</a>	Fedora [11]	Vital by VTLS (Commercial) [12]	By Browse	In future planning to move to Fedora or DigiTool
	RMIT University, Melbourne	Software by Virginia Polytechnic Institute		Lucene Two views • List View • Icon View	
EPrints	Monash University, Melbourne <a href="http://eprint.monash.edu.au">http://eprint.monash.edu.au</a>	Fedora	Vital by VTLS (Commercial)		
Deakin Research Online	Deakin University, Melbourne <a href="http://www.deakin.edu.au/opendoor/">http://www.deakin.edu.au/opendoor/</a>	Fedora	Fez [13]		Plan to go for DigiTool or Fedora in future
UTSePress Institutional Repository	University of Technology, Sydney <a href="http://epress.lib.uts.edu.au/dspace/">http://epress.lib.uts.edu.au/dspace/</a>	DSpace [14]	DSpace		Plan to implement mandatory submission of theses in digital format
Sydney eScholarship Repository	University of Sydney, Sydney <a href="http://ses.library.usyd.edu.au/">http://ses.library.usyd.edu.au/</a>	DSpace	DSpace		Plan to upgrade to DSpace 1.5
Demetrius	Australian National University, Canberra <a href="http://dspace.anu.edu.au/">http://dspace.anu.edu.au/</a>	DSpace	DSpace		Moved from ePrints to DSpace
UNSWorks/ADT	University of New South Wales, Sydney				
ARROW		Fedora	Vital by VTLS (Commercial)		

A. 2. Search Interface and results screen in different repositories

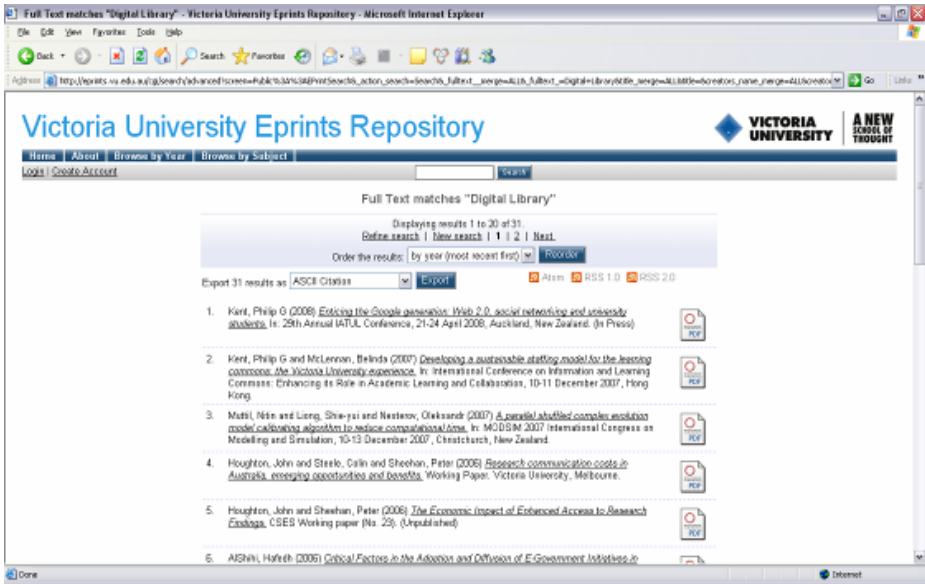


Fig. 5. Victoria University (VU)

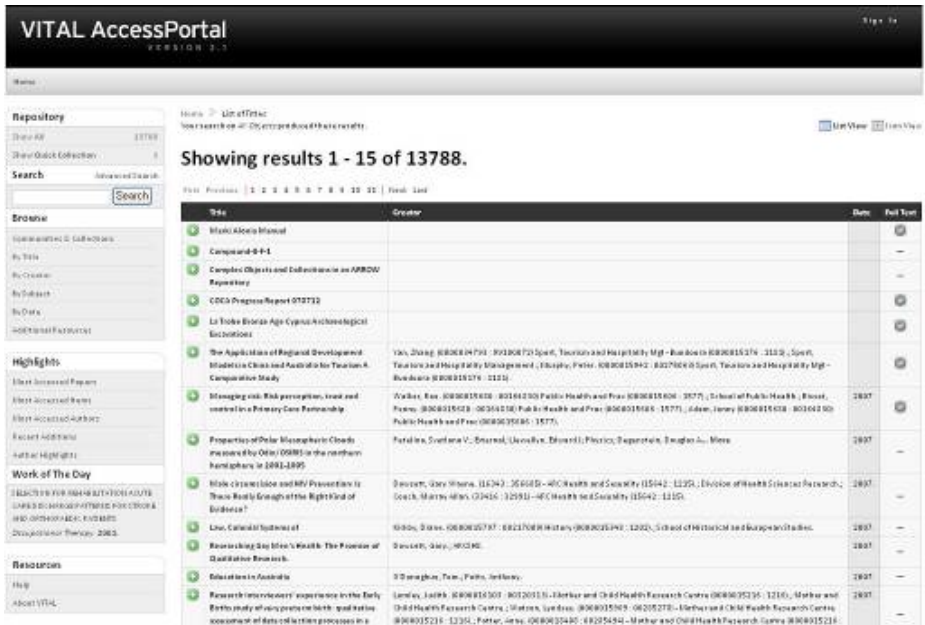


Fig. 6. La Trobe University

# Utilizing Semantic, Syntactic, and Question Category Information for Automated Digital Reference Services

Palakorn Achananuparp, Xiaohua Hu, Xiaohua Zhou, and Xiaodan Zhang

College of Information Science and Technology

Drexel University, Philadelphia, PA 19104

pkorn@drexel.edu, thu@cis.drexel.edu, xiaohua.zhou@drexel.edu,  
xzhang@cis.drexel.edu

**Abstract.** Digital reference services normally rely on human experts to provide quality answers to the user requests via online communication tools. As the services gain more popularity, more experts are needed to keep up with a growing demand. Alternatively, automated question answering module can help shorten the question-answering cycle. When the system receives a new user submitted question, the similarity of the user's request and the existing questions in the archive can be compared. If the appropriate match is found, the system then uses the associated answer to response to such request. Since a question is relatively short and two questions might contain very few words in common, the challenge is how to effectively identify the similarity of questions. In this paper, we focus on the problem of identifying questions that convey the similar information need. That is, our goal is to find paraphrases of the original questions. To achieve this, we propose a hybrid approach that combines semantic, syntactic, and question category to judge question similarity. Semantic and syntactic information is measured by taking into account word similarity, word order, and part of speech information. Information about the types of question is derived from a Support Vector Machine classifier. The experimental results demonstrate that our combined measures are highly effective in distinguishing original questions and their paraphrases, thus improving the potency of question matching task.

**Keywords:** Question similarity, sentence similarity, question categories, answer reuse, question answering, factoid questions, semantic and syntactic techniques.

## 1 Introduction

Digital reference services have gradually becoming a major part of the digital library services due to the popularity of well-known online services, such as Internet Public Library (IPL) and Ask Dr. Math. In a typical digital reference service, librarians are responsible for answering the users' requests via online communication tools, e.g. email, web form, etc. We believe the process of answering questions in digital reference services can be significantly expedited by automated question answering module. If the system can determine whether a submitted question has been asked before, it can match the users' request with the existing question & answer pairs in the archive. This approach provides a tremendous value to the service as it offers a real-time response to

the redundant questions and increase the librarians' availability to assist the users with a truly unique need. Since a question is normally represented in a short sentence text, the challenge is how to effectively identify the similarity of questions, thus improving the potency of question matching task. Due to the variability of natural language expression, the same information can be formulated in numerous ways. Therefore, most document similarity approaches are likely to assign a low similarity score to those questions.

In this paper, we propose a method to measure the similarity between questions by utilizing semantic, syntactic, and question category information. Semantic information was derived from a lexical resource while syntactic information was derived from word order and part of speech information. Categorical information of questions was provided by a trained question classifier. Additionally, we are interested in the investigating an agreement between different evaluation metrics. Specifically, we investigate whether a similarity measure that correlates higher with human judgment also leads to a better performance on precision/recall based metric. Next, we examine how different word similarity measures and their similarity threshold affect the overall performance of sentence-level similarity measure.

The paper is organized as follows. First, we describe our approach to determine question similarity in section 2. In section 3, we describe the experimental set up and discuss about the results in section 4. Then, we review related work in section 5. Finally, we conclude the paper in section 6.

## 2 The Hybrid Approach

Our proposed method is a hybrid one based on the combinations of three different components: *semantic similarity*, *syntactic similarity* and *question category similarity*. The combination of the first two components (*semantic* + *syntactic*) represent the *sentence similarity* component [11] while the addition of the third component, *question category*, transform the similarity measure into question similarity measure. To quantify the similarity between words in the sentence, semantic information was obtained from WordNet [5]. A part of speech tagger was used to acquire the structural information of the question phrases, i.e. word order and part of speech labels. While a deep NLP technique, e.g. complete parse trees, might provide greater syntactic information of the sentences, our reason to use shallow NLP technique, i.e. part of speech tagging, was to balance the trade offs between the effectiveness and efficiency of the similarity measure. Equation 1 below describes the question similarity function ( $S$ ) between two questions  $q_1$  and  $q_2$  as follow:

$$S(q_1, q_2) = \alpha \cdot (\gamma \cdot S_s(q_1, q_2) + \delta \cdot S_t(q_1, q_2)) + \beta \cdot S_c(q_1, q_2) \quad (1)$$

Four component coefficients were used to fine tune three similarity components. First, we optimized two sub-components within the sentence similarity component: *semantic similarity* ( $S_s$ ) and *syntactic similarity* ( $S_t$ ), through  $\gamma$  and  $\delta$ , respectively. Then, we controlled the influence of sentence similarity and question category similarity ( $S_c$ ) components via  $\alpha$  and  $\beta$ , respectively. All component coefficients have a real-number value ranging from 0 to 1.

To produce the actual question similarity score, each component will be replaced by the appropriate sentence similarity measures, which are described in section 2.2, and question category similarity measure, described in section 2.3. For example, either *sentence vector similarity* or *part-of-speech semantic similarity* measures can be plugged into the semantic similarity component. This results in a number of similarity measure combinations, in which we described them in section 2.4. Finally, most sentence similarity measures in section 2.4 rely on the comparison of individual words between two sentences. Such comparison requires word similarity measures which is described in the next section.

## 2.1 Word Similarity Measures

We adapted two existing measures to compute word similarity scores: Lin's universal similarity [12] and gloss overlap measures [1]. The two measures were chosen because of their superior performance to the conventional path-based similarity measures and their distinct approach to compute word similarity. Mainly, Lin's measure combines local similarity judgment with global term information from information content value while gloss overlap measure only computes word similarity on a local basis. The similarity value produced by both measures has a real-number value ranging from 0 (not similar) to 1 (identical).

### 2.1.1 Universal Similarity Measure

In this measure, the similarity between two words,  $w_1$  and  $w_2$  is determined by their information content and the path distance in WordNet hierarchies. Here, we used Resnik's formulation [19] of information content which defines the information content of concept  $c$  as the negative log likelihood function  $-\log(p(c))$ , where  $p(c)$  is the probability of encountering such concept  $c$ .

### 2.1.2 Gloss Overlap Measure

The Gloss overlap approach for measuring word similarity was first introduced by [9]. Our variation of gloss overlap similarity between two words is defined as the overlap between their glosses (dictionary definition) and their direct hypernym and hyponym in WordNet hierarchies [1]. The overall similarity measure is formulated as follow:

We empirically tested the correlation with human judgment for both measures on the selected noun pairs from the standard Rubenstein and Goodenough (R&G) data set used in [11] and found that both correlated highly with human judgment.  $sim_{ic}$  performed slightly better than  $sim_{gloss}$  ( $r_{lin}=0.924$  and  $r_{gloss}=0.901$ ).

## 2.2 Sentence Similarity Measures

We adopted the similarity measures used in [11] and [14] due to their efficiency in representing sentence-level text. All similarity measures used in this work rely on a pair-wise comparison between words in the two sentences. To select the best score for each word pairs, we performed a simple word sense disambiguation by choosing the maximum similarity score. The similarity score generated by all three measures has a real-number value ranging from 0 (not similar) to 1 (identical).

**2.2.1 Sentence Vector Similarity**

The motivation behind semantic measure was to distinguish sentences beyond their surface-text form by utilizing semantic similarity between words. From the vectorial model perspective, this means regular term weights, e.g. word frequency or TF-IDF, are replaced by semantic similarity scores. The process to compute the sentence vector similarity is described as follows. First, each sentence is converted into a sentence vector. Then the similarity between two sentences is derived from the cosine coefficient between the two sentence vectors. Each entry in the sentence vector is derived from computing word similarity score between word feature  $w_i$  and each word in the sentence. After that, the maximum score from the matching word that exceeds certain similarity threshold will be chosen. The idea behind sentence vector representation is very simple yet effective solution for a pair-wise comparison of sentences. The following example demonstrates the process to construct a sentence vector. Suppose sentence  $s_1$  and  $s_2$  are the two sentences to be compared;  $s_1 = \{w_1, w_2, w_3\}$  and  $s_2 = \{w_1, w_3, w_4\}$ , the sentence vector  $sv_1$  and  $sv_2$  are shown below:

	$w_1$	$w_2$	$w_3$	$w_4$
$sv_1$	1	1	1	$sim_m(w_4, s_1)$
$sv_2$	1	$sim_m(w_2, s_2)$	1	$sim_m(w_4, s_2)$

where  $sim_m(w_i, s_j)$  is a maximum word similarity score of  $w_i$  and the matching word in  $s_j$ . If the two words are lexically identical, then  $sim_m(w_i, s_j)$  is equal to 1.

**2.2.2 Word Order Similarity**

The ability to deal with different word compositions or morpho-syntactic variations in sentences is crucial for determining sentence similarity. Basic information, such as word order, can provide useful information to distinguish the meaning of two sentences. This is particular important in our case where single word token was used as a basic lexical unit. Without syntactic information, it is impossible to discriminate the sentence containing words “sale manager” and “office worker” from another sentence containing “office worker” and “sale manager” since both of them essentially share the same bag-of-word representation. Word order similarity is defined as the normalized difference of word order between the two sentences. It has been proved in [11] to be an efficient method to compute the similarity of word order. The formulation for word order similarity is defined as follow:

$$sim_{wo}(s_1, s_2) = 1 - \frac{\|r_1 - r_2\|}{\|r_1 + r_2\|} \tag{2}$$

$r_1$  and  $r_2$  is a word order vector of sentence  $s_1$  and  $s_2$ , respectively. The steps to build a word order vector are similar to sentence vector’s process. That is, a feature set of word order vector is taken from the individual words of the two sentences. Each entry in the word order vector is derived by comparing word feature  $w_i$  with each word in the sentence. If the two are identical, then we fill the entry of  $w_i$  with an index number (word position) of the corresponding word. Otherwise, we calculate word similarity score between  $w_i$  and the remaining words in the sentence and fill  $w_i$  entry with an index number of a matching word that gives a maximum similarity score.

### 2.2.3 Part-of-Speech Semantic Similarity

Unlike sentence vector similarity where word similarity is exhaustively computed over the possible word pairs, another way to measure semantic information between sentences is to compute word similarity between words of the same part of speech [5]. That is, a simple syntactic analysis (part of speech) of words is included in this measure. As such, this approach intuitively fits to handle a common form of paraphrase – lexical substitution. The Part-of-speech semantic similarity between two sentences is defined as a combined maximum similarity score from all word pairs in each part of speech class. In this formulation, we consider noun, verb, adjective, and adverb as the major part-of-speech classes. The overall sentence similarity is defined as follow:

$$sim_{ps}(s_1, s_2) = \frac{\sum_{w \in \{s_1\}} \max Sim(w, s_2) + \sum_{w \in \{s_2\}} \max Sim(w, s_1)}{|s_1| + |s_2|} \quad (3)$$

where  $\max Sim(w, s_2)$  is derived from selecting the maximum similarity score of  $w$  and the matching word in  $s_2$ , while  $\max Sim_m(w, s_1)$  is derived from selecting the maximum similarity score of  $w$  and the matching word in  $s_1$ . Then  $sim_{ps}(s_1, s_2)$  is computed by combining the sum of  $sim_m(w_i, s_2)$  and  $sim_m(w_j, s_1)$  from all four part of speech classes normalized by the length of  $s_1$  and  $s_2$ .

### 2.3 Question Category Similarity Measure

Questions differ from regular sentences in that they contain interrogative words. These words can be used to determine the *aboutness* of the questions. Given two questions with almost exact same words, the interrogative part acts as a surrogate of the categorical information that helps distinguish them. For instance, we can say that “*where was JFK assassinated?*” and “*when was JFK assassinated?*” are two different questions judging by different *wh*-pronouns: *where* (location) and *when* (time). Thus, we built our question category similarity measure around the idea that similar questions share the same interrogative words or question categories.

We define question category similarity as a cosine similarity between the question category vectors. Thus, the major step in our approach is the construction of question category vector. For the task of classifying questions into different types, we chose Support Vector Machine (SVM) as the underlying classifier as it has been shown in many literatures to be the best performer for question classification task [23]. In this work, we used SVMLight [8] as the implementation of the classifier.

We developed SVM classifier using linear kernel to predict the question categories. The features for the classifier include unigram, multiword collocations, and the hypernyms of the head nouns (the head of the noun phrases). Specifically, we restricted the head nouns to those following the interrogative words. For instance, a head noun of the question “*What tourist attractions are there in Reims?*” is *tourist*. A list of multiword collocations, including interrogative words, was compiled from the training example. For example, “how many”, “how much”, “what is a”, “what is the” were automatically identified and extracted by the aforementioned tool. In the testing stage, we simply used exact string match to identify multiword collocations. The hypernyms of the head noun serve as semantic features which increase the chance of semantically-similar concepts sharing common features. The method of extracting head nouns and their hypernyms are the same as the one in [15]. The classifier was built on the



UIUC dataset which is a superset of the TREC QA track [10] dataset. The UIUC dataset contains 5,500 training questions and 500 TREC-10 questions for testing. Their question class taxonomy contains two levels. The coarse level has six categories whereas the fine level has fifty categories. The classification precisions for coarse-grained and fine-grained taxonomies are 81.8% and 89.2%, respectively. In this study, we classified questions based on the fine-grained categories due to their superior performance. Moreover, we took a multi-label classification approach to categorize questions. As such, a question was classified into multiple categories.

We experimented with two approaches to build the question category vector. First, we constructed the vector based on *the ranked category*. In this approach, we use the ranks of predicted category as a feature set, starting from rank 1, 2, 3, etc. The other approach utilizes *the classification probability*. Here, we used the question categories as a feature set in the vector, starting from category 1, 2, 3, etc. Each entry in the question category vector contains the probability that a question will be classified into a given category.

## 2.4 The Combinations of Measures

By replacing the similarity components with appropriate similarity measures described in section 2.2 and 2.3, we derived a set of similarity measure combinations. First, three possible combinations of sentence similarity measures are shown in table 1. To simplify the task of tuning the semantic and syntactic components, we employed *ps* in a syntactic similarity component in one combination and a semantic similarity component in another since it considers both semantic and syntactic information in comparing sentence similarity. As a result, each combination represents a different flavor of similarity notions. For example, *sv+wo* judges similar questions on their word semantics and word order, *sv+ps* uses word semantics and part of speech information, and finally, and *ps+wo* employs word semantics from specific part of speech classes and word order. Next, we combined the each sentence similarity combination in table 1 with two variations of question category similarity measure described in the previous section. Each variation is based on the different approaches to construct question category vector. For example, *sv+wo+rank* represents sentence vector similarity + word order similarity + ranked category vector similarity. Ultimately, a total number of six combinations of question similarity measures were derived.

**Table 1.** The combinations of sentence similarity measures that represent sentence similarity component

Combination of Measures	Semantic Component	Syntactic Component
<i>sv+wo</i>	Sentence vector	Word order
<i>sv+ps</i>	Sentence vector	Part of speech
<i>ps+wo</i>	Part of speech	Word order

## 3 Experimental Evaluation

The experiment was structured into two parts. First, we investigated the effectiveness of the sentence similarity measures in terms of correlation with human judgment. In

the second part, we compared the performance of sentence similarity-only measures and the combined sentence and question category similarity measures on a set of paraphrased questions.

**Baseline:** Three measures were selected as the baseline comparisons: *Jaccard* coefficient, which is traditionally used as a distance measure when comparing the two strings, a standard *TF-IDF* term vector similarity, and a text semantic similarity measure based on the combined semantic similarity of words in the same part of speech and their IDF scores (henceforth *ps-IDF*) [5].

### 3.1 Sentence Similarity Experiment

The goal of sentence similarity experiment was to compare the performance of sentence similarity measures on the correlation with human judgment. To achieve that, we computed sentence similarity scores between sentence pairs using the three combinations of sentence similarity measures described in the previous section. Moreover, we investigated the three factors and their impact on the performance of each sentence similarity measure. These are *the underlying word similarity measure* (*lin* vs. *gloss*), *its threshold level* (from 0 to 1), and *the relative contribution of the semantic and syntactic components* ( $\gamma$  and  $\delta$ , respectively) indicated by the combinations of their coefficient values from 0 to 1.

### 3.2 Question Similarity Experiment

In this part of the experiment, we compared the performance between the sentence similarity component (semantic + syntactic) and the combined sentence similarity and question category similarity components (semantic + syntactic + question category). Additionally, we explored the effect of two variations of question category similarity measure (*rank* and *conf*) described previously on performance of overall question similarity measure. Furthermore, we investigated how the performance of question similarity measure is affected by the relative contribution of sentence similarity and question similarity components ( $\alpha$  and  $\beta$ , respectively). For cross-validation purpose, we also compared  $\gamma$  and  $\delta$  obtained from sentence pairs data set used in 3.1 with the optimal values derived from question pair data set in this experiment.

### 3.3 Data Sets

We conducted a sentence similarity evaluation on thirty sentence-pair data set published in [11]. Each sentence pair was derived from a definitive sentence of a subset of noun pairs from Rubenstein and Goodenough (R&G) data set. To evaluate the performance of the question similarity measures, we selected a set of 193 question pairs from TREC-9 question variants key. The variants key consists of fifty four original questions and their variants. The original questions are a subset of test questions used in TREC-9 QA experiment and were taken from the actual users' submissions. The question variants are the paraphrased questions that were constructed by human assessors to be semantically identical but syntactically different from the original questions. The total number of question pairs used in the experiment is 386 -- 193 pairs for testing paraphrased questions and another 193 pairs for testing

non-paraphrased questions. Although the data set is semi-artificial, it contains sufficient linguistic complexity to reflect the variability of nature language expressions. That is, there are various types of paraphrasing strategies [22] exhibited in the question variants, e.g., lexical substitution (*what kind of animal was Winnie the Pooh?* vs. *what species was Winnie the Pooh?*), morpho-syntactic variations (*what kind of animal was Winnie the Pooh?* vs. *Winnie the Pooh is what kind of animal?*, *who owns CNN?* vs. *CNN is owned by whom?*), interrogative reformulation (*how did Bob Marley die?* vs. *what killed Bob Marley?*), semantic inference (*What tourist attractions are there in Reims?* vs. *What do most tourists visit in Reims?*), with more than 50% of the paraphrases categorized into multiple categories.

**Table 2.** The composition of paraphrase categories in TREC-9 question variants

Paraphrase Category	Lexical Substitution	Morpho-Syntactic Variation	Interrogative Reformulation	Semantic Inference
# of questions	63	97	112	31

### 3.4 Preprocessing

The preprocessing procedure is described as follows. First, individual words in sentence/question text were extracted, part-of-speech tagged, but not stemmed to preserve their meaning. A set of functional words -- words that do not contain semantic content such as articles, pronouns, prepositions, conjunctions, auxiliary verbs, modal verbs, and punctuations, was removed. Cardinal numbers were not discarded. Then, word similarity scores for all possible word pairs were computed and the results were cached for later use.

### 3.5 Evaluation Criteria

Pearson's correlation coefficient was used to measure the correlation between human-judgment scores and algorithmic scores in the sentence similarity experiment. The correlation coefficients were tested at the significant level of  $p < 0.01$ . To evaluate the performance of our question similarity measures, we adapted the notion of rejection/recall used in [12] as it is a better representation of the task's performance. *Recall* is defined as the proportion of question pairs correctly judged to be similar compared to the total number of similar question pairs. *Rejection* is defined as the proportion of question pairs correctly judged to be dissimilar compared to the total number of dissimilar question pairs. Finally, to evaluate the combined performance of recall-rejection, we defined the harmonic mean of unigram recall and rejection ( $F_1$ ) similar to the one used in standard information retrieval evaluation.

## 4 Results and Discussion

### 4.1 Sentence Similarity

According to table 3, sentence similarity measures significantly outperformed the baseline measure ( $r = 0.85-0.88$ ) on the measures using *lin* as the word similarity

measure while sentence similarity measures that employed *gloss* as the word similarity performed poorer ( $r = 0.72-0.81$ ). Since R&G experiment is based on synonymy evaluation, *lin*'s notion of similarity fits the human judgment better. The baseline Jaccard coefficient and TF-IDF measures correlated reasonably well with the sentence data set ( $r_{Jaccard} = 0.81$  and  $r_{TF-IDF} = 0.87$ ) while text semantic similarity measure correlated the lowest ( $r_{ps-IDF} = 0.75$ ). This comes as no surprise since most similar sentences in the sentence pair data set contain the reasonable numbers of word overlaps, while the dissimilar sentences contain fewer common words, the naïve methods that operate at a surface text level were expected to generate a good result.

**Table 3.** The Pearson's correlation coefficient of each similarity measure with subject to human judgment on R&G-based sentence pair data set

Word Similarity Measure	<i>Lin</i>			<i>Gloss</i>		
Similarity Measure	<i>sv+wo</i>	<i>sv+ps</i>	<i>ps+ws</i>	<i>sv+wo</i>	<i>sv+ps</i>	<i>ps+wo</i>
Correlation Coefficient	0.85	0.87	0.88	0.72	0.79	0.81

Next, the effect of semantic/syntactic contribution differs on each similarity measure combination. *sv+wo* and *ps+wo* correlated the highest when the semantic component was weighted higher than the syntactic component ( $\gamma=0.8$  and  $\delta = 0.2$ ). In the case of *sv+ps*, the optimal result was met when the syntactic component was weighted higher ( $\gamma = 0.3$  and  $\delta = 0.7$ ). The result in *sv+wo* and *ps+wo* combinations are similar to the one reported in [11]. That is, in general, the semantic component plays a greater role than the syntactic component in determining the similarity between sentences. Furthermore, *sv+wo* correlated the lowest with human judgment compared to the other two combinations. Both *lin* and *gloss* correlated relatively well with human judgment at high word similarity threshold levels (greater than 0.5). Again, *lin* consistently outperformed *gloss* in all combinations, having *sv+ps* and *ps+wo* as the best overall measures.

## 4.2 Question Similarity

First, we tried to find the word similarity measure and threshold level that produces the best performance for the combined sentence similarity measures (*sv+wo*, *sv+ps*, and *ps+wo*). Any question pairs with similarity score exceed a threshold of 0.7 were considered to be a paraphrased pair. The result indicated that all sentence similarity combinations significantly outperformed all three baselines. Next, similarity measures using *lin* as the word similarity measure did not significantly outperform those using *gloss* in both recall and rejection at  $p<0.05$ . Within the same word similarity measure, lower word semantic similarity threshold (0) performed better than higher word similarity threshold (0.6),  $p<0.05$ . The result offers an interesting contrast to that of the sentence similarity experiment. While the higher word similarity thresholds correlated higher with human judgment than the lower word similarity thresholds, it was the latter that performed significantly better on recall and rejection metrics.

**Table 4.** The performance of the best overall measures and baselines on identifying TREC-9 question variant

Combination of Measures	<i>sv+ps</i>	<i>sv+ps+rank</i>	<i>ps+wo+conf</i>	<i>Jaccard</i>	<i>TF-IDF</i>	<i>ps-IDF</i>
Recall	0.79	0.88	0.98	0.24	0.50	0.30
Rejection	1.00	1.00	0.93	1.00	1.00	0.99
$F_1$	0.88	0.94	0.95	0.39	0.67	0.46

Next, we compared the effectiveness of the sentence similarity and question similarity measures at the optimal word similarity setting obtained from the above experiment. That is, we used *lin* as word similarity measure computed at the word similarity threshold of 0. The result is shown in table 4. Among the three sentence similarity combinations, *sv+ps* performed the best ( $F_1^{sv+ps} = 0.88$ ). The optimal semantic/syntactic coefficient values were similar to those in 4.1. This result reaffirmed that the optimal coefficient settings were applicable across data sets. Then, we reapplied the coefficient settings to the corresponding components in the question similarity measure. The inclusion of question category similarity measures (*rank* and *conf*) has significantly improved the overall performance to identify paraphrased questions. Among six question similarity combinations, measures that employ *conf* as the question category vector have significantly produced greater recalls than *rank* measures, however, with greater expense on rejection. *sv+ps+rank* is the best overall measure among *rank* combinations ( $F_1^{sv+ps+rank} = 0.94$ ) while *ps+wo+conf* is the best overall among *conf* combinations ( $F_1^{ps+wo+conf} = 0.95$ ) at the similarity threshold of 0.7.

Overall, the analysis of the experimental result on various parameter settings has shown that the best similarity measure consistently performed well across different evaluation metrics. Different types of word similarity measures did not produce a significantly different result in sentence and question similarity evaluation. Due to the fact that both measures utilize the same word coverage in WordNet, they ultimately produced a similar result regardless of their approaches. Specifically, WordNet contains 85% of the vocabulary space of the test data set, making it reasonably effective. Different word similarity thresholds yielded significantly different results on Pearson’s correlation and  $F_1$  metrics. Measures with higher word similarity threshold performed better in correlation metric while measures with lower word similarity threshold performed better in  $F_1$  metric. This shows that rejection/recall tended to over-penalize the similarity scores at higher word similarity thresholds. Different types of question category vector generated significantly different results. Overall, *conf* combinations produced the highest recall but suffered from a minimal rejection rate. The significant loss in rejection eventually outweighed the gain in recall. Finally, the optimal results were achieved by approximately equal contribution of the sentence similarity and question category similarity components.

## 5 Related Work

Several approaches to measure sentence-level similarity have been proposed recently [2][6][10][12][15][17][18]. Vector space model and lexical resources have been

applied to measure question similarity [4]. Although vector space model approaches work very well in document retrieval task, they are not suitable for short text matching because of small word overlaps, data sparseness, and lexical chasm problem [2]. The work by [13] is perhaps the most relevant to ours since they used semantic metrics and question type metric to judge the question similarity. Our approach is different from theirs in many aspects. First, we cover a broader range of similarity metrics (semantic, syntactic, and question category). Second, there are a number of differences between their question type similarity and ours. They treated question classification as a binary classification task while we consider the task as a multi-label classification. We believe this approach follows a more intuitive notion. Next, we automatically extracted multiword collocations and the hypernyms of the head nouns instead of manually constructing the feature set. Lastly, we used fine-grained question categories due to their superior accuracy in question classification task.

## 6 Conclusions

We have demonstrated that semantic, syntactic, and question category information is very effective in identifying paraphrased questions. Semantic and syntactic measures were helpful in handling synonyms, related words, and different word compositions. The addition of question category information has significantly improved the performance of the similarity measure by providing discriminative power from the interrogative words in the question sentences. We recognized certain shortcomings in the use of TREC-9 data set since it is partially artificial. Hence, it might be less noisy and contain fewer cases of lexical-syntactic variations. Moreover, most questions in TREC-9 data set are factoid questions which only cover a subset of those being queried a real-world reference service. The future works include improving the method to incorporate more contextual information into the similarity measure. Currently, we represented sentence and question phrases at the individual words level. We believe the performance can be improved by considering a more meaningful lexical unit such as multiword phrases. In addition, we plan to extend the word similarity measures to handle words that do not exist in WordNet taxonomy via other knowledge resources, e.g. web search, Wikipedia, etc. Furthermore, we plan to test our approach on other question-answering dataset.

**Acknowledgments.** This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196) and PA Dept of Health Grant (No. 239667).

## References

1. Achananuparp, P., Han, H., Nasraoui, O., Johnson, R.: Semantically enhanced user modeling. In: Proceedings of SAC 2007, pp. 1335–1339. ACM Press, New York (2007)
2. Achananuparp, P., Hu, X., Zhou, X., Zhang, X.: Utilizing Sentence Similarity and Question Type Similarity to Response to Similar Questions in Knowledge-Sharing Community. In: Proceedings of QAWeb 2008 Workshop, Beijing, China (2008)

3. Berger, A., Caruana, D., Cohn, D., Freitag, D., Mittal, V.: Bridging the lexical chasm: Statistical approaches to answer-finding. In: Proceedings of SIGIR, pp. 222–229 (2000)
4. Burke, R.D., Hammond, K.J., Kulyukin, V.A., Lytinen, S.L., Tomuro, N., Schoenberg, S.: Question answering from frequently asked question files: Experiences with the FAQ finder system. Technical report (1997)
5. Corley, C., Mihalcea, R.: Measuring the semantic similarity of texts. In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, Ann Arbor, Michigan, pp. 13–18 (June 2005)
6. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
7. Jeon, J., Croft, W.B., Lee, J.H.: Finding similar questions in large question and answer archives. In: Proceedings of ACM CIKM, pp. 84–90 (2005)
8. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of European Conference on Machine Learning, pp. 137–142 (1998)
9. Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In: Proceedings of the 5th annual international conference on Systems documentation, pp. 24–26 (1986)
10. Li, X., Roth, D.: Learning Question Classifiers. In: COLING 2002 (August 2002)
11. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering* 18(8), 1138–1150 (2006)
12. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proceedings of the Fifteenth international Conference on Machine Learning, San Francisco, CA, pp. 296–304 (1998)
13. Lytinen, S., Tomuro, N.: The Use of Question Types to Match Questions in FAQFinder. In: 2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases, pp. 46–53. AAAI Press, Menlo Park (2002)
14. Malik, R., Subramaniam, V., Kaushik, S.: Automatically Selecting Answer Templates to Respond to Customer Emails. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 1659–1664 (2007)
15. Metzler, D., Bernstein, Y., Croft, W., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: Proceedings of CIKM, pp. 517–524 (2005)
16. Metzler, D., Croft, W.B.: Analysis of Statistical Question Classification for Fact-based Questions. *Information Retrieval* 8(3), 481–504 (2005)
17. Metzler, D., Dumais, S.T., Meek, C.: Similarity Measures for Short Segments of Text. In: Amati, G., Carpineto, C., Romano, G. (eds.) *ECIR 2007*. LNCS, vol. 4425, pp. 16–27. Springer, Heidelberg (2007)
18. Murdock, V.: Aspects of sentence retrieval. Ph.D. Thesis, University of Massachusetts (2006)
19. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In: International Joint Conference for Artificial Intelligence (IJCAI 1995), pp. 448–453 (1995)
20. Sahami, M., Heilman, T.D.: A web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international Conference on World Wide Web, Edinburgh, Scotland, pp. 377–386 (2006)
21. Smadja, F.: Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1), 143–177 (1993)
22. Tomuro, N.: Interrogative Reformulation Patterns and Acquisition of Question Paraphrases. In: Proceedings of the Second international Workshop on Paraphrasing, pp. 33–40 (2003)
23. Zhang, D., Lee, W.S.: Question classification using support vector machines. In: Proceedings of SIGIR 2003, pp. 26–32. ACM Press, New York (2003)

# A Collaborative Approach to User Modeling for Personalized Content Recommendations

Heung-Nam Kim, Inay Ha, Seung-Hoon Lee, and Geun-Sik Jo

Intelligent E-Commerce Systems Laboratory,  
Department of Computer & Information Engineering, Inha University  
{nami, inay, shlee}@eslab.inha.ac.kr, gsjo@inha.ac.kr

**Abstract.** Recommender systems, which have emerged in response to the problem of information overload, provide users with recommendations of contents that are likely to fit their needs. One notable challenge in a recommender system is the cold start problem. To address this issue, we propose a collaborative approach to user modeling for generating personalized recommendations for users. Our approach first discovers useful and meaningful patterns of users, and then enriches a personal model with collaboration from other similar users. In order to evaluate the performance of our approach, we compare experimental results with those of a probabilistic learning model, a user-based collaborative filtering, and vector space model. We present experimental results that show how our model performs better than existing work.

## 1 Introduction

The prevalence of digital libraries and the development of Web 2.0 services enable end-users to be producers as well as consumers of contents. Even in a single day, an enormous amount of textual contents, such as news, research papers, blog articles, and wikis etc., is generated on the Web. It is getting more difficult to make a recommendation to a user about what she prefers among these contents automatically because of, not only their huge amount, but also the difficulty of automatically grasping her interests [12]. Recommender systems, which have emerged in response to the above challenges, provide users with recommendations of contents that are likely to fit their needs. There are two widely approaches among recommender systems, i.e., Content-based filtering (CB) [6] and Collaborative Filtering (CF) [2]. CF has an advantage over CB that is the ability to filter any type of items where it is hard to analyze the underlying content, e.g., music, videos, and photos. Because the filtering process is only based on historical information about whether or not a given target user has preferred an item before, analysis of the actual content, itself, is not necessarily required.

Nevertheless, CF suffers from a fundamental problem, namely the *cold start problem*, which can be divided into *cold start items* and *cold start users* [5]. Several researches have been proposed and challenged to address this problem [5, 6, 11]. In a CF-based recommender system, an item cannot be recommended until a large number of users have rated it before, known as the cold start item. This problem applies to new items generated every few minutes and can be partially improved by CB technology. In



the case of domains such as textual documents, CB has proven to be effective in locating textual contents relevant to a specific content information need [3, 13]. However, CB also encounters limitation for the cold start user, similar to CF. A cold start user describes a new user that joins a recommender system and has presented few opinions (i.e., the user has insufficient preference history). With these situations, the system is generally unable to make high quality recommendations.

Our aim was to build a robust user model that can be used for providing personalized recommendation both in terms of improving the performance and in dealing with the cold start problem. By capturing users' content of interest, we discover the preference patterns and terms existing in the user's content of interest. In addition, to partially overcome the cold start user problem, we propose an enrichment method of the personal model in collaboration with other similar users. The subsequent Sections are organized as follows: The next Section contains a brief review of our previous study and some notations. In Section 3, we describe a collaborative approach for modeling user interests and recommending contents. A performance evaluation is presented in Section 4. Finally, conclusions are presented and future work is discussed in Section 5.

## 2 Preliminaries

In this Section, we first review our previous study for modeling user preference [13]. Prior to this paper, we successfully built a personal user model that drives from the user's content of interest. Before going into further detail, some notations and definitions are introduced for understanding our approach. Let  $C = \{c_1, c_2, \dots, c_n\}$  be a set of all contents,  $T = \{t_1, t_2, \dots, t_m\}$  be a set of all index terms, and  $U = \{u_1, u_2, \dots, u_l\}$  be a set of distinct users. A content  $c_j$  is a set of terms, each of which may appear in multiple contents with different weights that quantify the importance of the term for describing the contents. In our study, a weight  $w_{i,j}$  associated with a pair  $(t_i, c_j)$  (i.e., a term  $t_i$  of a content  $c_j$ ) is computed by a fairly common type of *tf-idf* weighting scheme [1]. A co-occurrence pair  $(u_n, c_j)$  where  $u_n \in U$  is a user and  $c_j \in C$  is a content implies that user  $u_n$  viewed, collected, or clicked content  $c_j$ . We assume that contents clicked, read, or collected by a user are her interest contents. The first step in user modeling is to mine frequent term patterns from interest contents of each user. Since every user has different click-history, contents used for mining process should be selected individually for each user. Frequent patterns are a set of terms that appear frequently together in a set of interest contents of interest of a user. For example, if a set of terms {recommendation, collaborative, personalization, filtering} appear frequently together in a set of interest contents of a certain user, the set of those terms is a frequent pattern for the user. In the data mining research literature, the frequent patterns are typically defined as patterns that occur at least as frequently as a pre-determined minimum support (*min\_sup*) [8]. In our study, we applies the mining process based on the following assumption: each transaction corresponds to a interest content of a user, items in transaction are terms extracted from the content, and a transaction database corresponds to a set of interest contents of a user. Therefore, if the pattern support of pattern  $p_k$  (Definition 1), that is composed of at least  $l$  different terms, above *min\_sup*, i.e.,  $PS_u(p_k) > min\_sup$ , then pattern  $p_k$  is referred to a frequent term pattern. We denote a set of frequent term patterns for user  $u$  as  $E_u$ . Once the frequent

patterns are mined, we remove the patterns, which contain unnecessary terms, from  $E_u$  and model the user preference based on those patterns, collectively called Personalized Term Pattern. A formal description of a model for user  $u$ ,  $M_u$ , follows:  $M_u = \langle PTP_u, PT_u \rangle$ , where  $PTP_u$  models the interest patterns (Definition 3) and  $PT_u$  models the interest terms (Definition 4).

**Definition 1 (Pattern Support, PS).** Let  $I_u$  be a set of interest contents of user  $u$  and pattern  $p_k = \{t_1, t_2, \dots, t_n\}$  be a set of terms such that  $p_k \subseteq T$  and  $n \geq 2$ . A content  $c_j$  is said to contain pattern  $P_k$  if and only if  $p_k \subseteq c_j$ . *Pattern support* for pattern  $p_k$  in  $I_u$ , written as  $PS_u(p_k)$ , is the ratio of contents in  $I_u$  that contain pattern  $p_k$ . That is,  $PS_u(p_k) = f_u(p_k) / |I_u|$ , where  $f_u(p_k)$  indicates the occurrence frequency of pattern  $p_k$  in  $I_u$ .

**Definition 2 (Pattern Weight, PW).** *Pattern weight* of  $p_k$  for user  $u$ , denoted as  $PW_u(p_k)$ , indicates the importance of each term in representing the pattern and is computed as follows:

$$PW_u(p_k) = \frac{1}{|p_k|} \cdot \sum_{i \in p_k} \mu_{i,u}$$

where  $\mu_{i,u}$  is the mean of weight for term  $t_i$  in  $I_u$  and is computed as follows:

$$\mu_{i,u} = \frac{1}{|I_u(i)|} \times \sum_{j \in I_u(i)} w_{i,j}$$

where  $I_u(i)$  is a set of interest contents for user  $u$  containing term  $t_i$  and  $w_{i,j}$  is the weight of term  $t_i$  in content  $c_j$ .

**Definition 3 (Personalized Term Pattern, PTP).** *Personalized term pattern* is defined as a frequent term pattern whose *pattern weight* is greater than a threshold value  $min\_pw$ , i.e.,  $p_k \in E_u$  and  $PW(p_k) > min\_pw$ . A set of *personalized term patterns* for user  $u$  is denoted as  $PTP_u$  such that  $PTP_u = \{ (p_k, PS_u(p_k)) \mid PW_u(p_k) > min\_pw \wedge p_k \in E_u \}$ .

**Definition 4 (Personalized Term, PT).** *Personalized Term* is a term that occurs within *personalized term patterns*. The set of *personalized terms* for user  $u$  is denoted as  $PT_u$ . In addition, the vector for  $PT_u$  is represented by  $\vec{pT}_u = (\mu_{1,u}, \mu_{2,u}, \dots, \mu_{t,u})$ , where  $t$  is the total number of personalized terms and  $\mu_{i,u}$  is the mean of weight for term  $t_i$ .

### 3 Collaborative User Modeling for Content Recommendation

In this Section we describe how to enrich a user model for a given target user. The model  $M_u$  described in Section 2 is referred to the initial user model for user  $u$ . This model can be applied immediately to generate content recommendations. However, diverse patterns for user  $u$  are not able to be discovered thorough the mining process in the case that the user has clicked few contents, known as the cold start user. With this situation, initial personalized term patterns may not be sufficient to represent user preference, and thus our approach is generally unable to make high quality recommendations. In addition, when we only use the initial model for recommendations, it is hard to recommend novel contents that the target user might enjoy outside their

usual set of interest contents. For the above reasons, we propose an enrichment method of the model for the target user via the personalized term patterns of like-minded users.

### 3.1 Content-Based Neighborhood Formation

The main goal of neighborhood formation is to identify a set of user neighbors,  $k$  *nearest neighbors*, which is defined as a group of users exhibiting interest terms close to those of the target user. In a typically CF-based recommender system, it encounters serious limitations for finding a set of users, namely *sparsity problem* [6, 11]. The sparsity problem occurs when available data is insufficient to identify similar users (neighbors) due to an immense amount of contents. In practice, even though users are very active, the result of reading contents is just a few of the total number of contents. Accordingly, it is often the case that there is no intersection at all between two users and hence the similarity is not computable at all. Even when the computation of similarity is possible, it may not be very reliable, because of insufficient information processed. To this end, in our study, we select the best neighbors by using personalized terms,  $PT$ , of each user. In order to find  $k$  nearest neighbors, cosine similarity, which quantifies the similarity of two vectors according to their angle, is employed to measure the similarity values between a target user and every other user. As noted in Definitions 4, the personalized terms of two users,  $u$  and  $v$ , are represented as  $t$ -dimensional vectors,  $\vec{PT}_u$  and  $\vec{PT}_v$  respectively. Therefore, the similarity between two users,  $u$  and  $v$  is measured by Equation (1)

$$sim(u,v) = \cos(\vec{PT}_u, \vec{PT}_v) = \frac{\sum_{k=1}^t \mu_{k,u} \times \mu_{k,v}}{\sqrt{\sum_{k=1}^t \mu_{k,u}^2} \times \sqrt{\sum_{k=1}^t \mu_{k,v}^2}} \quad (1)$$

The similarity score between two users is in the range of  $[0, 1]$  and the higher score a user has, the more similar she is to the target user. After computing all-to-all similarity between users, we define the set of nearest neighbors of each user  $u$  as an ordered list of  $k$  users  $\mathcal{N}(u) = \{v_1, v_2, \dots, v_k\}$  such that  $u \notin \mathcal{N}(u)$ , and  $sim(u, v_1)$  is maximum,  $sim(u, v_2)$  is the next maximum and so on [4].

### 3.2 Collaborative Enrichment of User Preference

We explain the general idea of the enrichment process in the following: Let  $\mathcal{N}(u) = \{v_1, v_2, \dots, v_k\}$  be a sorted neighbor list of target user  $u$ ,  $PTP_u$  be a set of personalized term patterns for user  $u$ , and  $PTP_v$ ,  $v \in \mathcal{N}(u)$ , be a set of personalized term patterns for neighbor user  $v$  of user  $u$ . Firstly, we choose neighbor user  $v$  according to descending the similarity between target user  $u$  and neighbors. For each pattern  $p_i$  in  $PTP_u$ , *specific patterns* of  $p_i$  in  $PTP_v$  are identified. Given two patterns  $p_i$  and  $p_j$ ,  $p_i$  is said a *general pattern* of  $p_j$  if and only if  $p_i$  is a subset of  $p_j$ , i.e.,  $p_i \subset p_j$ . In contrary,  $p_j$  is called a *specific pattern* of  $p_i$ . For example, Let  $p_1 = \{t_1, t_2, t_3\}$  be a personalized term pattern for user  $u$  such that  $p_1 \in PTP_u$ , and  $PTP_v = \{p_2, p_3, p_4\}$  be a set of PTP for user  $v$  such that  $p_2 = \{t_1, t_2, t_4\}$ ,  $p_3 = \{t_1, t_2, t_3, t_4\}$ , and  $p_4 = \{t_1, t_2, t_3, t_5\}$ . Since pattern  $p_3$  and  $p_4$  contain entire terms of pattern  $p_1$ , they are said the specific pattern. Several specific patterns that occur in PTP of the neighbor  $v$ ,  $PTP_v$ , may be found. To make the

enrichment efficient, we only consider specific patterns which have the higher pattern support than that of the general pattern. Assume that the pattern support for  $p_1, p_3$ , and  $p_4$  is 0.3, 0.35, and 0.28, respectively (i.e.,  $PS_u(p_1)=0.3$ ,  $PS_v(p_3)=0.35$ , and  $PS_v(p_4)=0.28$ ). In this case, only pattern  $p_3$  is used for enriching the model of user  $u$ , if it is not PTP for user  $u$ . This pattern such as  $p_3$  is called a collaborative PTP for target user  $u$ . Finally, a set of collaborative patterns is identified from  $k$  nearest neighbors, with respect to target user  $u$ . Note that the collaborative PTP for the target user is not allowed to be redundant. That is, if same patterns, which are already enriched by neighbor  $v$ , are also discovered from another neighbor  $h$  such that  $sim(u,v) \geq sim(u,h)$  and  $v \neq h$ , those patterns are pruned. The enriched model for user  $u$  is defined as a triple  $M^*_u = \langle PTP_u, PT_u, CPTP_u \rangle$  where  $PTP_u$  is a set of personalized term patterns for user  $u$ ,  $PT_u$  is a set of personalized term for user  $u$ , and  $CPTP_u$  is a set of collaborative personalized term patterns for user  $u$ , respectively.

**Definition 5 (Collaborative PTP, CPTP).** Let  $p_i$  be a personalized term pattern for target user  $u$ ,  $p_i \in PTP_u$ , and  $p_j$  be a personalized term pattern term for neighbor  $v$  such that  $p_j \in PTP_v$ , and  $v \in \mathcal{N}(u)$ . We define the set of collaborative personalized term patterns for user  $u$ , denoted as  $CPTP_u$ , as the set of neighbor patterns  $p_j$  such that  $p_i \subset p_j$ ,  $p_j \notin PTP_u$ , and  $PS_u(p_i) \leq PS_v(p_j)$ .

### 3.3 Generating Content Recommendation

After the model is enriched, we are ready to provide recommendations for new contents that a user has not yet clicked or read. Based on the enriched model for each user, we recommend *top-N* ranked contents to the user that she might be interested in reading. *Top-N recommendation* is one of the recommendation schemes offering the target user  $u$  the ordered set of items  $X_u$  such that  $|X_u| \leq N$  and  $X_u \cap I_u = \emptyset$  [7]. To this end, the most important task in personalized recommendation is to generate a prediction, this is, attempting to speculate upon how a certain user would prefer unseen contents. In our study, we consider matched patterns, that is, how many interest patterns in a user model are contained in the new content. Formally, the numeric score of the target user  $u$  for the content  $c_n$ , denoted as  $P_{u,n}$ , is obtained as the following:

$$P_{u,n} = \frac{\sum_{p_k \in (PTP_u \cup CPTP_u)} B_n^{p_k}}{N_u} \cdot \frac{\sum_{p_k \in (PTP_u \cup CPTP_u)} |p_k| \times \omega_u^{p_k} \times B_n^{p_k}}{\sum_{p_k \in (PTP_u \cup CPTP_u)} \omega_u^{p_k}} \tag{2}$$

where  $N_u$  is the total number of patterns in both  $PTP_u$  and  $CPTP_u$ , and  $B_n^{p_k}$  is binary variable for determining whether pattern  $p_k$  occur in content  $c_n$  or not. That is,  $B_n^{p_k}$  is 1 if pattern  $p_k$  appear content  $c_n$  and 0 otherwise.  $\omega_u^{p_k}$  represents the weighted pattern support of  $p_k$  for user  $u$ , which is given by:

$$\omega_u^{p_k} = \begin{cases} PS_u(p_k) & \text{if } p_k \in PTP_u \\ PS_v(p_k) \times sim(u,v) & \text{if } p_k \in CPTP_u, p_k \in PTP_v \end{cases} \tag{3}$$

Once the prediction of the target user for contents, which she has not yet read, are computed, the contents are sorted in order of descending predicted value  $P_{u,n}$ . Finally, a set of  $N$  ordered contents that have obtained the higher values are identified for user  $u$ .

Then, those contents are recommended to user  $u$ . The main concept of the prediction dictates that patterns with numerous occurrences in the model of the target user are a good estimate of preference for the selected content. This scheme provides some advantages, the ability not only to make a recommendation for new contents that no one has yet read but also to support serendipitous recommendations. That is, it can make contents containing patterns that are valuable to the target user, but are not discovered from her interest contents, to be at a higher rank in the recommended content set.

## 4 Experimental Evaluation

In this Section, we empirically evaluate the proposed approach and compare its performance against the performances of the benchmark algorithms. The experimental data is taken from NSF (National Science Foundation) research award abstracts [14]. The original dataset is too large to be used for experiments, and thus we selected award abstracts whose topic is highly related to computer science. The selected dataset contained 974 unique abstracts (i.e., contents) and 9,823 unique terms as obtained from the abstracts. In addition, we collected 3,894 click-histories (i.e., interest contents of the users) of the total contents clicked by 30 users. To evaluate the performance of the recommendations, we randomly divided the dataset into *a training set* and *a test set*. The clicked contents of the users were split into *a test set* with exactly 30 contents per user in the test set (i.e., 900 contents) and *a training set* with the remaining contents (i.e., 2,994 contents) was used to learn and build a model of each user.

In order to compare the performance of our approach, a user-based CF, which is described in [4] (denoted as *UCF*), a probabilistic learning algorithm, which applies the multinomial event model of a *naïve Bayes assumption* (denoted as *NB*) [9], and a TF-IDF vector-based algorithm, which is employed in [3] (denoted as *VT*), were implemented. For the content recommendation process, in the case of *NB*, contents were ranked using the calculated probability value whereas they were ranked using the calculated cosine similarity for *VT*. For *UCF*, the proximity between users was measured by cosine-based similarity and contents were ranked using the weighted sum using the similarity as the weight. The *top-N* recommendation of our strategy (denoted as  $M^*$ ) was then evaluated in comparison with the benchmark algorithms. We adopted two evaluation measures that are defined as follows:

**Hit Rate (HR).** In the context of *top-N* recommendations, *hit-rate*, a measure of how often a list of recommendations contains contents that the user is actually interested in, was used for the evaluation metric [13]. The *hit-rate* for user  $u$  is defined as:

$$HR(u) = \frac{|T_u \cap X_u|}{|T_u|}$$

where  $T_u$  is the content list of user  $u$  in the test data and  $X_u$  is a *top-N* recommended content list for user  $u$ . Finally, the overall HR of the *top-N* recommendation for all users is computed by averaging these personal  $HR(u)$  in test data.

**Reciprocal Hit Rank (RHR).** One limitation of the *hit-rate* measure is that it treats all hits equally regardless of the ranking of recommended contents. In other words, content that is recommended with a top ranking is treated equally with content that is

recommended with an  $N$ th ranking. To address this limitation, therefore, we adopted the *reciprocal hit-rank* metric described in [7]. The *reciprocal hit-rank* for user  $u$  is defined as:

$$RHR(u) = \sum_{c_n \in (T_u \cap X_u)} \frac{1}{rank(c_n)}$$

where  $rank(c_n)$  refers to a recommended ranking of content  $c_n$  within the *hit set* of user  $u$ . That is, hit contents that appear earlier in the *top-N* list are given more weight than hit contents that occur later in the list. Finally, the overall RHR for all users is computed by averaging the personal  $RHR(u)$  in test data. The higher the RHR, the more accurately the algorithm recommends contents.

## 4.1 Experimental Results

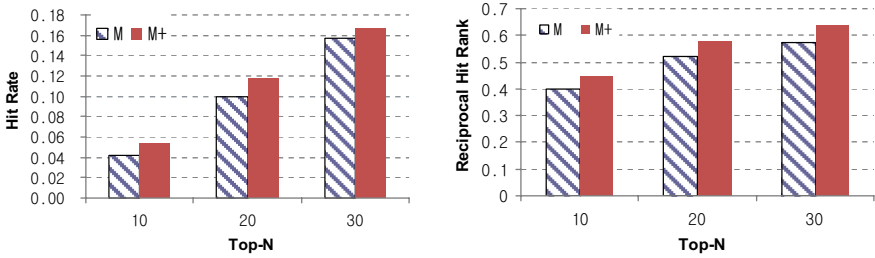
The performance evaluation is divided into two dimensions. The effectiveness of the model enrichment is first evaluated in comparison with the initial user model, and then the accuracy of content recommendations is evaluated in comparison with the benchmark methods.

### 4.1.1 Effect of the Model Enrichment

The following experiment investigates the effect of the enriched model through the neighborhood. According to the results of our previous study [13],  $min\_sup$  value was set to 0.1 (i.e., 10%) and  $min\_pw$  value was set to 0.79 which is the average weight of all terms in the dataset. As noted a number of previous studies, the size of neighborhood influences the recommendation quality of neighborhood-based algorithms. Therefore, different numbers of user neighbors were used for the model enrichment: 10, 20, and 30. We set the number of recommended items  $N$  to 30 for each user. As seen from the results of Table 1, the overall recommendation accuracy tend to be improved slightly. However, unlike CF-based approaches, increasing the neighborhood size did not practically affect the accuracy. Interestingly, the reciprocal hit rank was rather worse when the neighborhood size was 30 (RHR of 0.6376); whereas, its RHR was 0.6381 when the size was 10. These results were affected by the fact the neighborhood with a small size provide the collaborative PTP enough. Recall that the patterns are selected for enriching the collaborative PTP according to the nearest-order, and thus the redundant patterns generated by farthest neighbors are pruned. Another reasons might be that we were only looking for a small number of recommended contents (i.e,  $N=30$ ) and a number of users in the dataset was not sufficiently large. That is, once the number of nearest neighbors is relatively large, the rank of recommended contents for each user is not nearly changed by any further increases in the number of nearest neighbors. In consideration of computation cost, the neighborhood size for enriching the model was set to 10 in subsequent of experiments.

**Table 1.** HR and RHR as the neighborhood size grows ( $N=30$ )

Neighbors:	Hit Rate (HR)			Reciprocal Hit Rank (RHR)		
	10	20	30	10	20	30
M+	0.1666	0.1682	0.1682	0.6381	0.6389	0.6376



**Fig. 1.** Comparison of the accuracy achieved by the initial model and the enriched model

We continued to examine the effect of the enriched model  $M^+$  of each user, comparing the results achieved by the initial model  $M$  of each user. We performed an experiment with  $N$  values of 10, 20, and 30. Examining the average number of the collaborative PTP in the users, we found that 188 patterns had been enriched for each user. Figure 1 presents the results of the experiment. The results demonstrate that the enriched model shows the improved HR and RHR on all occasions, compared to the initial model. Particularly, with respect to the results of RHR, the enriched model provides considerably improved RHR, compared to the initial model. For example, when  $N$  is 30, the enriched model obtains a RHR of 0.6381 whereas the initial model demonstrates a RHR of 0.5712. This is particularly important since users practically tend to click on contents with higher ranks. We can conclude that using collaborative models brings significant advantages in terms of improving the recommendation accuracy.

### 4.1.2 Comparisons with Other Methods

For evaluating the  $top-N$  recommendation, we calculated the hit rate (HR) and the reciprocal hit rank (RHR) achieved by  $NB$ ,  $VT$ ,  $UCF$ , and  $M^+$ . We selectively varied the number of returned items  $N$  from 10 to 30 in an increment of 10. For  $UCF$ , a number of users in the dataset was not sufficiently large, and thus all users in the dataset were used as neighbors of the target user. Table 2 summarizes the results of RHR and HR. In general, with the growth of recommended items  $N$ , HR and RHR tend increase. In addition, HR for all algorithms is unsatisfactorily low at a small number of  $N$  (i.e.,  $N=10$ ). Nevertheless, comparing the results achieved by  $M^+$  and the benchmark algorithms, HR of the former found to be superior to that of the benchmark algorithms in all cases. Overall,  $M^+$  performs 6% better than  $NB$ , 2.5%, better than  $VT$ , and 2.7% better than  $UCF$ . Similar conclusions, except for  $N=10$ , can be made by looking at the RHR results as well. Interestingly, with respect to RHR,  $M^+$  and  $VT$  significantly outperformed  $NB$  and  $UCF$ . That is, when a relatively small number of contents were recommended,  $M^+$  and  $VT$  caused a more proper contents to be at a higher rank in the recommended contents. So they can provide better contents for a target user than the other methods. Overall,  $M^+$  achieves 27.1%, 3.8%, and 10.8% improvement in terms of RHR on average, compared to  $NB$ ,  $VT$ , and  $UCF$ , respectively. Ideally, recommendation algorithms should provide a wide range of desirable contents for users. Therefore, we continued to analyze the number of contents for

which the methods, except for *NB*, could not provide predictions for each user at all (i.e., prediction value of the target user for the content was zero). Recall that *NB* and *VT* is a class of CB whereas *UCF* is a class of CF. Strictly speaking, our approach is closely connected with CB due to the dependence of content characteristics (i.e., content-based user models, content-based neighbors, content-based enrichments, and content-based recommendations). As a result, contents of 59 for *UCF*, 22 for *VT*, and 22 for *M<sup>+</sup>* were not able to be predicted on average, respectively. As noted previously, such results were caused by the fact *UCF* could only make predictions for contents that at least a few users had clicked. On the other hand, *VT* and *M<sup>+</sup>* could only make predictions for contents that contained the terms in the target user model although they did not suffer from the cold start items at all.

**Table 2.** Comparisons of HR and RHR as the value of *N* increases

<i>Top-N:</i>	Hit Rate (HR)			Reciprocal Hit Rank (RHR)		
	10	20	30	10	20	30
NB	0.0303	0.0515	0.078	0.2436	0.2863	0.3190
VT	0.0515	0.0878	0.1242	0.4545	0.5245	0.5681
UCF	0.0484	0.0909	0.1181	0.3800	0.4627	0.4954
M+	0.0545	0.1181	0.1666	0.4445	0.5800	0.6381

## 5 Conclusions and Future Work

In this paper, particularly for textual contents, we presented a new and unique method for modeling user interests via collaborative approach of users and for providing enhanced recommendation accuracy. The major advantage of the proposed modeling method is that it supports not only the identification of useful patterns of each user but also the enrichment of valuable patterns of neighbors. As noted in our experimental results, our model obtained better recommendation accuracy, compared to the benchmark methods. Moreover, we also observed that our method can provide more suitable contents for user preference even though the number of recommended items is small. There are several interesting research issues to address in order to successfully apply our approach to a practical environment. The proposed method cannot be applied to domains, where contents are not easily analyzed by automated processes. In addition, there remain common issues that have been mentioned in keyword-based analysis: polysemy and synonymy. A semantic user model is one of the interesting issues that we plan to consider for addressing this problem in the future.

## References

1. Salton, G., Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. *Information Processing and Management* 24, 513–523 (1988)
2. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM* 40, 77–87 (1997)



3. Chen, L., Sycara, K.: WebMate: Personal Agent for Browsing and Searching. In: Proc. of the 2nd Int. Conf. on Autonomous Agents and Multi Agent Systems, pp. 132–139 (1998)
4. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.T.: Analysis of recommendation algorithms for e-commerce. In: Proceedings of ACM E-Commerce, pp. 158–167 (2000)
5. Schein, A.I., Popescul, A., Ungar, L.H., Pennock, D.M.: Methods and Metrics for Cold Start Recommendations. In: ACM Conference on Research and Development in Information Retrieval, pp. 253–260 (2002)
6. Melville, P., Mooney, R.J., Nagarajan, R.: Content-Boosted Collaborative Filtering for Improved Recommendations. In: Eighteenth national conference on Artificial intelligence, pp. 187–192 (2002)
7. Deshpande, M., Karypis, G.: Item-based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems* 22, 143–177 (2004)
8. Han, J., Pei, J., Yin, Y.: Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery* 8, 53–87 (2004)
9. McCallum, A., Nigam, K.: A Comparison of Event Models for Naïve Bayes Text Classification. In: AAAI 1998 Workshop on Learning for Text Categorization (1998)
10. Flesca, S., Greco, S., Tagarelli, A., Zumpano, E.: Mining User Preferences, Page Content and Usage to Personalize Website Navigation. *World Wide Web, Internet and Web Information System* 8, 317–345 (2005)
11. Degenmms, M., Lops, S.G.: A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction* 17, 217–255 (2007)
12. Das, A., Datar, M., Garg, A.: Google News Personalization: Scalable Online Collaborative Filtering. In: Proceedings of the 16th international conference on World Wide Web, pp. 271–280 (2007)
13. Kim, H.N., Ha, I.A., Jung, J.G., Jo, G.S.: User Preference Modeling from Positive Contents for Personalized Recommendation. In: Corruble, V., Takeda, M., Suzuki, E. (eds.) DS 2007. LNCS (LNAI), vol. 4755, pp. 116–126. Springer, Heidelberg (2007)
14. Pazzani, M.J., Meyers, A.: NSF Research Awards Abstracts 1990-2003, <http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html>

# Using a Grid for Digital Preservation

José Barateiro<sup>1</sup>, Gonçalo Antunes<sup>2</sup>, Manuel Cabral<sup>2</sup>,  
José Borbinha<sup>2</sup>, and Rodrigo Rodrigues<sup>3</sup>

<sup>1</sup> LNEC - Laboratório Nacional de Engenharia Civil, Lisbon, Portugal

<sup>2</sup> INESC-ID, Information Systems Group, Lisbon, Portugal

<sup>3</sup> Max Planck Institute for Software Systems, Kaiserslautern and Saarbrücken,  
Germany

jbarateiro@lnec.pt, {goncalo.antunes,manuel.cabral}@tagus.ist.utl.pt  
jlb@ist.utl.pt, rodrigo.rodrigues@inesc-id.pt

**Abstract.** Digital preservation aims at maintaining digital objects and data accessible over long periods of time. Data grids provide several functionalities required by digital preservation systems, especially when massive amounts of data must be preserved, as in e-Science domains. We propose the use of existing data grid solutions to build frameworks for digital preservation. In this paper we survey the main threats to digital preservation, which are used to identify a central point of failure in the metadata catalog of the iRODS data grid solution. We propose three extensions to the iRODS framework, to overcome the shortcomings of iRODS when used as a digital preservation system.

**Keywords:** Digital Libraries, Digital Preservation, Data Grids, e-Science.

## 1 Introduction

Physical artifacts like printed works or drawings carved in stone can survive for centuries. These stable objects are testimonials of past generations and an important asset to the future. In contrast, digital objects are unstable, requiring the execution of continuous actions to make it possible to interpret them in the future.

Digital preservation is defined as the ability of two or more systems or components to exchange and use information [6]. Digital preservation stresses the time dimension of this interoperability, focusing on the requirement that data or digital objects must remain authentic and accessible to users and systems over a long period of time, thus maintaining their value. Achieving this goal may require specific investments in an infrastructure for storing, maintaining, and managing data. Such costs may be prohibitive for small organizations, or organizations that do not have a steady revenue, like university libraries, research laboratories, or non-profit organizations.

Project GRITO [7] tries to lower the cost of digital preservation by harnessing the spare storage of grid clusters in Portuguese universities and research institutions. To achieve this goal we propose to build a heterogeneous storage

---

<sup>1</sup> <http://grito.intraneia.com>

framework that will integrate two classes of members: (*i*) exclusive storage clusters, comprising systems dedicated to digital preservation, which are likely to be under the administration of the data owner; (*ii*) extended clusters, as existing grid clusters, primarily used for data processing, but whose spare disk, CPU and bandwidth can be also used to support preservation services.

Project GRITO appears in the context of the international project SHAMAN<sup>2</sup> - Sustaining Heritage Access through Multivalent ArchiviNg, whose goal is to develop integrated solutions to long-term preservation of massive data collections, especially engineering and scientific data. Important requirements include the support for migration strategies, with a strong focus on preserving authenticity and integrity.

Therefore GRITO addresses the digital preservation problem from a bottom-up perspective, focusing on detailed technical issues, while SHAMAN represents a top-down perspective, addressing business and organizational models, but not ignoring the related technical challenges.

A common ground between these two initiatives is the decision to use the iRODS<sup>3</sup> data grid technology as a storage substrate for digital preservation. In this paper, we analyze the main threats to digital preservation, and whether iRODS is designed to withstand them. Furthermore, we identify a central point of failure in the iRODS architecture that could undermine our preservation goals, and we propose a solution that takes advantage of the extensibility present in the design of iRODS.

The remainder of this paper is organized as follows. Section 2 explains or motivation for the use of data grid technology for digital preservation. In section 3 we propose a taxonomy of threats to digital preservation. In section 4 we describe the iRODS data grid, pointing-out in section 5 its possible vulnerabilities for digital preservation, according to our taxonomy. In section 6 we propose an extension to iRODS to avoid those vulnerabilities. Finally, in section 7 we list the open issues of the proposed extension and conclude.

## 2 Motivation

The complexity of digital preservation increases with the fact that each type of digital object has its own specific requirements. For instance, the preservation of audio files requires having to deal with compression and complex encodings, unlike the preservation of XML files. Several communities, like biology, medicine, geographical sciences, engineering or physics, manage large amounts of structured datasets of data captured by sensors, physical or mathematical simulations generated by large computations, and also specialized documents reporting work progress and conclusions to researchers. That information can be represented in a wide range of formats (e.g., a researcher can use a specific input and output format, and a specific program to produce simulations) and include a large number of relations that are not expressed in the data models. Moreover, the

<sup>2</sup> <http://www.shaman-ip.eu>

<sup>3</sup> <https://www.irods.org>

collaborative environment of the scientific community, and associated services and infrastructures, usually known as *e-Science* (from "enhanced Science") [7], implies that interoperability and data sharing are required.

## 2.1 Data Grids

In recent years, there have been research efforts to define a new type of systems that deal with the large scale management, sharing and processing of data. These were commonly called Data Grids [5]. Data Grids offer a distributed infrastructure and services that support applications that deal with massive data blocks stored in heterogeneous distributed resources [3].

Grid computing is growing fast. Many applications of this technology exist, and Grid frameworks are already common in scientific research projects, enterprises, and other environments that require high processing power while using low-cost hardware. A possible definition of a Grid computing system [4] is one where: (i) resources are subjected to decentralized control; (ii) standard, open, and general purpose protocols and interfaces are used; and (iii) nontrivial qualities of service are delivered (e.g., combined throughput or response time).

In Data Grids, data is organized into collections or datasets, and is replicated, managed and modified using a specific management system. Information about replicas is usually organized in a replica catalog.

In summary, the common characteristics of a Data Grid can be described as: (i) **Massive Datasets:** a Data Grid allows the management and access to enormous quantities of data, in the order of terabytes or even petabytes (e.g., scientific projects such as the Southern California Earthquake Center, can generate, in a single simulation, up to 1.3 million files and 10 terabytes of data [8]); (ii) **Logical Namespace:** the requirements for scalability imply the use of virtual names for resources, files and users; (iii) **Replication:** scalability and reliability require high availability and redundancy; (iv) **Authorization and Authentication:** Due to the high value and frailty of the data, authentication and authorization mechanisms must be enforced to comply with authenticity and integrity requirements.

## 2.2 Data Grids and Digital Preservation

Grids are built using middleware software that makes fundamental aspects such as file management, user management and networking protocols, completely transparent. These goals are also shared by digital preservation systems.

SRB - Storage Resource Broker<sup>4</sup>, is a grid technology that has been operational for more than a decade, and is used in many research projects, storing petabytes of managed data. However, SRB is a generic grid infrastructure, and any modification to the management of data needs to be hard coded. The iRODS data grid is being developed by the same team that worked on SRB. The purpose is to create a system with an adaptive middleware that simplifies the task

---

<sup>4</sup> <http://www.sdsc.edu/srb>

of modifying how data is managed, or creating new policies tailored to a particular application, while retaining the good practices and lessons learned from SRB. However, neither SRB nor iRODS address specific requirements for digital preservation, which is the problem being addressed by this paper.

### 3 Digital Preservation Threats

In this section we present a taxonomy of threats to digital preservation based on several papers that point out different threats [1,2,10].

Our taxonomy is presented in Table 1. Component failures enclose the technical problems in the infrastructure’s components. Management failures are the consequences of wrong decisions. Finally, disasters and attacks correspond, respectively, to non-deliberate and deliberate actions affecting the system or its components.

Some threats cannot be detected immediately, remaining unnoticed for a long time. For instance, a damaged hard disk sector can remain undetected until a data integrity validation or hard disk check is performed. Moreover, we can not assume threat independence. For instance, a natural disaster like an earthquake can produce other threats.

**Table 1.** Threats to preservation systems

Component failures	Management Failures
Media faults	Organization failures
Hardware faults	Economic failures
Software faults	Media/Hardware obsolescence
Communication faults	Software obsolescence
Network services failures	
Disasters	Attacks
Natural Disasters	External attacks
Human operational errors	Internal attacks

In the next sections we further divide each threat into a set of possible specific events.

#### 3.1 Component Failures

**Media faults** occur when a storage media fails partially or totally, losing data through disk crashes or “bit rot”. Other hardware components can suffer **hardware faults** by transient recoverable failures, like power loss, or irrecoverable failures, such as a power supply unit burning out.

Similarly, **software faults**, usually known as *bugs*, can cause abrupt failures in the system. For instance, a firmware error can cause a data loss in hard drives. **Communication faults** occur in packet transmission, including detected errors (e.g., IP packet error) and undetected checksum errors. Other **network services failures**, such as DNS problems, can compromise the system availability.

### 3.2 Management Failures

An organization responsible for a preservation system may become unable to continue operating at the desired level due to sudden financial limitations (**economic failure**), political changes or any other unpredictable reason (**organization failure**). Moreover, failures can also occur due to incompetent management.

A different kind of management failure occurs when, even if the internal components do not fail over time, they become obsolete and unable to interact with the exterior components. Thus, unforeseen **media, hardware** or **software obsolescence** limits the system interoperability.

### 3.3 Disasters

**Natural disasters** such as earthquakes or fires can cause failures in many components simultaneously. For example, an earthquake may cause a data center to be destroyed or a wide-scale power failure. Accidentally **human operational errors** might introduce irrecoverable errors. For instance, people often delete data by mistake. Additionally, humans can cause failures in other components such as hardware (accidentally disconnecting a power cable) or software (uninstalling a needed library).

### 3.4 Attacks

Attacks might encompass deliberate data destruction, denial of service, theft, modification of data or component destruction, motivated by criminal, political or war reasons, including fraud, revenge or malicious amusement. Systems connected to public networks are especially exposed to **external attacks**, such as those caused by viruses or worms. Similarly, **internal attacks** might be performed by internal actors (e.g., employees) with privileged access to the organization and to the physical locations of the components.

## 4 iRODS Overview

The iRODS system is an open-source storage solution for data grids based on a distributed client-server architecture. A database in a central repository, called iCAT, is used to maintain, among other things, the information about the nodes in the Grid, the state of data and its attributes, and information about users. A rule system is used to enforce and execute adaptive rules. This system belongs to the class of adaptive middleware systems, since it allows users to alter software functionalities without any recompilation [9]. Figure 1 shows the UML [11] deployment diagram of iRODS. Note that the iCAT database only resides in the central node.

iRODS uses the storage provided by the local file system, creating a virtual file system on top of it. That virtualization creates infrastructural independence, since logical names are given to files, users and resources.

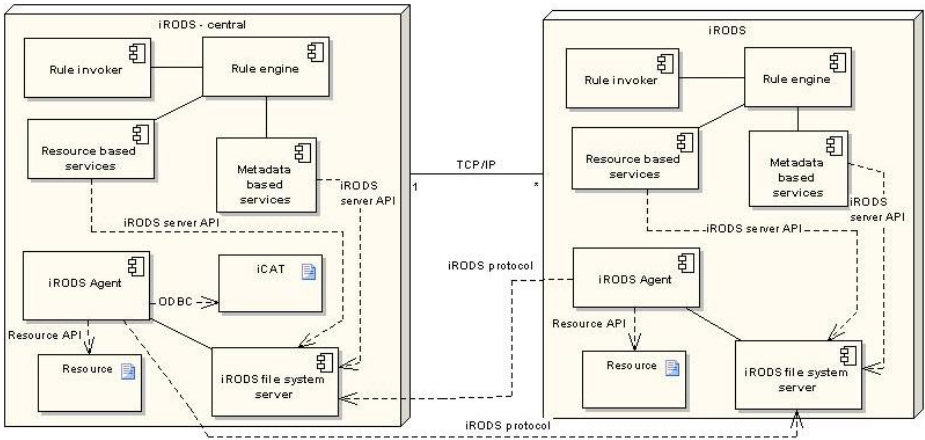


Fig. 1. iRODS deployment diagram

Management policies are mapped into rules that invoke and control operations (micro-services) on remote storage media. Rules can be used for access control, to access another grid system, etc. Middleware functions can be extended by composing new rules and policies.

## 5 Vulnerabilities in iRODS for Digital Preservation

iRODS presents some vulnerabilities if it is used as the basis for a digital preservation system.

In particular, the iCAT stores crucial information like the localization of nodes, the mapping between logical names and physical objects, information about rules, collections, data, metadata, etc. Consequently, iRODS is unable to work without the iCAT catalog, which means that it is a central point of failure. An unrecoverable failure in the metadata repository can cause total data loss, even if the data stored on other nodes remains intact. Table 2 summarizes how the digital preservation threats listed in Section 3 affect the overall iRODS system if these threats affect the iCAT.

Table 2. Threats to digital preservation in iRODS

Component failures	Loss	Management Failures	Loss
Media faults	Partial	Organization failures	None
Hardware faults	Partial	Economic failures	None
Software faults	Partial	Media/Hardware obsolescence	None
Communication faults	None	Software obsolescence	Total
Network services failures	None		
<b>Disasters</b>		<b>Attacks</b>	
Natural Disasters	Total	External attacks	Total
Human operational errors	Partial	Internal attacks	Total

Communication faults, network service failures, organization failures and economic failures can not directly affect the iCAT catalog. Media/hardware obsolescence can be easily avoided, as the iCAT catalog is managed by the open-source PostgreSQL database management system, which is able to run on several operating systems and hardware/software configurations. Consequently, a programmed replacement or migration of any obsolete component can surpass this threat.

Media, hardware and software faults just partially affect the iCAT catalog, since efficient short-term recovery strategies (e.g., iCAT backups, redundant RAID storage, etc.) can be used to recover from these types of failures. In these scenarios, the main issue may be the identification of the failure. For instance, a bit-rot in an iCAT file may become undetected for a long period of time, affecting part of the preservation system. We also consider that human operational errors can partly affect the iCAT catalog and consequently the preservation system.

Serious losses can occur from natural disasters, software obsolescence and external or internal attacks. An earthquake can destroy the centralized iCAT repository. If iCAT backups are also destroyed, a critical loss occurs affecting all the nodes in the system, because all the metadata was stored in the central repository. Natural disasters, external and internal attacks can affect the entire system if the event corrupts/destroys the iCAT repository and the short-term recovery support.

Since PostgreSQL is able to run on several hardware configurations, the media/hardware obsolescence is not a critical threat for digital preservation using iRODS. However, if PostgreSQL becomes obsolete, the iRODS system becomes unable to access the metadata stored in iCAT, which can potentially produce complete data loss. Thus software obsolescence turns the system dependent on a specific hardware/software configuration and consequently also fragile to hardware/software obsolescence. However, in these types of scenarios, the crux of the threat is software obsolescence.

## 6 Extending iRODS

In order to reduce the threats to iCAT, we propose an extension to iRODS, comprising of three new services: (*i*) iCAT Replication Service (iRep), consisting of replicating the iCAT directory in all the nodes of the data grid; (*ii*) iCAT Recovery Service (iRec), to recover the iCAT catalog in case of corruption or failure of the central node; and (*iii*) Audit Service, consisting of a system check that compares the iCAT with its replicas, and alerts any discrepancy, if detected.

### 6.1 Replication and Recovery

Figure 2 presents the UML activity diagram modeling the process of replicating the iCAT to other nodes in the data grid. The metadata repository is scanned for modifications. If there are any modifications, the system evaluates the rules to define if the replication should be postponed or proceed with a full or partial export. For instance, operations on metadata elements about nodes and data



have a higher priority than operations on elements about users. Moreover, we also distinguish between types of operations. Therefore, a delete operation is not as critical as an insert, because the loss of new metadata (e.g., mapping between logical and physical name) may imply the loss of new data in the grid. Based on the type of the iCAT element modified and on the operation performed (delete, update or insert) we define a priority level to the iCAT replication. After low priority modifications have been made, the iCAT replication is postponed. For high priority levels, we also evaluate the current workload of the system, which determines if the export should be full or just partial.

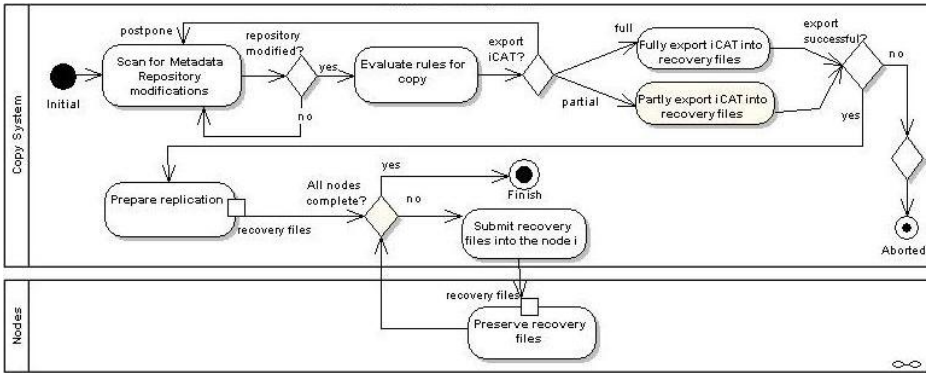


Fig. 2. Activity diagram of the iCAT replication process performed by iRep

If the conditions are met, the current iCAT contents are partly (recent modifications and nearest records) or fully (all data, schema, and the list of nodes with the replicas) exported from the repository into a set of recovery files. Then, the recovery files are replicated to all the other storage nodes registered in the data grid. Local nodes are responsible for the preservation of recovery files, which are stored in a specific area outside the control of the data grid.

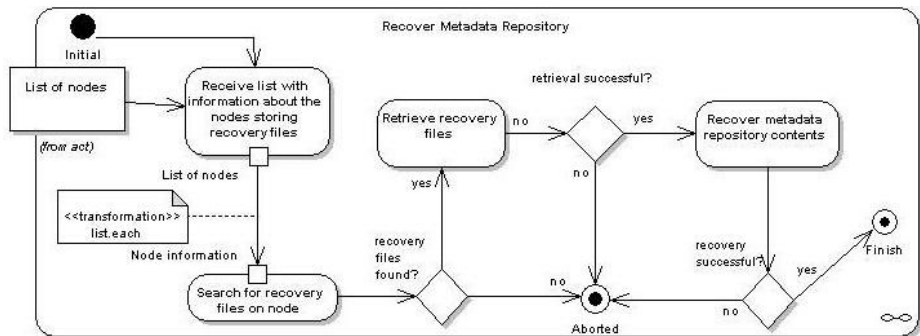


Fig. 3. Activity diagram of the iCAT recovery process performed by iRec

If a problem occurs and the iCAT repository becomes unavailable, a recovery process can be performed, as described in Figure 3. This process is executed by a component external to iRODS. Depending on the scenario, it may be necessary to install a new instance of iRODS (e.g., if the node crashed permanently) or just to recover the iCAT catalog. To proceed with the iCAT catalog recovery, a list of the nodes that are storing recovery files should be given as an input (it can be retrieved from any node that had survived). Then, all the nodes are asked to send the recovery files. Those that are able to do it (e.g., those that survived, in case of a major disaster) will send the recovery files to the central node, where are compared for integrity validation, and the iCAT catalog is rebuilt.

When the recovery process is completed, the data grid moves into a state where all the data stored in undamaged nodes is available again.

## 6.2 Audit

The audit process checks the integrity of the iCAT catalog. When this process is started, the iCAT catalog remains accessible and the list of nodes in the data grid is obtained from the current iCAT contents. The central node exports the current contents of the iCAT catalog into a set of recovery files. Then, the recovery files are submitted to all nodes, asking them to compare the submitted files with the latest version preserved in the local node. Thus, the iCAT check is computed locally, using parallel processing provided by the Grid infrastructure.

In case of discrepancies, the local node produces a log with the list of detected inconsistencies and sends it to the central node. Finally, the central node is responsible for notifying the administrator (e.g., by email) confirming the success of the audit process or sending the log files. Note that the audit process does not produce any modification in either the iCAT or the recovery files stored in data grid nodes. Consequently, the system state remains unchanged with this process.

Our solution executes both the replication, recovery and audit processes as iRODS modules implemented with external micro-services. We can materialize these modules as new components (replicate, recover and audit), thus extending the iRODS architecture represented in the deployment diagram of Figure 1. We decided to implement these as micro-services because: (i) it keeps the module implementation external to the iRODS core; and (ii) they can use the iRODS API to access resources and the iCAT.

## 7 Conclusion and Open Issues

Digital preservation in e-Science may require the use of data grids to manage the large and continuously growing amount of data. However, current data grids do not support natively key digital preservation techniques such as, for example, auditing. iRODS is a good starting point for building a preservation solution based on data grids because of its extensibility, due to the possibility to include new micro-services and rules.

This paper presented a taxonomy of threats to digital preservation. Based on that, we identified that iRODS had a central point of failure in the metadata

catalog. Consequently, we presented an extension to the iRODS system to handle digital preservation threats to the metadata catalog.

The proposed solution still has some open issues that must be addressed. For example, the format used to export the schema and contents of iCAT must be defined in a way that makes itself appropriate for preservation. In this moment we are using the XML format, but a better defined XML-Schema needs to be defined.

Another open issue is to establish the policies that define when the repository should be replicated. On the one hand, for performance reasons, it is unfeasible to replicate the repository on every single change to the iCAT database. On the other hand, a long replication period (e.g., daily) may imply important updates being lost in case of an iCAT failure. Thus, this process should be balanced with the normal operations on the grid and user-definable by a set of rules. For instance, the administrator may be able to define the maximum number of pending non-replicated transactions, the admissible workload to perform a replication in parallel with normal operations, etc.

We are using and validating the proposed solution in the context of project GRITO, focusing on data objects from the National Digital Library<sup>5</sup> and scientific data provided by the Portuguese National Laboratory of Civil Engineering<sup>6</sup>. The case of the scientific data will be further analyzed in the context of project SHAMAN.

## Acknowledgments

This work is partially supported by the projects GRITO (FCT, GRID/GRI/81872/2006) and SHAMAN (European Commission, ICT-216736), and by the individual grant from FCT (SFRH/BD/23405/2005) and LNEC to José Barateiro.

## References

1. Baker, M., Keeton, K., Martin, S.: Why traditional storage systems don't help us save stuff forever. In: 1st IEEE Workshop on Hot Topics in System Dependability, June 30 (2005)
2. Baker, M., Shah, M., Rosenthal, D.S.H., Roussopoulos, M., Maniatis, P., Giuli, T.J., Bungale, P.P.: A fresh look at the reliability of long-term digital storage. In: EuroSys, pp. 221–234 (2006)
3. Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., Tuecke.: The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications* 23, 187–200 (2000)
4. Foster, I.: What is the grid? A three point checklist. *GRIDToday* 1(6) (July 2002)
5. Hey, T., Trefethen, A.E.: The uk e-science core program and the grid. In: International Conference on Computational Science (1), pp. 3–21 (2002)
6. IEEE. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries (1990)

<sup>5</sup> <http://www.bn.pt>

<sup>6</sup> <http://www.lnec.pt>

7. Miles, S., Wong, S.C., Fang, W., Groth, P., Zauner, K.-P., Moreau, L.: Provenance-based validation of e-science experiments. *Web Semant.* 5(1), 28–38 (2007)
8. Moore, R.: Digital libraries and data intensive computing. In: China Digital Library Conference, Beijing, China (September 2004)
9. Rajasekar, A., Wan, M., Moore, R., Schroeder, W.: A prototype rule-based distributed data management system. In: HPDC workshop on Next Generation Distributed Data Management, Paris, France (2006)
10. Rosenthal, D.S.H., Robertson, T., Lipkis, T., Reich, V., Morabito, S.: Requirements for digital preservation systems: A bottom-up approach. CoRR, abs/cs/0509018 (2005)
11. Unified, U.: modeling language specification, version 1.4.2 formal/05-04-01. ISO/IEC 19501 (January 2005)

# A User-Oriented Approach to Scheduling Collection Building in Greenstone

Wendy Osborn<sup>1</sup>, David Bainbridge<sup>2</sup>, and Ian H. Witten<sup>2</sup>

<sup>1</sup> Department of Mathematics and Computer Science  
University of Lethbridge  
Lethbridge, Alberta, Canada  
`wendy.osborn@uleth.ca`

<sup>2</sup> Department of Computing Science  
University of Waikato  
Hamilton, New Zealand  
`{davidb,ihw}@cs.waikato.ac.nz`

**Abstract.** We propose a user-oriented approach for the automated and scheduled maintenance of Greenstone digital library collections. Existing systems require the user either to add new data manually to a collection, or to have programming knowledge in order to use existing application programming interfaces (APIs) in order to automate scheduled collection updates. The Greenstone Scheduler can automate the construction of any existing collection, and schedule the construction to occur periodically. This is accomplished through incorporating a module specific to this purpose into the Greenstone Librarian Interface.

## 1 Introduction

A wide range of applications generate multimedia documents, such as images and video on a daily basis. For example, many municipalities have a photo-radar program to catch vehicles traveling over the speed limit, or traveling in bus lanes during rush hour. A vast number of pictures of vehicle license plates are created every day. If these images are organized into a digital library, for instance, the collection would need to be updated regularly to incorporate new images. Another example of an application that requires periodic updating includes a resource of information across several post-secondary institutions [6].

When documents and metadata are added to a digital library collection on a regular basis, such as hourly or daily, an automated and scheduled approach to collection maintenance is preferred. An automated approach should not be time consuming for the user, leaving their time available for other important tasks.

Existing digital library software such as DSpace [8] and Fedora [2] require that items be added manually to the collection. In Fedora, data is retrieved at the time of viewing. However, the location needs to be manually configured. Further, although Fedora and DSpace do provide application programming interfaces (APIs) to extend functionality, programming knowledge is required for using an API and setting up tools based on it.

A module for automating and scheduling collection maintenance in Greenstone has been proposed recently [5]. The Scheduler can automate the construction of any existing collection, and schedules the construction to occur on an hourly, daily or weekly basis. In addition, the owner of a collection is still free to update the collection manually whenever they want. Further, the Scheduler interacts with the existing task scheduling mechanism on the host system, which keeps the Scheduler minimal, yet powerful.

In this paper, we present the incorporation of the Scheduler into the Greenstone Librarian Interface [10], a user-friendly graphical interface for creating digital library collections. By providing a user interface for the Scheduler, we improve its usability and provide further abstraction of the scheduling process.

## 2 Greenstone

Greenstone [11] is a software suite for creating digital library collections and making them available globally via the internet or removable media. A collection can contain documents of different formats. Over thirty formats are supported through a plugin architecture including images, postscript, PDF, audio, and formatted and unformatted text that extract full text and metadata where possible. A Greenstone collection can be customized in many ways. For example, a collection owner can customize the types of documents that can appear in the collection, the appearance of the interface of the collection, and the indices and classifiers that other users can use to access the collection.

A collection is created or modified by following four steps:

1. *document addition*. New documents that will become part of a collection are copied into the import directory for the collection.
2. *document importation*. The documents in the import folder are now processed for the collection by specifying an import command. Different importing options can be specified by the user. All imported documents are located in the archive directory.
3. *accessor creation*. Indices and classifiers are created by using a build command. The resulting indices and classifiers are placed in the build directory.
4. *collection activation*. Finally, the collection is activated by copying the indices and classifiers from the building directory to the index directory.

## 3 Greenstone Librarian Interface

The Greenstone Librarian Interface [10] is a graphical user interface that provides a user-friendly method to build and configure Greenstone collections. It incorporates the four steps above. In addition, the Librarian Interface allows a user to create new collections, select metadata sets, configure which document types to allow, and select import and build command options.

Importing converts source document into a canonical XML format; building processes the canonical XML format to create the necessary full-text index and metadata files.

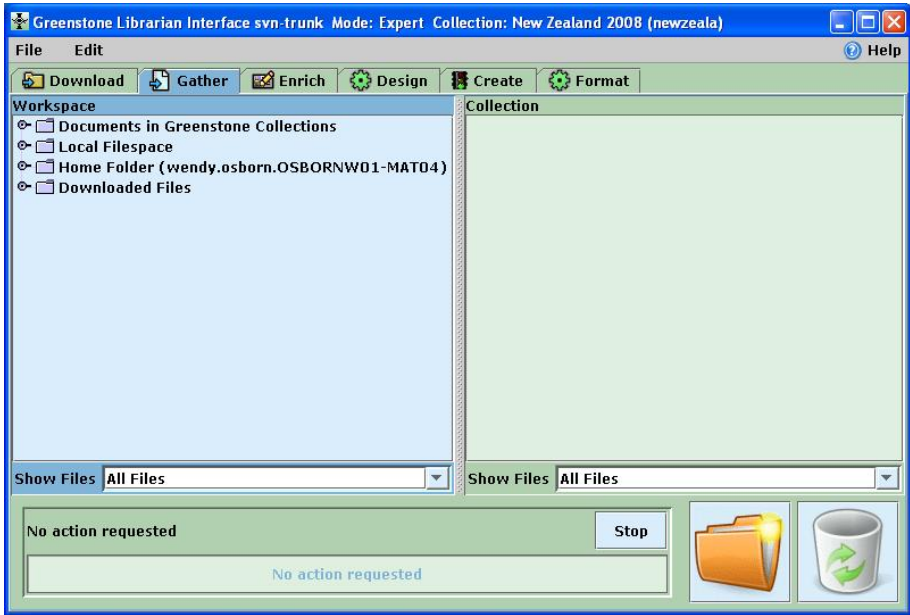


Fig. 1. The Greenstone Librarian Interface

### 3.1 Command-Line Scheduling

The Scheduler is a command-line program that is responsible for both creating a building script for a specific collection and for interacting with the scheduling service on the operating system [5]. The original version of the Scheduler takes as command-line arguments the name of the collection, the import and build commands (and all of their required arguments), and the frequency of execution (hourly, daily or weekly). The output from the Scheduler is a customized Perl script that rebuilds the collection, and modifications to the scheduling service to execute the script at the frequency specified. The scheduling service employed by the Scheduler is Cron [3], because it is available for all major platforms. Unix, Linux and Mac OS X run Vixie Cron [9]. Versions of Cron that exist for Windows includes Pycron [7].

For example, suppose we want to schedule a daily build of the collection *pics*. A call to the Scheduler would resemble the following:

```
schedule.pl pics "import.pl -removeold pics"
              "buildcol.pl -removeold pics" daily
```

Figure 2 shows the resulting build script and corresponding *crontab* record for the collection *pics*.

```
#!/usr/bin/perl
$ENV{'GSDLHOME'}="/home2/gsd1/gsd1";
$ENV{'GSDLOS'}="linux";
$ENV{'GSDLLANG'}="";
$ENV{'PATH'}="/bin:/usr/local/sbin:/usr/local/bin:/sbin:/usr/sbin:
/usr/bin:/usr/X11R6/bin:/usr/local/gsd1/bin/script:
/usr/local/gsd1/bin/linux";
system("import.pl -removeold pics");
system("buildcol.pl -removeold pics");
system("\rm -r /gsdl/collect/pics/index/*");
system("mv /gsdl/collect/pics/building/*
/gsd1/collect/pics/index/");
system("chmod -R 755 /gsdl/collect/pics/index/*");

00 0 * * * /gsdl/collect/pics/gsd1.pl
```

Fig. 2. Sample Building Script and Crontab Record [5](#)

## 4 Scheduling in the Librarian Interface

The Scheduler is a minimal, yet powerful tool for maintaining Greenstone collections. It hides the details concerning the collection building process, and also the details for scheduling a Cron task. However, the Scheduler has two main limitations. The first is that the user is still required to know the syntax and command-line arguments for both the import and build commands in order to perform scheduling. The second is that, currently, notification of collection building success—or failure—is still dependent on the version of Cron (and indirectly, the operating system it is running on) that is used.

The Librarian Interface provides a user-friendly tool for building collections, including configuring the import and build commands. Therefore, it is ideal to extend the Librarian Interface to allow the configuration of the Scheduler as well. Figure [4](#) depicts the Librarian Interface extension to support the scheduling of collection builds. Currently, the Create panel is displaying most of the arguments (i.e. Schedule Options) for the Scheduler.

Notice that no explicit options exist for the import and build commands. This is because the import and build commands are created based on the arguments selected by the user from the Import Options and Build Options. Therefore, the user no longer needs to know the exact syntax of the import and build commands!

## 5 Example: Collecting Pictures While Traveling

Before we discuss extensions to both the scheduler and the Greenstone Librarian Interface, we present a simple application of scheduling from the Librarian Interface. In this scenario, we have a traveler who wants to post pictures of their trip in a Greenstone collection for her friends to view. Instead of waiting until



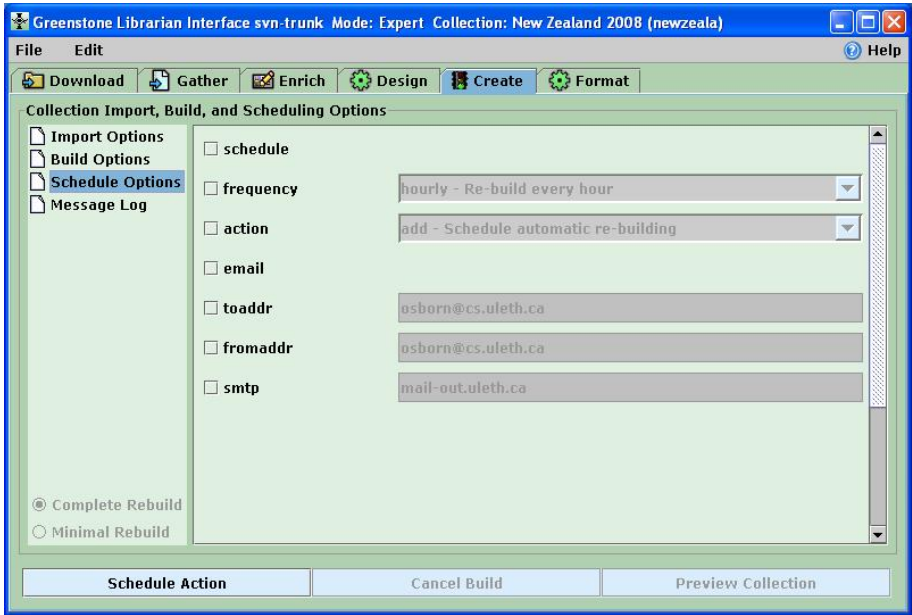


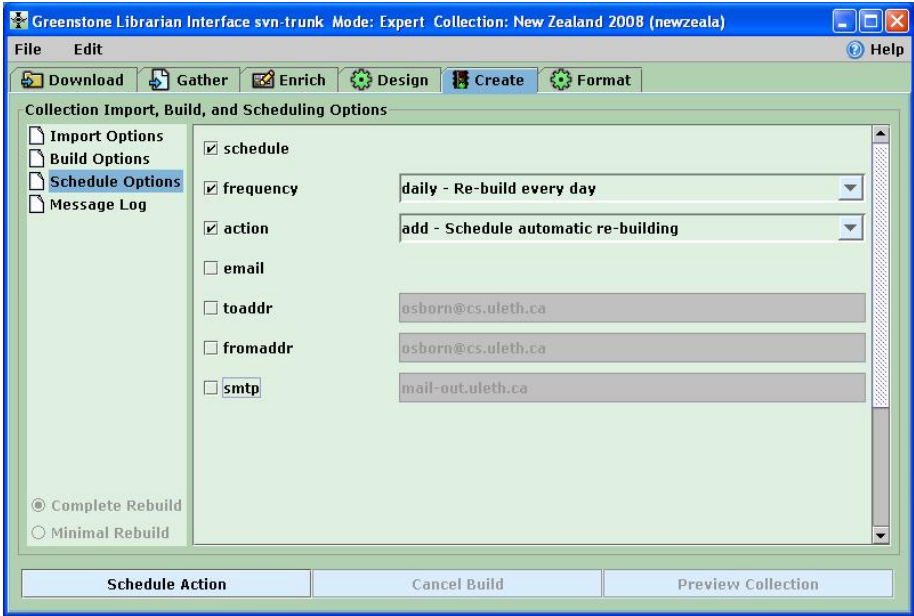
Fig. 3. The Schedule Options Pane

the end of the trip, the traveler wants to post her pictures from each day, incrementally adding to the collection on a daily basis. The traveler does not want to worry about obtaining the Librarian Interface to rebuild the collection while traveling. Instead, she simply wants to upload the pictures to the import folder of her collection, and have her collection rebuilt automatically and on a daily basis. This can be accomplished by setting up a scheduled, automatic rebuild of the collection of travel photos from the Librarian Interface.

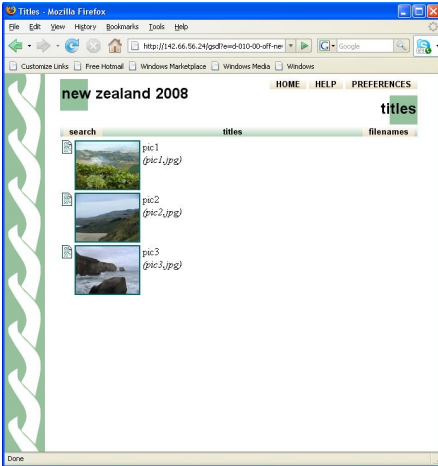
First, before departing, the traveler runs the Librarian Interface and creates a new collection. Then, the user selects the Create tab to display the collection creation pane. From here, clicking on Schedule Options will display the available options for setting up a scheduled, automatic collection build of the collection of travel pictures. Figure 4(a) depicts the available scheduling options, which are displayed with default and derived values as appropriate. Here, the traveler selects schedule, which indicates that she wants to set up a scheduled, automatic build. Also, she selects a frequency of hourly and an action of add (or, to create a new scheduled build).

Next, the traveler clicks on Schedule Build. This will set up the building script for the collection of travel pics, as well as the *crontab* record that will indicate to Cron that the collection is to be rebuilt daily. The collection is now ready to be re-built while the traveler is away.

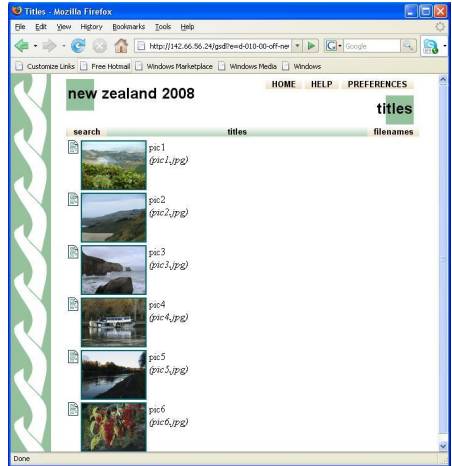
At the end of the first day of travel, she uploads three pictures, which are added to the collection when the collection is re-built automatically overnight. The updated collection is depicted in Figure 4(b). The next day, the traveler



(a) Scheduling Options Pane



(b) Build - First Day



(c) Build - Second Day

Fig. 4. Collection Build Scheduling

uploads three more pictures. When the collection is re-built overnight, these pictures are added to the existing collection. Figure 4(c) depicts the updated collection with the new pictures.

Although not shown here, the user can switch to the Import Options and Build Options and select any options that are required for collection importing

and building. These options are incorporated into the automatic building script that is created for the collection. In addition, the user can manually build and configure their collection as many times as necessary to confirm the right sequence is being performed, before setting up a scheduled, automatic build.

## 6 Implementation

In order to enable scheduled, automatic building from the Librarian Interface, several issues needed to be addressed. Some issues require modification of the Greenstone scheduler itself, while others are modifications to the Librarian Interface. We present and discuss the top issues here.

### 6.1 User Modes

The Librarian Interface has four modes of use— Library Assistant, Librarian, Library Systems Specialist, and Expert—stepwise increasing the functionality available to the user [10]. It was important to determine which modes would have access to the Scheduling Options, and for those modes that were granted access, how much access each would be granted. It was decided that Library Systems Specialists and Experts would be provided access to the Scheduling Options. In addition to determining which options to make available to each user mode, it is also necessary to determine how the Scheduler would interact with the existing import and build functionality in the Librarian Interface.

**Mode Level for Options.** The Scheduler was first modified so that the passing of command-line arguments conformed with that of other Greenstone commands, such as import and build. This also allowed us to easily specify which user mode could have access to each argument when it is added as a Schedule Option in the Librarian Interface. The Expert user mode is granted full access to all Schedule Options. The Librarian Systems Specialist is only granted limited access to options. This user mode can only specify whether or not to schedule a collection build with the default values—a frequency of hourly, and no sending of email.

**Scheduling and Building.** Another important decision that needed to be made was how the Scheduler would interact with the collection building functionality of the Librarian Interface. The question asked was, should scheduling be done at the same time as collection building, or be a completely separate task?

For Expert mode, the functionality for Scheduling is separate from that of collection building. This is because an expert user may want to configure and manually re-build their collection before scheduling an automatic re-build of it. For Library System Specialist mode, the Scheduling functionality is done at the same time as collection building. If the specialist chooses to schedule, a new scheduled build is created. If the specialist chooses not to schedule, any existing scheduled builds are deleted.

## 6.2 Cron Event Logs

It is important to maintain logs that keep track of the outcome of a scheduled collection build. Both Vixie Cron and Pycron maintain a log that keeps track of the attempted execution of all scheduled tasks. Neither scheduling service keeps track of the success or failure of a scheduled task, nor do they keep track of the output of a task. In addition, to view logs created by Vixie Cron requires the user to have root access.

Another desirable feature of maintaining logs is to be able to only record output that is considered important, and to disregard all other task output. For example, if the output from the input and build commands of Greenstone is required, but the output from moving indices and classifiers is not considered important, the log should reflect this.

Therefore, the Greenstone scheduler is modified in two ways to handle the logging of building script output. The first is to create a custom log for each execution of the building script. The collection building script creates a unique filename every time it is executed by using a timestamp. The second is to specify in the collection building script which actions will have its output redirected to the logfile. The actions whose output is to be ignored will have its output redirected to the ‘bit bucket’ (e.g. `/dev/null/` in Linux).

## 6.3 Email Notification

It is also important than an email notification service be provided, which will inform users of the outcome of their scheduled collection build. We handle email notification from the Scheduler and Librarian Interface for the following reasons:

1. *User notification.* Whether Cron notifies users about the outcome of a scheduled task depends on the its implementation. For Vixie Cron, the outcome of a task (i.e. output from either successful task completion, or error output) is emailed to the owner of the task. Pycron does not send email notification.
2. *Flexibility of Notification.* In Vixie Cron it is possible to suppress email notification, either by setting an environment variable to null, or by redirecting all output to a file or the ‘bit bucket’. However, this is normally an all-or-nothing event—either all output, or none, is sent by email. Similar to logging, a desirable feature would be to send email that contains only the most important parts of the building process, and ignores other parts of the building process.
3. *Greenstone Email Support.* Greenstone comes with a Perl email script, that is a wrapper for the Perl `sendmail` command. The email script is platform-independent. Therefore, it is ideal to use it to provide a uniform way to send email concerning the execution of a scheduled task. In addition, the script does not require the piping of task output directly to it, but instead can send the contents of a file.

Therefore, both the Librarian Interface and the Scheduler are modified so that email notification is handled in a uniform and user-friendly manner across all operating systems.

First, options have been added to the Scheduler that are required for the Perl email script—specifically, a flag to specify that email will be sent (-email), the sender (-fromaddr), receiver (-toaddr) and the email server that will be used to send the email (-smtp). Also, the corresponding fields exist in the Schedule Options pane of the Librarian Interface. In order to assist users in using the email features of scheduling, the Librarian Interface attempts to populate the fields -toaddr, -fromaddr, and -smtp in the Schedule Options pane by consulting the configurations for the Librarian Interface and Greenstone. If suitable values are available from these sources, they are assigned to the appropriate fields.

Second, the output from the building script must be captured and re-directed to the Perl email script. The capturing of output already takes place, in the event log. This serves as the file that the email script will send to the user. Also, since it contains only the output that is considered important, this will be reflected in the email message as well.

Finally, the Scheduler is modified so that, if specified, the generated build script will send an email message containing the contents of the log to the specified recipient. An added bonus is that if email is not specified, the log can still be consulted by the user if required.

## 6.4 Scheduled Building in Isolation

An important modification to the Scheduler is to ensure that a scheduled build is completed in its entirety without interference from another scheduled build. A build may take a significant amount of time depending on the size of the collection—from seconds for a small one to 33 hours for a collection containing 20 GB of raw text and 50 GB of metadata [1]. To handle this, the Perl script for the collection checks for a lock file, which indicates that a collection build is underway. If the file exists, the collection owner is notified via email and information is placed in the event log. Otherwise, a lock file is created before the scheduled build begins, and is removed when the build finishes. This ensures that multiples builds do not occur concurrently.

## 7 Conclusion

In this paper, we propose and discuss the incorporation of the Greenstone Scheduler into the Librarian Interface. This overcomes two limitations of the Schedule—the requirement to know the syntax and command-line arguments for both the import and build commands, and the inconsistency of notification of collection building success or failure. Providing an interface to the Scheduler improves its usability and provide further abstraction of the scheduling process from the user.

Some future directions of work include the following. The first is to allow the user to select a specific period of time (e.g. 20 minutes after the hour) for their collection to be re-built. Currently, collection building occurs at the top of the hour (hourly), at midnight (daily) and on Sunday at midnight (weekly). The

second is support for dependencies between fields in the Librarian Interface. For example, if a user selects the email option, it requires -toaddr, -fromaddr, and -smtp. Currently, the user must ensure that these are also selected, as it is not done automatically.

## References

1. Boddie, S., Thompson, J., Bainbridge, D., Witten, I.H.: Coping with very large digital collections using greenstone. In: Proceedings of the ECDL Workshop on Very Large Digital Libraries (September 2008)
2. Lagoze, C., Payette, S., Shin, E., Wilper, C.: Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries* 6(2), 124–138 (2006)
3. Nemeth, E., Snyder, G., Hein, T.R.: *Linux Administration Handbook*. Prentice-Hall, Englewood Cliffs (2007)
4. New Zealand Digital Library Project. Greenstone digital library software (last visited, July 2008), <http://www.greenstone.org>
5. Osborn, W., Fox, S.: Automatic and scheduled maintenance of digital library collections. In: Proceedings of the 2nd IEEE International Conference on Digital Information Management (ICDIM 2007) (October 2007)
6. Osborn, W., Fox, S., O'Shea, S.: A unified resource for post-secondary program information. In: Proceedings of the 2008 International Conference on Information Resources Management (Conf-IRM) (2008)
7. Schapira, E.: Pycron cron - great cron for windows (last visited, July 2008), <http://sourceforge.net/projects/pycron>
8. Tansley, R., Bass, M., Smith, M.: Dspace as an open archival information system: Status and future directions. In: Proceedings of the 10th European Conference on Digital Libraries (ECDL 2005) (September 2006)
9. Vixie, P.: Vixie cron for FreeBSD (last visited, July 2008), <http://www.freebsd.org/cgi/cvsweb.cgi/src/usr.sbin/cron/>
10. Witten, I.H.: Creating and customizing collections with the Greenstone Librarian Interface. In: Proceedings of the International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society (March 2004)
11. Witten, I.H., Bainbridge, D.: *How to Build a Digital Library*. Morgan Kaufmann, San Francisco (2002)

# LORE: A Compound Object Authoring and Publishing Tool for the Australian Literature Studies Community

Anna Gerber and Jane Hunter

University of Queensland  
St Lucia, Queensland, Australia  
(617) 3365 1092  
{agerber, jane}@itee.uq.edu.au

**Abstract.** This paper presents LORE (Literature Object Re-use and Exchange), a light-weight tool which is designed to allow scholars and teachers of Australian literature to author, edit and publish compound information objects encapsulating related digital resources and bibliographic records. LORE enables users to easily create OAI-ORE-compliant compound objects, which build on the IFLA FRBR model, and also enables them to describe and publish them to an RDF repository as Named Graphs. Using the tool, literary scholars can create typed relationships between individual atomic objects using terms from a bibliographic ontology and can attach metadata to the compound object. This paper describes the implementation and user interface of the LORE tool, as developed within the context of an ongoing case study being conducted in collaboration with AustLit: The Australian Literature Resource, which focuses on compound objects for teaching and research within the Australian literature studies community.

**Keywords:** OAI-ORE, IFLA FRBR, Bibliographic Ontologies, Named Graphs.

## 1 Introduction and Background

Within the discipline of literature research and teaching, the ability to relate disparate digital resources in a standardized, machine-readable format has the potential to add significant value to distributed collections of literary resources. Such compound objects can be used to: track the lineage of derivative works which are based on a common concept or idea; or to relate disparate objects that are related to a common theme; or to encapsulate related digital resources for teaching purposes. For example, one might want to relate the original edition of *Follow the Rabbit-Proof Fence* to the illustrated edition, a radio recording and a digital version of the film based on the novel – and enable them to be retrieved and presented, with their relationships visualized, regardless of their location.

Our objective is to provide a tool to enable such an encapsulation and subsequent re-use and visualization, by building on the efforts of two previous digital library initiatives:

- The IFLA Functional Requirements for Bibliographic Records (FRBR)[1]
- The OAI-Object Reuse and Exchange (OAI-ORE) [2]

The ability to easily share and exchange such compound mixed-media digital objects will facilitate collaborative eScholarship and discussion amongst researchers of Australian literature.

The IFLA Functional Requirements for Bibliographic Records (FRBR) is a 1998 recommendation of the International Federation of Library Associations and Institutions (IFLA) to restructure catalog databases to reflect the conceptual structure of information resources. It uses an entity-relationship model of metadata for bibliographic resources that supports four levels of representation: work, expression, manifestation and item. It also supports three groups of entities: products of intellectual or artistic endeavour (publications); those entities responsible for intellectual or artistic content (a person or corporate body); and entities that serve as subjects of intellectual or artistic endeavor (concept, object, event, and place).

The Open Archives Initiative Object Reuse and Exchange (OAI-ORE) is an international collaborative initiative, focusing on an interoperability framework for the exchange of information about Digital Objects between cooperating repositories, registries and services. OAI-ORE aims to support the creation, management and dissemination of the new forms of composite digital resources being produced by eResearch and to make the information within these compound digital objects discoverable, machine-readable, interoperable and reusable.

Named Graphs [3] are endorsed by the OAI-ORE initiative [4] as a means of publishing compound digital objects that clearly states their logical boundaries. When applied to compound objects, the nodes in the Named Graph correspond to the individual aggregated resources, and the arcs correspond to typed relationships between those resources. In the terms of the OAI-ORE, compound objects correspond to ORE aggregations, and the Named Graphs that describe them to ORE resource maps. Resource maps and their component nodes and arcs are all web resources which can be identified and unambiguously referenced by HTTP URI handles, thus providing a basis for reuse and exchange. Our hypothesis is that OAI-ORE Named Graphs provide the ideal mechanism for representing literary compound objects that encapsulate the entities and relationships expressed by the IFLA FRBR. They do this in a way that is discipline-independent but provides hooks to include rich semantics, metadata and discipline-specific vocabularies, ontologies and rules.

In developing LORE, we aim to apply OAI-ORE to eScholarship within the discipline of Australian Literature through a case study involving the creation, exchange and re-use of compound digital objects for the purposes of teaching and research within the Australian literature studies community. In addition, the LORE services enable users to label the nodes and arcs within an OAI-ORE compound object using an ontology of classes and relationships which is based on the IFLA FRBR, but extended to support new types of entities and relationships, specific to certain sub-communities.

## 2 Case Study

AustLit [5] is a non-profit collaboration between the National Library of Australia and twelve Universities. It provides an important resource for scholars undertaking research into many aspects of Australian literary heritage and print culture history. AustLit serves the research, teaching and general information-seeking communities as a source of information about Australian literary works and the people and



organizations involved in their creation or publication (agents), and also as a mechanism for the extraction and dissemination of research data.

AustLit supports the activities of numerous distributed research sub-communities around Australia and the world. These communities include: Black Words, focusing on Australian Indigenous literary cultures and traditions; Australian Popular Theatre, focusing on variety theatre of the 19<sup>th</sup> and early 20<sup>th</sup> Century; Australian Responses to Asia, focusing on Australian creative works about or referring to Asia; and many others, each built around an area of specialist literary research.

The AustLit data model is based on IFLA FRBR. The data model adopts the core work, expression and manifestation record structure, with enhancements to represent additional metadata required by AustLit's specialist research communities. AustLit's implementation of IFLA FRBR has also been extended with event-awareness based on the Harmony ABC ontology [6], to represent events such as the creation of a work, realization of an expression or embodiment of a manifestation and the associated agents and event metadata [7]. The records are stored using a highly normalized Oracle database schema based on the structure of RDF and Topic Maps [8].

### 3 Objectives

In the current AustLit system, authoring and editing of records is restricted to AustLit staff and a few key members of the research sub-communities who have been trained to use the complex data entry interface. Sub-communities cannot create their own additions or extensions to the data model to record specialized research data – they must request changes to be made to the underlying AustLit database on their behalf. As specialized research activity within the AustLit community has increased, the proliferation of additions to the shared underlying data model has increased the complexity of the AustLit user interface (UI) for all users, making it even less accessible to scholars who have not been trained in its use.

Hence, our primary objective is to provide an intuitive method for the AustLit community to collaboratively author scholarly content that can be integrated with existing research practices and tools, including the AustLit web portal. More specifically, we aim to develop an easy-to-use, light-weight, in-browser tool to enable literary scholars to easily author machine-processable and human-understandable compound objects in a standardized format. Additional objectives are:

- to enable the authors to enter metadata describing these compound objects to enable their easy discovery and re-use;
- to enable the publishing of these compound objects in open access repositories so they can be readily shared and re-used;
- to enable the lineage of derived intellectual products to be documented and visualized through these compound objects.

### 4 Related Work

The SCOPE system was developed by Cheung et al [9] specifically to enable the authoring and publishing of Scientific Compound Objects – and to document the provenance of related scientific outcomes (e.g., data, models, publications). Our aim

is to develop a similar tool specifically for scholars of literature and related products. Although some previous work has been done using RDF to represent multimedia and hypertext presentations for e-Humanities applications (e.g., CULTOS [10]), this work does not combine the advantageous features of both OAI-ORE and IFLA-FRBR to capture or label the precise relationships between entities. An overview of other implementations and applications of IFLA FRBR is provided in [11]. Of particular relevance are extensions to the Greenstone digital library software, to provide a visual interface to enable librarians to enrich digital libraries with FRBR data [12], and an alerting service [13]. Existing efforts to implement OAI-ORE for eScience and eScholarship include FORSITE [14], eChemistry [15] and UIUC [16].

A significant focus of e-Humanities tools development has been on scholarly mark-up and annotation tools to attach interpretations to individual objects or parts of objects (e.g., a paragraph in a journal article). LORE takes the paradigm of annotation a step further, enabling authors to annotate or tag the relationships between multiple objects or resources with tags from a specific ontology (an extended OWL version of the IFLA FRBR model).

## 5 Implementation

In order to address our aims of providing a light-weight tool that can be assimilated with the AustLit web portal and existing research practices, we have adopted an implementation approach based on Web 2.0 technologies. LORE is implemented as a Mozilla Firefox extension, and a standalone version can also be installed on any web server to provide cross-browser support. Both versions are implemented using AJAX technologies (Asynchronous JavaScript and XML). UI elements in the Firefox version are implemented in XUL. The Firefox extension provides some advantages over the standalone version, as it can be customized to suit individual research needs via user preferences; it allows us to bypass security sandboxing that might be imposed on server-hosted JavaScript within institutional environments; and the extension framework provides an ideal mechanism for software distribution and updates which is familiar and accessible to our target community. LORE uses RESTful web services on a Sesame 2 RDF data store, running as a Tomcat Java Servlet, for storing and querying Named Graphs representing the compound objects.

The types for intra-aggregation relationships as well as metadata terms for aggregated objects are specified via an OWL ontology, which is configured at start-up from the user preferences. By examining all of the topics and topic relationships from the AustLit database, we developed an OWL ontology that describes the existing AustLit data model, which is based on IFLA FRBR. We use this as the default ontology in our case study. The AustLit ontology is presented in Appendix A. We also tested LORE with the FRBR RDF model [17] as well as other Bibliographic ontologies [18].

## 6 User Interface

Figure 1 shows the editing interface provided by the LORE Firefox extension. OAI-ORE resource maps are displayed in a graphical form, as in Figure 1, as well as RDF/XML, shown in Figure 2, which may be selected and copied for use with any

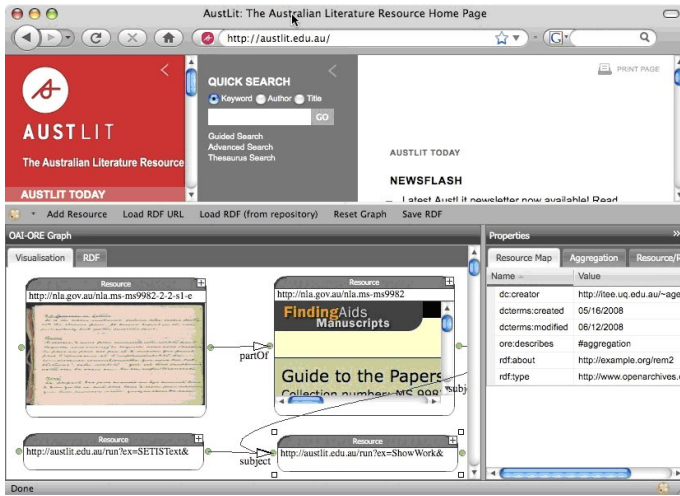


Fig. 1. Compound object editing interface: compound object about Patrick White’s *Voss*

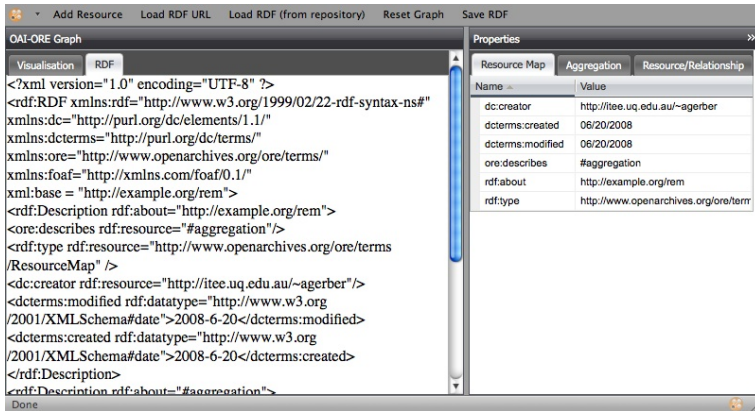


Fig. 2. RDF/XML view of resource map

RDF-enabled system. In the graphical representation, the *nodes* represent the individual atomic resources aggregated via the resource map and the *arcs* represent typed relationships between the aggregated resources.

Each node in the graphical view provides an interactive preview of the resource that it represents, allowing resources to be easily visually distinguished. The preview can be collapsed so that only the URI of a resource is visible, to conserve screen space in larger resource maps, as seen in the bottom two nodes of Figure 1. When a node is expanded, it may be resized by clicking and dragging with the mouse on the node border. This feature means that the preview can be made large enough for the user to view and interact with the resource directly from within the LORE editor rather than having to individually load aggregated resources via the main browser window.

Metadata about each resource aggregated by the resource map is displayed and can be edited via the *Resource/Relationship* tab in the properties view in the right-hand panel when the resource is selected. The other tabs in the properties sidebar allow metadata about the resource map and the aggregation that it represents to be specified.

An ontology that encodes the relationships and metadata terms specific to the application domain can be specified in the tool preferences. For our case study, we use this preference to load bibliographic ontologies including FRBR and our OWL representation of the AustLit data model. The preferences are also used to configure default values for the creator and the repository where compound objects are stored.

By default, only the metadata required by OAI-ORE is displayed in the properties view when creating a new compound object. A user may specify additional metadata by selecting from the properties menu, as shown in Figure 3. The metadata terms selectable from the menu are those from Dublin Core [19] and DCMI Metadata Terms [20], plus optional terms specified by OAI-ORE, as well as selected terms from FOAF [21] and for resources, datatype properties from the domain ontology.

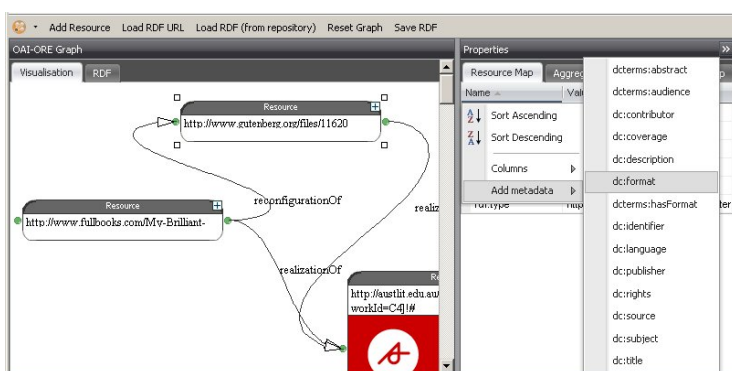


Fig. 3. Additional metadata menu

The types for the relationships represented by the arcs in the visualization are indicated by labels on each arc, and in the *Resource/Relationship* properties view. The type of a relationship can be changed by editing in the properties view or by selecting from the context menu of each arc. The arc context menu types are populated by the object properties from the domain ontology. Figure 4 shows some of the object properties from the AustLit ontology populating the arc context menu.

Resources to be added to the resource map may be discovered by browsing or searching via the Firefox browser. Clicking on the OAI-ORE logo in the status bar toggles the editor between hidden and visible, so that the full browser window can be used for resource discovery, whilst the resource map being constructed or modified remains readily accessible throughout the browsing session. A resource loaded in the browser can be added to the resource map by selecting *Add Resource* from the menu.

The menu also provides options for saving and loading compound objects. Resource maps that have been created or modified using LORE are saved as named graphs in the RDF repository specified in the preferences by selecting *Save RDF*. Resource maps can be loaded directly from a URL locating an RDF/XML representation

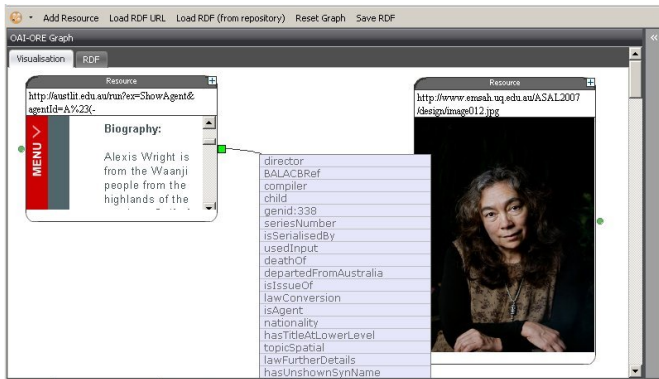


Fig. 4. Resource relationship types from the AustLit ontology

of the resource map by selecting *Load RDF URL*. Alternatively, resource maps can be loaded from the default RDF repository, by means of querying the repository for a named graph by its identifier through the *Load RDF (from repository)* menu option.

## 7 Discussion

### 7.1 User Interface Feedback

Feedback from the AustLit researchers with whom we have been collaborating has been very positive. They particularly liked the interactive node previews, the direct integration of the editor with the browser through the Firefox extension and the ease with which they could customize the relationship types and metadata supported by the editor by loading different domain ontologies. They would like to see support for additional types of arcs with distinguishing visualizations, e.g. differently shaped line decorators or arrow types, use of colors or line types, to highlight specific relationships, as well as explicit support for bi-directional relationships. The ability to identify a relationship context within a resource (for example, to relate a section of a document to a region within an image) was flagged as a priority. Some usability issues with the current interface were also identified, for example, the UI for adding additional metadata is not intuitive and will require redesign, and some of the terminology used in the UI is too technical for our target users. Feedback so far has been collected through walkthroughs and user testing involving three AustLit researchers. More extensive usability testing will be conducted as part of Aus-e-Lit.

### 7.2 Considerations Arising from the Case Study

The case study has raised the following issues for further consideration:

Objects can only be added to a compound object in the LORE editor if they can be loaded from a URL in the web browser. This is consistent with OAI-ORE, as it is designed for creating aggregations of web-accessible resources. However, not all objects that researchers may want to encapsulate in a compound object will have URIs that resolve to a web resource (URIs may be used purely as identifiers not as resource

locators), and not all objects can be identified by URIs, for example objects that exist within institutional repositories using local identifiers.

Because the IFLA FRBR ontology (and by extension, the AustLit ontology) is quite complex, it can be difficult for a Literary scholar who is not an expert in this model to apply the metadata terms and relationship types from the ontology to relate resources. For example, while they may understand the distinction between a FRBR *manifestation* and *item* in the case of a physical publication such as a book, they may not understand how these concepts apply to a digital resource, and may apply relationships or metadata from multiple levels (work, expression, manifestation or item) to a single resource. Strategies for addressing this issue could include adding more semantic checks to the UI to assist users in applying the ontology terms correctly, or simplifying the domain ontologies based on community needs and understanding.

### 7.3 Limitations and Future Work

The LORE tool that has been developed is a proof-of-concept prototype that demonstrates how OAI-ORE can be used to author, edit and publish compound objects directly from within the familiar interface of the web browser. Continued development of LORE is part of the Aus-e-Lit project, which aims to address the eResearch needs of scholars of Australian Literature. Further effort will be undertaken as part of this project to improve the robustness and usability of the system and to overcome existing limitations including:

- Fedora [22] support is being implemented so that the OAI-ORE objects can be published to a Fedora repository in addition to the existing RDF datastore functionality.
- We plan to add a search and retrieval interface, to enable discovery of compound objects through searches over the aggregated objects as well as over metadata terms and relationships. In addition to searches based on user-input, searches will be automated based on browser activity, for example, if the user navigates to a URL in the web browser, compound objects that aggregate the resource identified by that URL will also be displayed in the search results interface.
- LORE currently only supports uni-directional relationships. We will investigate distinguishing bi-directional, symmetric, transitive and reflexive relationships from the domain ontology within the editing UI.
- A rule engine can be used to infer additional indirect relationships between aggregated objects. For example, if a transitive property is asserted between objects A and B and also between objects B and C, we can infer that the same property exists indirectly between objects A and C. We intend to investigate how semantic inferring capabilities can improve LORE's compound object search and editing UI.
- The current implementation only allows a single domain ontology for the metadata terms and relationship types to be configured, and requires the user to configure that ontology directly by URL. This is not ideal for our target users, and we intend to incorporate ontology discovery by enabling querying a Metadata Schema Repository.
- We intend enabling users to attach Creative Commons licenses to the compound objects, prior to publishing them, so the permitted types of re-use can be specified.

The types of relationships that may be created between aggregated objects, and metadata that may be attached to those objects, as specified in the AustLit ontology will

need to be refined and extended in specialized ontologies for use by AustLit's specialist research sub-communities. We plan to evaluate these ontologies and the LORE tool more thoroughly through user feedback and case studies taken from the research sub-communities and from within the wider Australian literature studies community.

## 8 Conclusions

In this paper, we describe LORE, a light-weight tool for authoring and publishing OAI-ORE compliant compound objects that use the IFLA FRBR model to represent bibliographic relationships. LORE enables literary scholars to create, exchange and re-use compound objects for the purposes of teaching and research, to describe works, agents and related digital resources, plus their associated metadata and typed relationships. The continued development and evaluation of LORE in the context of the case study will provide an essential component of the cyber-infrastructure requirements of the Australian literary studies community, as well as other literary scholars globally.

## Acknowledgements

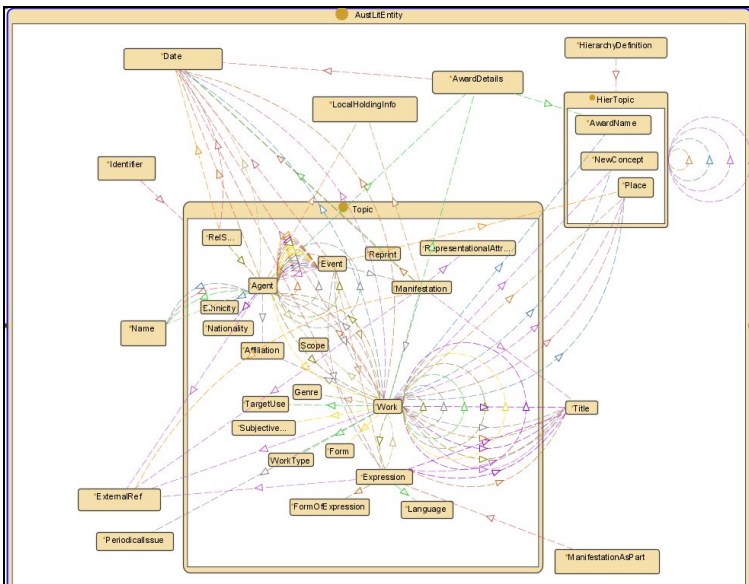
Aus-e-Lit is funded by DEST through the National eResearch Architecture Taskforce (NeAT), part of the National Collaborative Research Infrastructure Strategy (NCRIS).

## References

1. Functional requirements for bibliographic records: Final report (1998), <http://www.ifla.org/VII/s13/frbr/frbr.pdf>
2. Open Archives Initiative - Object Reuse and Exchange (OAI-ORE), <http://www.openarchives.org/ore/>
3. Jeremy, J.C., et al.: Named graphs, provenance and trust. In: 14th International Conference on World Wide Web, pp. 613–622. ACM Press, Chiba (2005)
4. Lagoze, C., Van de Sompel, H.: Compound Information Objects: The OAI-ORE Perspective (2007), <http://www.openarchives.org/ore/documents/CompoundObjects-200705.html>
5. AustLit: The Australian Literature Resource, <http://austlit.edu.au>
6. Doerr, M., Hunter, J., Lagoze, C.: Towards a Core Ontology for Information Integration. *Journal of Digital Information* 4(1) (April 2003)
7. Kilner, K.: The AustLit Gateway and Scholarly Bibliography: A Specialist Implementation of the FRBR. *Cataloguing and Classification Quarterly* 39(3/4) (2005)
8. Fitch, K.: Taking RDF And Topic Maps Seriously. In: AusWeb 2002, The Eighth Australian World Wide Web Conference, Queensland, Australia (2002)
9. Cheung, K., Hunter, J., Lashtabeg, A., Drennan, J.: SCOPE - A Scientific Compound Object Publishing and Editing System. In: 3rd International Digital Curation Conference, Washington DC (2007)
10. CULTOS, <http://www.cultos.org/>
11. Babeu, A.: Building a "FRBR-Inspired" Catalog: The Perseus Digital Library Experience (January 2008), <http://www.perseus.tufts.edu/~ababeu/PerseusFRBRExperiment.pdf>

12. Buchanan, G.: FRBR: Enriching and integrating digital libraries. In: Proc. ACM/IEEE Joint Conference on Digital Libraries, pp. 160–269. ACM, Chapel Hill (2006)
13. Buchanan, G., Hinze, A.: A Generic Alerting Service for Digital Libraries. In: Proc. ACM/IEEE Joint Conference on Digital Libraries, pp. 131–140. ACM, Colorado (2005)
14. FORSITE, <http://foresite.cheshire3.org/>
15. Van Noorden. R.: Microsoft Ventures into Open Access Chemistry. In: Chemistry World (January 2008), <http://www.rsc.org/chemistryworld/News/2008/January/29010803.asp>
16. Cole, T.W.: OAI-ORE experiments at the University of Illinois Library at Urbana-Champaign (April 2008), <http://www.openarchives.org/ore/meetings/Soton/Cole-OAI-ORE-Roll-Out-OR08.pdf>
17. Davis, I., Newman, R.: Expression of Core FRBR Concepts in RDF (2005), <http://vocab.org/frbr/core>
18. A Selection of Bibliographic Ontologies - OAI-ORE Wiki, <http://foresite.cheshire3.org/wiki/CommunitiesHome/ScholarlyWorksHome/BibliographicOntologies>
19. Dublin Core Metadata Element Set, Version 1.1, <http://dublincore.org/documents/dces/>
20. DCMI Metadata Terms, <http://dublincore.org/documents/dcmi-terms/>
21. FOAF Vocabulary Specification 0.91, <http://xmlns.com/foaf/spec/>
22. Lagoze, C., et al.: Fedora: An architecture for complex objects and their relationships. International Journal on Digital Libraries 6(2), 124–138 (2006)

**Appendix: AustLit Ontology**



**Fig. 5.** A subset of the OWL object properties from the AustLit Ontology



# Consolidation of References to Persons in Bibliographic Databases

Nuno Freire, José Borbinha, and Bruno Martins

Instituto Superior Técnico, Technical University of Lisbon,  
Av. Rovisco Pais  
1049-001 Lisboa, Portugal  
{nuno.freire,jlb,bruno.g.martins}@ist.utl.pt

**Abstract.** Entity resolution is the process of determining if, in a specific context, two or more references correspond to the same entity. In this work, we address this problem in the context of references to persons as they are found in bibliographic data, specifically in the case of consolidating multiple datasets. Our solution follows the extraction, transformation and loading (ETL) process, typical in data warehouses. It computes the similarities of the attribute values for the references, and employs a decision tree to decide when the references match. We describe the characteristics of these references within bibliographic datasets, and how we explored those characteristics by developing new similarity metrics to improve the quality of the consolidation process. We evaluated our work by designing an experiment with data from four national libraries. The results show that the proposed similarity metrics contribute significantly to the consolidation process.

**Keywords:** Entity resolution, bibliographic metadata, similarity metrics, machine learning.

## 1 Introduction

Entity resolution is the process of, given a specific context, determining if two or more references correspond to the same entity. In real world datasets, objects are described by a set of attributes that are specific to the context of the business. When these attributes don't comprise a unique identifier, the references may be ambiguous. An object might have multiple different representations, and each representation might match the description of multiple objects (i.e., reference and referent ambiguity). The variations found in the descriptions may have multiple origins, such as misspellings, typing errors, different conventions for abbreviations, naming varying over time, heterogeneous data schemas, etc.

This paper describes the solution that was developed within the DIGMAP<sup>1</sup> project to address the consolidation of references to persons found in bibliographic data. The project's focus is on developing an architecture of services for virtual digital libraries of old maps. The main data sources are national libraries and other relevant online

---

<sup>1</sup> <http://www.digmap.eu>

resources and collections with metadata available by OAI-PMH<sup>2</sup>. DIGMAP deals specifically with automated methods for enriching all the harvested the metadata with structured spatio-temporal information [8, 9, 13], as well as with the consolidation of all references to persons. This enriched metadata is afterwards used to improve the end-user experience while searching and browsing inn this virtual digital library.

The article follows with an introduction to the general concepts and most important related works. Sections 3 and 4 describe respectively the implemented consolidation process and its evaluation. The final section describes future work and presents conclusions.

## 2 Concepts and Related Work

Our work is an application of techniques that are typically used in entity resolution processes, to consolidate the references to persons found in bibliographic databases. This section introduces the entity resolution problem and the characteristics of the bibliographic data that was the focus of our work.

### 2.1 Entity Resolution

Entity resolution is a common problem to many different research communities, although the term used is not always the same. Common designations include record linkage, record matching, merge-purge, data de-duplication, instance identification, database hardening, name matching, reference reconciliation, reference disambiguation, and object consolidation [1, 2, 3].

Different communities have been proposing several techniques, but most frequently we find applications of algorithms from machine learning, artificial intelligence and data mining. Across these communities, several variations of the problem have been formulated. Usually, they differ in the way the resolution process is performed (resolve one entity; resolving several entities; match object representations with a list of disambiguated objects), and on characteristics of the data sets (if it is performed in one or several datasets; the level of detail and overlap between the data schemas of the datasets). Of these variants, we identified three that are related to our work: record linkage, reference disambiguation, and object consolidation.

Record linkage is the problem that often arises when multiple database tables (from different data sources) are merged to create a single database. Alternatively, the problem of object consolidation arises most frequently when the dataset being processed is constructed by merging various data sources into a single unified database. The difference between the two problems is that while record linkage deals with records in a table, object consolidation deals with entities/objects – a semantic concept of a higher level [3]. Methods of record linkage typically assume the availability of many attributes in each object, which can be used effectively in resolving the entity. Object consolidation addresses the cases where very few attributes are available, therefore exploring other sources of evidence for resolving the entities.

---

<sup>2</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html>

As for reference disambiguation, the goal is to match object representations with the list of possible objects which is known to be disambiguated. The requirement of having such a clean list of objects limits the applicability of reference disambiguation. As a rule, each instance of the reference disambiguation problem can be formulated as an instance of the object consolidation problem, while the reverse is not true. That is, the object consolidation problem is more general.

## 2.2 References to Persons in Bibliographic Data

Entity resolution is very depending of the context. The processes need to be adapted to the data that they are being applied to, in order to achieve acceptable results. Our work focused on exploring the structural and semantic richness of the bibliographic data as it exists in the library catalogues. In particular, we looked for characteristics that can be used for helping in the resolution of references to persons.

In bibliographic records, references to persons are found as authors or contributors of works, and some times as the subject of the work. The value of having these references as complete as possible is recognized by cataloguing rules, which indicate that the references should contain, besides the name of the person, the birth and death years [4]. However, these dates do not always exist, as they are not always known of the cataloguers. It is also often the case that these dates are known as approximations of real dates – cataloguing rules comprise conventions for these cases. Another common practice is, when the birth and death dates are not known, to use flourishing dates to indicate the year in which the person is know to have produced its first intellectual work. Although this information regarding is not always fully structured, the common conventions used when encoding the information allow it to be reliably parsed automatically.

Another characteristic of the bibliographic data that we explored in our work was the fact that each bibliographic record contains the year of publication, or the range of years during which the work was published.

## 2.3 Related Work

Two projects have specifically addressed the consolidation of references to persons in bibliographical data. In project LEAF<sup>3</sup> (Linking and Exploring Authority Files), only the information found in authority files<sup>4</sup> was used in the resolution process. Their process was based only on exact matching of the person name (or one of the name variants) and the birth and death years [5].

A second project is VIAF<sup>5</sup> (The Virtual International Authority File), which addresses the consolidation of person references at three major authority databases from the USA, France and Germany. Similarly to our work, this project explores the bibliography associated with the person references to help in the resolution process. However, no wide results and further details of the consolidation process are available, and neither do the authors present an evaluation of the proposed approach.

---

<sup>3</sup> <http://www.crxnet.com/leaf/index.html>

<sup>4</sup> [http://en.wikipedia.org/wiki/Authority\\_control](http://en.wikipedia.org/wiki/Authority_control)

<sup>5</sup> <http://www.oclc.org/research/projects/viaf/>

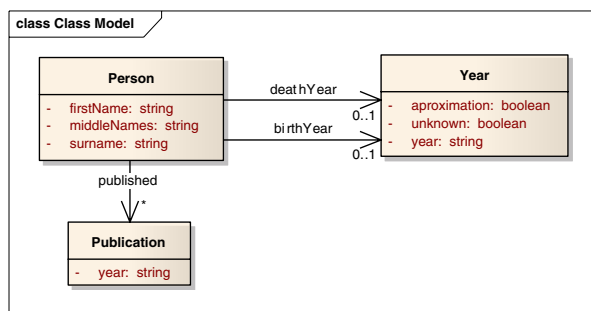
Our work has some overlap with the work carried out in citation indexing systems, that autonomously index the citations found in research papers [2, 6, 7]. These works are based on similar techniques but address bibliographic data with different characteristics of that from libraries.

### 3 The Reference Consolidation System

The person reference consolidation system was built as an ETL (Extraction, Transformation and Loading) process, a typical approach for building data warehouses.

The process starts with the preparation of data for further processing. This step comprises tasks for selecting the relevant data from the bibliographic data sources, parsing the data, and doing initial transformations so that it conforms to a standardized schema. In this way, the structural heterogeneity across the references to persons found in the data sources is reduced.

The standardized data schema is shown in Figure 1. Person names are parsed into three fields that contain the first name, the middle names and the surname. Fields for birth and death years indicate if the years are given as approximations, if they are unknown, or if they refer to a date of flourishing instead of birth/death. The schema also holds the year of publication for all the works where the reference was found.



**Fig. 1.** The standardized data schema of the resolution system

As typical in ETL processes, the final decision to match two references is made by reasoning on the similarity scores obtained by comparing the individual fields. This, however, is not a trivial task due to typographical variations of string data, missing data and other mistakes made when cataloguing.

In this work, we experimented with several string comparison techniques for measuring the similarity of the names. Several comparison techniques were suitable for the kinds of variations of data found in the names of persons: typographical errors, missing names and use of initials, and language variations of writing names. We obtained the best results with the Jaro-Winkler similarity metric [10].

In addition, we developed additional metrics that measure the likelihood of two references referring to the same person, with basis on their dates, name initials and years of publication. These metrics are detailed in the following subsections.

The final decision to match two references is made by reasoning on the comparison scores obtained with the similarity metrics. We used several machine learning techniques for classification, and trained them with a manually classified data set, which is described in Section 4. Due to space limitations, the comparison results are not included in this paper, but we chose to use Alternate Decision Trees [11] since it obtained the best results.

### 3.1 Comparison Metrics for Years

When the birth and death years in the references are unknown or just approximate, cataloguers use conventions to indicate these cases. For example, an approximated decade or century can be registered as “197-“ or “15--“. The fact that years are not always represented as numbers means that a similarity metric for numerical data could not be applied. For this reason we experimented with the Levenshtein distance [12] and developed two metrics that take into consideration the conventions used to represent these years.

#### 3.1.1 Similarity between Two Years

This metric tries to score the likelihood of two references referring to the same year in which a person was born or died. The inputs are two strings, which can contain an exact year or may just specify a decade or century.

The output gives a score between 0 and 1, where 1 means the two references are exactly the same, while 0 means that the references point to years very far apart. Values closer to 1 mean that the references point to years (or time periods) that are very close.

This metric applies different formulas to calculate the similarity, depending on the type of time period (year, decade or century). The formulas are applied as follows (yearA, and yearB are the input):

- When yearA and yearB are precise years (ex: 1975 and 1980):

$$\text{SimilarityOfYears}(\text{yearA}, \text{yearB}) = 1 - \frac{|\text{yearA} - \text{yearB}|}{2} \times 0,01$$

- When one parameter is a precise year, and the other a decade or century (ex: 18-- and 1893):

$$\text{SimilarityOfYears}(\text{yearA}, \text{yearB}) = 1 - \frac{|\text{yearC} - \text{yearApprox}|}{2} \times 0,01$$

where:

- yearC:** is the reference which specifies a year
- yearApprox:** is the middle of the time period in the other reference (ex: for an input of “18--“ the value is “1850”; for “189-“ the value is “1895”)

- When both parameters represent a decade (ex: 184-, 199-):

$$\text{SimilarityOfYears}(\text{yearA}, \text{yearB}) = 0,98 - \frac{|\text{yearA} - \text{yearB}|}{3} \times 0,02$$

- When both parameters represent a century (ex: 18--, 19--):

$$\text{SimilarityOfYears}(\text{yearA}, \text{yearB}) = 0,9 - |\text{yearA} - \text{yearB}|^3 \times 0,02$$

- When parameters represent a decade and a century (ex: 184-, 19--):

$$\text{SimilarityOfYears}(\text{yearA}, \text{yearB}) = 0,98 - |\text{dist}(\text{decade}, \text{century})|^2 \times 0,05$$

where:

- dist(x,y)**: is distance, measured in decades, from the decade to the century.
- decade**: is the reference which specifies a decade
- century**: is the reference which specifies a century

### 3.1.2 Similarity between Life/Flourishing Period References

This metric tries to score the likelihood of two references to a person's life period. The inputs are the years of birth and death, or alternatively the years of flourishing.

The metric can be seen as an extension of the one described in Section 0, also giving a similar output. The formulas used when the periods refer to the life of the person are:

- When the start and end of each period is specified:

$$\text{SimilarityOfPeriod}(\text{periodA}, \text{periodB}) = (\text{SimilarityOfYears}(\text{periodA.birthYear}, \text{periodB.birthYear}) + \text{SimilarityOfYears}(\text{periodA.deathYear}, \text{periodB.deathYear})) / 2$$

- When only the start of the periods is specified:

$$\text{SimilarityOfPeriod}(\text{periodA}, \text{periodB}) = \text{SimilarityOfYears}(\text{periodA.birthYear}, \text{periodB.birthYear})$$

- When only the end of the periods are specified:

$$\text{SimilarityOfPeriod}(\text{periodA}, \text{periodB}) = \text{SimilarityOfYears}(\text{periodA.deathYear}, \text{periodB.deathYear})$$

When both periods refer to the dates of flourishing, the metric computes the similarity score according to the following formula:

$$\text{SimilarityOfFlourishingPeriods}(\text{periodA}, \text{periodB}) = (\text{SimilarityOfPeriod}(\text{periodA}, \text{periodB}) + 1) / 2$$

## 3.2 Similarity of Name Initials

This metric tries to score the likelihood of two references by comparing the initials of the full name. It gives a score between 0 and 1, where 1 means the initials are exactly the same, while 0 means that the initials are very different.

The metric considers the number of initials that are omitted (omt) from one of them and the number of initials that are different (dif). For example "J.W.S." and "J.T.S."

has one difference in the middle initial and no omissions, while “J.S.” and “J.W.S.” have one omission and no differences. The final similarity score is given according to the following formula:

$$\text{SimilarityOfInitials}(\text{initialsA}, \text{initialsB}) = 1 - (\text{diff} \times 0,3 + \text{omt} \times 0,03)$$

### 3.3 Similarity of the Bibliography

This metric tries to score the likelihood of two references by comparing the time span of their publications. A positive score means that the bibliographies are likely to be from the same person and the higher the score the higher the likelihood. A negative score means that the bibliographies are unlikely to belong to the same person and the lower the score the higher the unlikelihood.

The algorithm of this metric uses the following intermediate values:

<p><b>totalTimeSpan:</b> the total time span of both bibliographies  <b>overlapPct:</b> the percentage of the time span overlap between the two bibliographies  <b>gap:</b> the size of the gap between the two bibliographies (in years)  <b>pubsOnSameYearPct:</b> the percentage of publications on the same year</p>
--

With the above values the similarity score is given as follows:

```

if (overlapPct == 0){
  if (gap>50 or totalTimeSpan>120 or (totalTimeSpan>80 and
    gap>30) ){
    score= - (((min(200, totalTimeSpan)-80)/120)x0,5 +
      ((min(80, gap)-40)/40)x0,5 )
  } else {
    score=(1-totalTimeSpan/100)x0,5+(1-gap/50)x0,5
  }
}else{
  if (overlapPct<50 and totalTimeSpan>80){
    score= - (((min(200, totalTimeSpan)-80)/120)x0,5 +
      ((min(80, gap)-40)/40)x0,5 )
  } else if (totalTimeSpan<70 or overlapPct>55 or (totalTime
    Span<90 and overlapPct>50) ){
    score=(max(100, totalTimeSpan)/100)x0,2+
      min(1, overlapPct/100 +
        pubsOnSameYearPct/100/3)x0,5
  }
}

```

## 4 Evaluation

The proposals were evaluated with data from four European national libraries, partners in DIGMAP. All libraries provided the bibliographical database of their cartographic materials, and the corresponding authority file. Together, the four databases totalled 17.106 bibliographic records, where we found 5.734 references to persons.

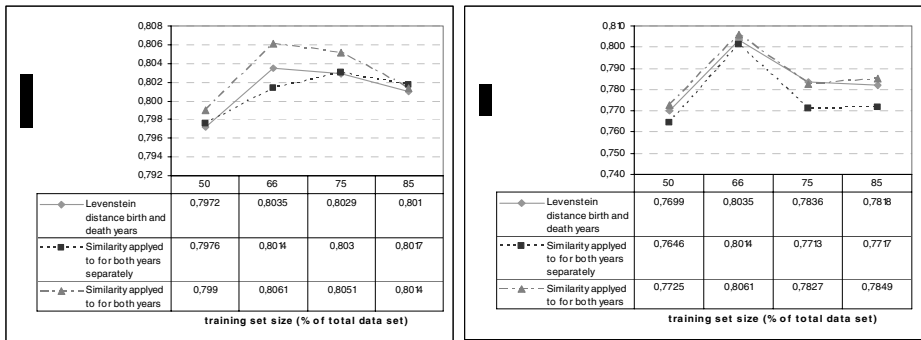
From this collection, a smaller set of records was selected. We calculated the similarity of the surnames and first names of all persons, and selected those that had a minimal similarity on the first name or surname with at least one other reference. This selection process created a sub collection of 1.667 references to persons. These references were manually resolved by a professional librarian with access to the original bibliographic records, so all the ambiguities could be solved.

The smaller collection served both as training set for the decision tree, and also to evaluate the precision and recall of the full set of metrics.

For each test, we present below the obtained precision and recall. The values shown are the result of executing the consolidation process 100 times and averaging the precision and recall values obtained. For each execution, the decision tree was trained with a random part of the training set (50%, 66%, 75% and 85%), and the precision and recall values were measured by executing the trained decision tree on the rest of the training set.

The precision and recall performance of some metrics was measured by comparing the results obtained by a base comparison and those obtained by adding the metric to the base comparison. This base comparison consisted on the following fields: firstName, middleNames, surname, birthYear, deathYear.

Our first test aimed at choosing the best way to use the years of birth and death in the resolution process. We compared three techniques: Levenstein distance of birth and death years; similarity between two years applied to birth and death years (described in Section 0); and similarity between life/flourishing period references (described in Section 3.1.2). The results are given in Figure 2, showing that the similarity between life/flourishing periods gave higher precision and recall, resulting in a F-measure of  $\sim 0,791$  ( $\sim 0,002$  above Levenstein distance).



**Fig. 2.** Precision and recall of similarity metrics for death and birth years

The similarity metric for comparing the initials of the full names was evaluated by comparing the results of the base comparison with the results obtained by also using the initials metric. The results are given in Figure 3, and they show a gain in precision of  $\sim 0,007$  and of  $\sim 0,025$  in recall. The F-measure was  $\sim 0,807$  which was an increase of  $\sim 0,018$  over the base comparison results.



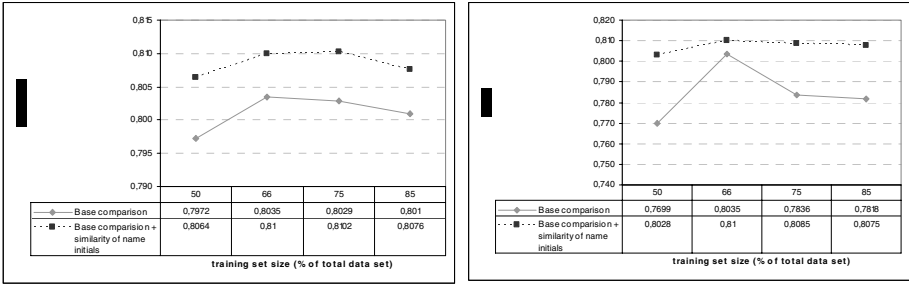


Fig. 3. Precision and recall of the full name initials similarity metric

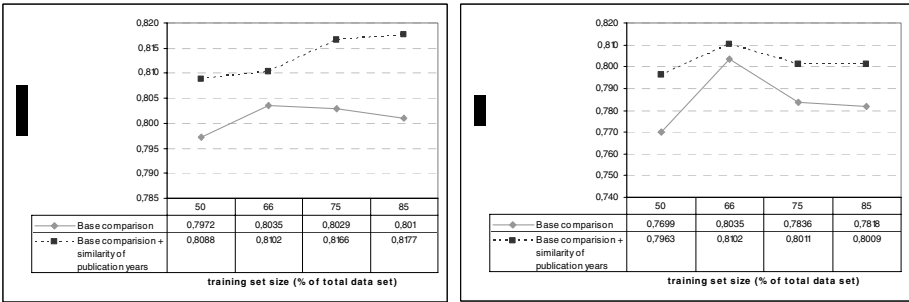


Fig. 4. Precision and recall of the publication years' similarity metric

The similarity metric for comparing the bibliographies was evaluated by comparing the results of the base comparison with and without this metric. The results are shown in Figure 4, and they show a gain in precision of  $\sim 0,012$  and  $\sim 0,017$  in recall. The F-measure was  $\sim 0,805$  which was an increase of  $\sim 0,017$  over the base comparison results.

Altogether the best results were obtained with the combination of the metrics that gave the best results. The final comparison used in the DIGMAP consolidation process was measured against the base comparison with the results shown in Figure 5. They show an average gain in precision of  $\sim 0,032$  and  $\sim 0,049$  in recall. The F-measure was  $\sim 0,833$  representing an increase of  $\sim 0,044$  over the base comparison results.

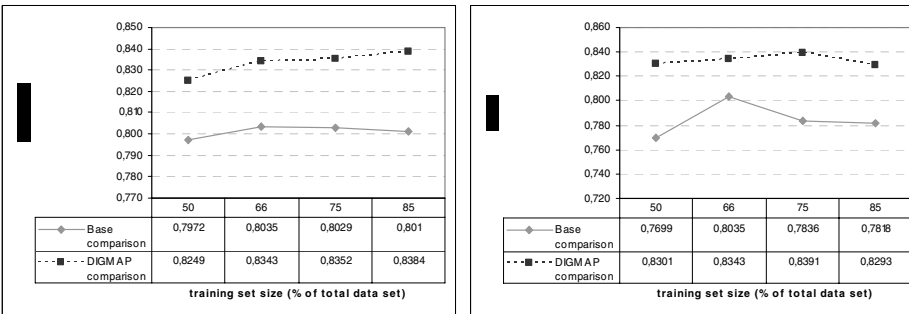


Fig. 5. Precision and recall of the DIGMAP consolidation system

## 5 Conclusions and Future Work

In this paper we argue that the structural and semantic richness of the bibliographic data, as it exists in the library catalogues, can be used with success in automatic entity resolution processes. We have demonstrated this claim using the years of birth and death (considering the conventions used in libraries for expressing this information), the initials of the full name, and the flourishing years. Currently ongoing work explores how to further use the context in which the references to persons are found, in order to improve the consolidation system. In particular, we will explore co-authors, subjects, places of publication, and titles similarities.

Our work is a step to achieve a more ambitious objective: to design an object consolidation solution to transform traditional bibliographic databases (especially those expressed according to the MARC family of formats) in new schemas aligned with the Functional Requirements for Bibliographic Records (FRBR) conceptual model<sup>6</sup>.

## References

1. Elmagarmid, A.K., Ipeirotis, P.G., Verykios, V.S.: Duplicate Record Detection: A Survey. *IEEE Transactions on knowledge and data engineering* 19(1), 1–16 (2007)
2. Dong, X., Halevy, A., Madhavan, J.: Reference reconciliation in complex information spaces. In: *Proceedings of the 2005 ACM SIGMOD international Conference on Management of Data. SIGMOD 2005*, pp. 85–96. ACM, New York (2005)
3. Chen, Z., Kalashnikov, D.V., Mehrotra, S.: Exploiting relationships for object consolidation. In: *IQIS 2005*, pp. 47–58. ACM, New York (2005)
4. ALA, CLA, CILIP. *Anglo-American Cataloguing Rules: 2002 Revision* (2002)
5. Kaiser, M., Lieder, H.J., Majcen, K., Vallant, H.: *New Ways of Sharing and Using Authority Information. D-Lib Magazine* 9(11) (2003), <http://www.dlib.org/dlib/november03/lieder/11lieder.html>
6. Lawrence, S., Giles, C.L., Bollacker, K.D.: Autonomous Citation Matching. In: *Proceedings of the Third International Conference on Autonomous Agents*. ACM, New York (1999)
7. Pasula, H., Marthi, B., Milch, B., Russell, S., Shpitser, I.: Identity Uncertainty and Citation Matching. In: *Advances in Neural Information Processing* (2002)
8. Martins, B., Manguinhas, H., Borbinha, J.: Extracting and Exploring Semantic Geographical Information from Textual Resources. In: *Proceedings of the Second IEEE International Conference on Semantic Computing (ICSC)* (2008)
9. Manguinhas, H., Martins, B., Borbinha, J., Siabato, W.: The DIGMAP Geo-Temporal Web Gazetteer Service. In: *Third ICA Workshop on Digital Approaches to Cartographic Heritage* (2008)
10. Jaro, M.A.: Advances in record linking methodology as applied to the 1985 census of Tampa Florida. *Journal of the American Statistical Society* 64, 1183–1210 (1989)
11. Freund, Y., Mason, L.: The Alternating Decision Tree Algorithm. In: *Proceedings of the 16th International Conference on Machine Learning*, pp. 124–133 (1999)
12. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 707–710 (1966)
13. Martins, B., Freire, N., Borbinha, J.: Using XML Technologies for Complex Data Transformations in Geo-referenced Digital Libraries. In: *International Conference on Asia-Pacific Digital Libraries 2008* (2008)

---

<sup>6</sup> <http://www.ifla.org/VII/s13/frbr/frbr.pdf>

# On Visualizing Heterogeneous Semantic Networks from Multiple Data Sources

Maureen<sup>1</sup>, Aixin Sun<sup>1</sup>, Ee-Peng Lim<sup>2</sup>, Anwitaman Datta<sup>1</sup>, and Kuiyu Chang<sup>1</sup>

<sup>1</sup> School of Computer Engineering, Nanyang Technological University, Singapore  
{maureen,axsun,anwitaman,askychang}@ntu.edu.sg

<sup>2</sup> School of Information Systems, Singapore Management University, Singapore  
eplim@smu.edu.sg

**Abstract.** In this paper, we focus on the visualization of heterogeneous semantic networks obtained from multiple data sources. A semantic network comprising a set of entities and relationships is often used for representing knowledge derived from textual data or database records. Although the semantic networks created for the same domain at different data sources may cover a similar set of entities, these networks could also be very different because of naming conventions, coverage, view points, and other reasons. Since digital libraries often contain data from multiple sources, we propose a visualization tool to integrate and analyze the differences among multiple social networks. Through a case study on two terrorism-related semantic networks derived from Wikipedia and Terrorism Knowledge Base (TKB) respectively, the effectiveness of our proposed visualization tool is demonstrated.

## 1 Introduction

### 1.1 Motivation

A semantic network refers to a set of concepts or entities, possibly of different types, connected by relationships. In the digital library context, semantic networks have always been a useful paradigm for representing knowledge found in text and database records which in turn helps users to more effectively and quickly search and navigate information. Some often cited examples of semantic networks in digital libraries include author co-citation networks [2], keyword co-occurrence networks [10], etc. In this paper, we focus on social networks as kinds of semantic networks found in text collections and databases. For large social networks, visualization tools will be required to assist users in viewing, searching and analyzing entities and relationships in the networks as well as locating the documents or database records containing the sub-networks users are interested in. In this paper, we therefore describe our proposed interactive tool that supports social network visualization and data access based on network navigation.

As digital libraries often include data taken from different sources, the social networks obtained from one source may look very different from other sources

even when they share some common entities and relationships. This heterogeneity is often caused by different *naming conventions*, *attribute format*, *coverage*, and *view points* adopted at different sources. For example, the (*first name, last name*) person name format may be used in source *A*, while source *B* uses the (*last name, first name*) name format. Person entities from *A* may have a phone attribute but person entities from *B* may not have it. As the social networks can be contributed by different sets of users, they may not cover the same set of entities and relationships. Furthermore, the users who are responsible for creating content at different sources may assign different type labels or attribute values to the same entity or relationship due to varying view points. Given these heterogeneity issues, a visualization tool is necessary to integrate multiple social networks together via entity (and relationship) resolution as well as attribute merging and to keep the unresolved and resolved entities distinctive in the user interface.

With the recent advances in social computing and the wide availability of social software (e.g., wikis and blogs), it is increasingly easy to find semantic networks or even social networks of specific domains defined over Web content or publicly accessible databases. For example, Wikipedia, the largest encyclopedia on the Web collaboratively created by millions of users, provides rich article content about interlinked entities, thereby providing additional semantics about their relationships (e.g., topic category labels of articles).

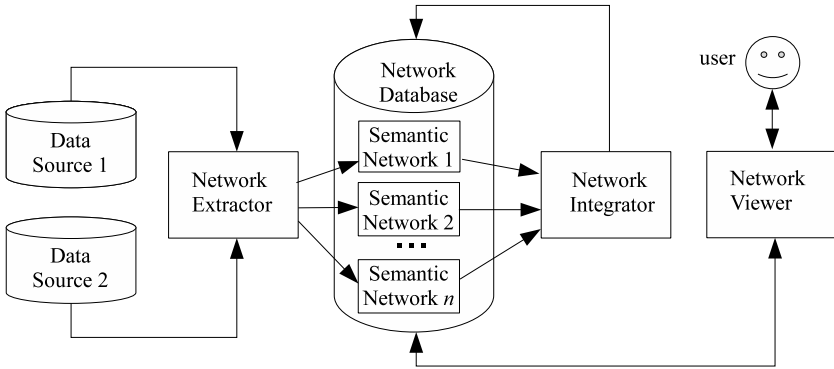
## 1.2 Objective and Contribution

The main objective of this research is to develop an interactive tool for visualizing semantic networks from multiple data sources. Other than viewing and navigating network entities and relationships, the visualization tool will assist users in exploring the underlying data (documents or database records) from which the networks are obtained, and comparing the entities, relationships, and network connectivities between semantic networks.

Figure 1 depicts the system architecture of the visualization tool. It consists of a **network extractor** that extracts semantic networks from data sources. The extracted network information is stored in the **network database**. The **network integrator** is responsible for taking two or more heterogeneous semantic networks and integrating their entities and relationships. These integrated semantic networks are then stored in the network database. The **network viewer** provides an interactive interface for users to retrieve semantic networks, navigate, and access semantic networks and their underlying text or database records.

In this paper, we describe our visualization tool built based on the above system architecture and summarize the research contributions as follows:

- We have defined a database schema for modeling semantic networks and the entity matchings among different semantic networks. This database schema is designed to be generic enough to handle as many different types of semantic networks as possible.
- We have developed a working prototype visualization tool using TouchGraph API [14], a graphical user interface programming package for graph



**Fig. 1.** System Architecture for Visualization Tool

visualization. We use color and shape to distinguish the different data sources and entity types.

- We have applied our tool to a case study involving two terrorism related social networks from (a) Wikipedia and (b) Terrorism Knowledge Base (TKB). TKB was provided on the Web and maintained by the Memorial Institute for the Prevention of Terrorism (MIPT). In this case study, the social network derived from Wikipedia represents the common web user knowledge in the terrorism domain, in which users acquire information from news articles and other online sources (some of them are mentioned as references in Wikipedia articles). TKB on the other hand is an expert maintained knowledge base containing information about terrorist groups and members. This case study leads to some interesting observations of the integrated social networks, which help users identify discrepancies between TKB and Wikipedia social networks.

### 1.3 Paper Outline

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 discusses the modeling and integrating of semantic networks. The visualization interface is given in Section 4 followed by a case study in Section 5. Finally, Section 6 concludes this paper.

## 2 Related Work

There have been several work on the visualization of different kinds of network graphs. For example, Vizster provides visualization functions for exploration, search, and analysis of online social networks [5]. A survey of visualization techniques for ontology networks is reported in [8]. Gene network visualization is addressed in [3]. All of the aforementioned network visualization tools cannot handle multiple networks as they are confined to display only a single network.

In our research, we propose the idea of visualizing multiple semantic networks by integrating them together. The integrated network allows us to better understand the global network connectivities and compare any variations across different networks.

With reference to the survey by Katifori *et al* in [8], our visualization tool adopts both *context plus focus* and *distortion* techniques. Our semantic network graphs when displayed in the network viewer require a combination of context and focus. That is, each graph has a node serving the focus (central node) surrounded by other nodes with edges connected to it. Some of the other visualization tools that use this technique include TGVizTab [1], MoireGraphs [7], OntoRama [4], OntoViz [11], and OZONE [13].

A closely related work to our visualization tool is Protege [12], a framework for ontology creation, editing, and visualization. Under the Protege framework, several visualization tools have been developed including the above-mentioned tool such as TGVizTab and OntoViz. Our tool is similar in the graph visualization aspect but differs in usage and data storage aspects. In particular, we aim to use our visualization tool for multi-modal social networks which are stored in a relational database.

### 3 Modeling and Integration of Semantic Networks

#### 3.1 Semantic Network Representation

A semantic network comprises typed entities and relationships. Our network data model supports a configurable set of *entity types* and *relationship types*. Each entity type defines a set of attributes shared by all entities belonging to the type; each entity type may have one or more *relationship type* with other entity types. For example, in our case study, the semantic network created in the terrorism domain involves two entity types: **Terrorist Group** and **Terrorist**. **Terrorist Group** entity type has attributes: name, location, and date. **Terrorist Group** may be related to itself by an **Associated With** relationship type, and to **Terrorist** entity type by a **Has Leader** relationship type. At the instance level, the **Terrorist Group** entity *Al-Qaeda* has an **Associated With** relationship with *Yemen Islamic Jihad* (an entity of **Terrorist Group**) and a **Has Leader** relationship with *Osama bin Laden*, a **Terrorist** entity. In our visualization, like many others, each entity is depicted as a *node*, and each relationship is depicted as a directed *edge* connecting the related pair of nodes.

To store the semantic networks from different data sources in our network database, we define meta-data to describe the data sources and their mappings. Each data source is an instance of the **Source** entity type, identified by its *SourceID*. Each data source is also given a *SourceName* and it consists of one or more *EntityType* instances. *EntityType* instances can be related to other *EntityType* instances through some relationships. Other than *EntityTypeID* and *EntityType-Name*, each *EntityType* instance may have attributes defined through *Attribute* instances. Specifically, each *Attribute* instance is given a *AttributeName*, *Order*, and *IsMultivalued* flag. The *Order* value indicates the relative position at

which the attribute will be subsequently displayed by the network viewer. The *IsMultivalued* flag is a boolean value indicating whether the attribute allows set values. An *Attribute* instance is also assigned a *Domain* instance, which defines the *DataType*, *MinValue* and *MaxValue* of the the attribute value. Our default *Domain* instances include integer, character strings, date, and float numbers, which are supported by most existing database systems. A user defined *Domain* instance will have its enumerable domain values given by the multi-valued attribute *UserDefined*. As in many database systems, by separating domain information from the attribute definition, the same *Domain* instance can be shared among different *Attribute* instances. The *EquivalentTo* relationship is used to store matching entities discussed in the following subsections.

### 3.2 Semantic Network Integration

To integrate different heterogeneous semantic networks, the mapping of entities and relationships between networks need to be addressed. There are two kinds of entity matching, namely *inter-source* and *intra-source* entity matchings. The former refers to finding matching entities from different data sources, while the latter detects matching entities from the same data source.

**Inter-Source Entity Matching.** In this kind of entity matching, we aim to find common real-world entity with different names from different data sources, i.e., *synonyms*. When the difference between two synonyms is minor, they can be detected by a simple name similarity test. An example of this is a terrorist group known as *Harakat-ul-jihad-i-islami* and *Harakat-ul-jihad-ul-islami* defined in TKB and Wikipedia, respectively. We measure the similarity between them using edit distance, i.e., the minimum number of operations (character insertion, deletion, or substitution) required to transform one name into another. When the edit distance between two entity names is smaller than a specified threshold (30% of the shortest name length in our case study), we flag entities as candidate synonyms for subsequent human verification. Fuzzy search provided by Lucene is utilized in our implementation to automate the above matching process. However, for synonyms that are very different, name similarity tests fail due to their low similarity score. For example, a terrorist group known as *Black Widow* in TKB is known as *Shahdika* in Wikipedia. One can only tell they are synonyms by reading the content of the Wikipedia article and the corresponding TKB database record, as well as referring to external knowledge. For such kind of synonyms, manual matching is adopted in our current implementation.

**Intra-Source Entity Matching.** Each real world entity is supposed to be represented by a unique entry in a data source. However, this assumption does not always hold as the same entity may be labeled differently in a single data source. Some data sources may store these different names of the same entity and their mappings within their databases or markup articles. We propose an intra-source entity matching scheme that derive matching of entity names from the same source by referring to matching entity names in other sources. For example,

**Table 1.** Entities in TKB matching ASALA in Wikipedia

---

1. Armenian Secret Army for the Liberation of Armenia (ASALA)
2. Third of October Group
3. Ninth of June Organization
4. New Armenian Resistance (NAR)
5. September-France

---

in Table 1, all the five groups in TKB match a single Wikipedia article called *Armenian Secret Army for the Liberation of Armenia* (ASALA). The reason is that TKB lists the groups: (*Third of October Group*, *Ninth of June Organization*, *New Armenian Resistance*, and *September France*) as possible sub-groups or ad-hoc groups of the more established group named ASALA. These mappings of different names to the same entity can be applied to find matching entity names in Wikipedia.

## 4 Network Visualization Interface

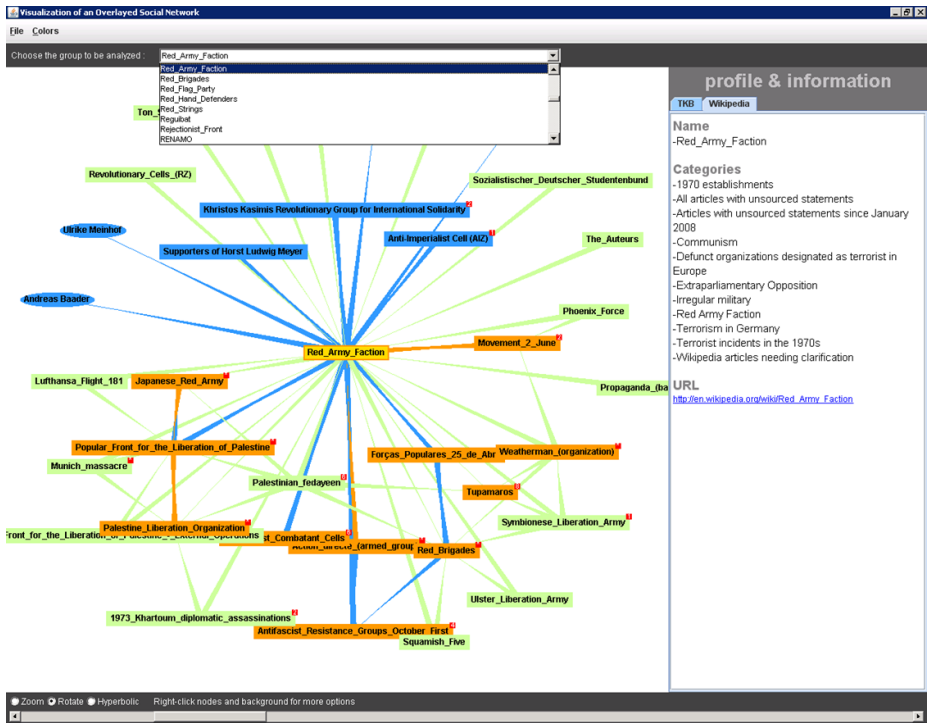
As shown in the system architecture, our network viewer provides visualization functions for semantic networks stored in a network database. The main visualization functions include: (a) loading and displaying multiple semantic networks; (b) browsing the attributes of nodes; and (c) constructing a subnetwork as part of the data analysis process. The visualization interface has been implemented using TouchGraph [14], an open-source library in Java for creating and displaying networks through interactive user interface.

### 4.1 Interface Design

The network viewer user interface is shown in Figure 2. A drop-down-list at the top section provides users a list of entities to be selected for analysis. Once an entity is selected, its entity profile and attribute information will be displayed on the right section. Since an entity may appear in multiple data sources, its information is obtained from all data sources and is shown in the respective source's tabbed pane. The *balloon graph view* [6] is chosen in Touchgraph to display a semantic network containing the selected entity at the center of the network.

We use *color* and *shape* to distinguish the data source(s) and the entity type, respectively. In the example given, exclusive information from TKB are shown in blue. Green is assigned to exclusive Wikipedia sources, and orange is assigned to overlapping sources (i.e., those that appear in both TKB and Wikipedia). Note that the color scheme can be configured by users. Moreover, all **Terrorist** entities are represented by ellipses and **Terrorist Group** entities are depicted as rectangles. For instance, as shown in Figure 2, an entity named *Andreas Baader* belongs to the **Terrorist** entity type from TKB. The corresponding node is a blue ellipse. Another entity named *Red Army Faction* belongs to **Terrorist Group** and





**Fig. 2.** Look and feel of our visualization tool

can be found in both TKB and Wikipedia. The corresponding node is an orange rectangle. Tools including zooming, rotating, etc are provided at the bottom of the interface.

**4.2 Database Configuration**

Other than visualizing semantic networks, our visualization tool also supports configuration of the data sources, entity types and their attributes to minimize user effort in maintaining the databases. The wizard dialog allows user to (a) add new data source to the network database, and (b) create new entity types. Screen captures are not shown due to page limits. All of the above operations affect the network database content. As soon as a user completes configuration using this wizard, the necessary tables in the network database will be automatically built and/or updated. Users may then import, view, insert, edit, remove, and export network data in the network database. To allow semantic network data to be portable across applications, we adopted eXtensible Markup Language (XML) for data import and export operations [9]. These functions are provided mainly for those users who are less familiar with database systems.

## 5 Case Study

In this section, we demonstrate the usefulness of our visualization graph for social network analysis through a case study. Following our earlier discussion, our case study involves two semantic networks both consisting of terrorism related entities and relationships from TKB and Wikipedia respectively. The semantic network derived from Wikipedia represents the common web user knowledge in the terrorism domain while the one from TKB represents the expert understanding of the domain. Here, we would like to find out how the knowledge of experts differ from that of the public.

For Terrorist Group, 858 entities and 1179 relationships were extracted from TKB; 998 entities and 2302 relationships from Wikipedia. Among them 305 entities and 259 relationships appear in both sources. For Terrorist entity type, 1463 entities have been extracted from TKB together with 1374 relationships between Terrorist and Terrorist Group. For Wikipedia, since there is no particular category label for extracting terrorists, extracting terrorists from Wikipedia remains challenging. In this case study, we hence mainly focus on the differences among terrorist groups. As shown in Figure 3, the selected terrorist group *Tanzim* is shown at the center of the network. Those nodes that only appear in one data source are clearly indicated by their colors. Recall that all information derived from TKB are shown in blue and that from Wikipedia in green; and orange is used for information derived from both sources. It is therefore interesting to observe differences in relationships among entities that appear in both data sources. For example, according to TKB, *Tanzim* is related to *Badr Organization*, *Al-Aqsa Martyr's Brigades*, *Fatah*, and *Popular Resistance Committee*. On the other hand, according to Wikipedia *Tanzim* is related to all these groups except *Badr Organization*. Furthermore, there are no relationships between *Badr*

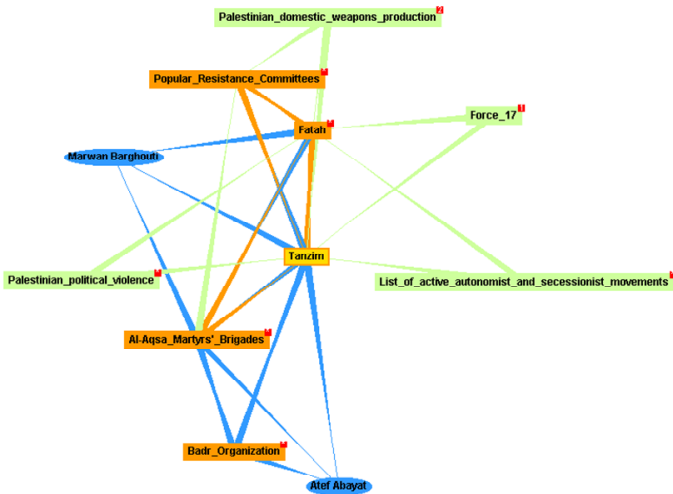


Fig. 3. Network of Tanzim based on information from TKB and Wikipedia

*Organization and Al-Aqsa Martyr's Brigades* in Wikipedia whereas in TKB such a relationship exists. Also, we have observed that according to Wikipedia, there is a relationship between *Popular Resistance Committee* and *Al-Aqsa Martyr's Brigades* whereas it is not mentioned in TKB. This specific example illustrates that in the homeland security domain, the knowledge of the public can be quite different from that of domain experts. Understanding how this can happen is another interesting topic that can be further investigated.

## 6 Conclusion

In this paper, we proposed a tool for visualizing heterogeneous semantic networks obtained from multiple data sources. The modeling of metadata for the entities and relationships contained in semantic networks and their mappings are described. In order to have an easy way of analyzing the integrated network and comparing their differences as well, we have delivered a visualization interface using TouchGraph API. A case study on two semantic networks obtained from TKB and Wikipedia is reported to illustrate the differences in the understanding of terrorism related information from the public and the expert domains.

The future work for this visualization tool is to embed the system with functionality to query the graph using faceted search technique [15]. Faceted search is basically a method for refining search results by categories. For example, given a library of terrorism from our database, faceted search will enable a user to pare down the search results using attributes such as location of incident, date of event, terrorist's nationality and so on. Thus, this method will allow the user to browse and navigate the information to find what he/she really wants.

We will also continue working on ways to minimize manual effort for entity matching. Some of the possible enhancements like the ability to zoom in/out of complex networks with a fish-eye view, retrace steps during browsing using back/forward buttons, load/save the current network view for selected node, are among the list of candidate features to be incorporated into our visualization tool.

## Acknowledgement

This work was supported by A\*STAR Public Sector R&D, Singapore, Project Number 062 101 0031.

## References

1. Alani, H.: Tgviztab: An ontology visualization extension for protege. In: Proc. of Knowledge Capture (K-Cap 2003), Workshop on Visualization Information in Knowledge Engineering, Sanibel Island, Florida (2003)
2. Chen, C.: Visualizing semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 401–420 (1999)

3. Ciccicarese, P., Mazzocchi, S., Ferrazzia, F., Sacchia, L.: Genius: a new tool for gene networks visualization. In: Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP) Proceedings, pp. 107–111 (2004)
4. Eklund, P.W., Roberts, N., Green, S.: Ontorama: Browsing an RDF ontology using a hyperbolic-like browser. In: Proc. of the First International Symposium on CyberWorlds (CW 2002) Theory and Practices, Seattle, Washington, pp. 405–411 (2002)
5. Heer, J., Boyd, D.: Vizster: Visualizing online social networks. In: Proc. IEEE Symposium on Information Visualization, Minneapolis, MN, USA (2005)
6. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics* 6(1), 24–43 (2000)
7. Jankun, K.T.J., Kwan, L.M.: Moiregraphs: Radial focus plus context visualization and interaction for graphs with visual nodes. In: Proc. of IEEE Symposium on Information Visualization, Seattle, Washington, pp. 20–21 (2003)
8. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods—a survey. *ACM Comput. Surv.* 39(4), 10 (2007)
9. Lear, A.C.: XML seen as integral to application integration. *IT Professional* 1(5), 12–16 (1999)
10. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools* 13(1), 157–169 (2004)
11. OntoViz, <http://protegewiki.stanford.edu/index.php/OntoViz>
12. Storey, M., Lintern, R., Ernst, N.: Visualization and protege. In: 7th International Protege Conference (2004)
13. Suh, B., Bederson, B.B.: Ozone: A zoomable interface for navigating ontology information. In: Proc. of Advanced Visual Interfaces. ACM, New York (2002)
14. Touchgraph, <http://www.touchgraph.com/>
15. Yee, K.-P., Swearingen, K., Li, K., Hearst, M.: Faceted metadata for image search and browsing. In: CHI 2003: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 401–408. ACM, New York (2003)

# Using Mutual Information Technique in Cross-Language Information Retrieval

Syandra Sari<sup>1</sup> and Mirna Adriani<sup>2</sup>

<sup>1</sup> Faculty of Computer Science, University of Indonesia affiliation with  
Faculty of Industrial Technology, Trisakti University  
Jl Kyai Tapa No.1, Jakarta 11440  
syandra\_sari@trisakti.ac.id

<sup>2</sup> Faculty of Computer Science, University of Indonesia  
Depok Campus, Depok 16424  
mirna@cs.ui.ac.id

**Abstract.** This paper describes the Indonesian-English cross language information system, where we investigated the problem of cross language document retrieval for Indonesian-English. Our work based on Indonesian-English parallel corpus and applied mutual information technique. Our parallel corpus was built using a machine translation tool. Indonesian queries were translated into English by means of bilingual word list that created using mutual information technique. Experimental results demonstrated that translating Indonesian queries into English using this bilingual word list achieved 41.9% of the monolingual queries.

**Keywords:** Cross-language information retrieval, mutual information, parallel corpus.

## 1 Introduction

The World Wide Web and the number of machine readable texts accessible via CD-ROMs have been growing rapidly. Now documents are usually provided in a limited number of languages. Hence the task of information retrieval needs to be expanded, so that users can retrieve textual information across languages. One application, commonly termed “cross language information retrieval (CLIR)”, is the retrieval task where the user presents queries in one language to retrieve documents in another language. CLIR needs to transform queries and documents into a common representation, so that monolingual IR techniques can be applied. There are two main approaches in CLIR. The first approach translates queries into the document language, while the second approach translates documents into the query language. McCarley applied a statistical MT method to both query and document translation methods for English-French CLIR and vice versa. He showed that the relative superiority between query and document translation methods varied depending on the source and target language pair [11]. Either query translation or document translation, three significant approaches that are used for translation in CLIR process are machine translation[1], dictionary [8] and parallel or comparable corpus [12].

Machine translation, bilingual dictionary and parallel or comparable corpus are not always available for each language pair. For some languages such as English, Spanish, French, we can easily find good quality machine translation, bilingual dictionary and parallel or comparable corpus for them. But for some other languages, example Indonesian language, machine translation, bilingual dictionary and parallel or comparable corpus are still hard to find and if there are some resources or tools the quality are still poor and need to be developed.

While Indonesian is spoken by more than 250 million people, our work is aimed at evaluating language resources and tools available for Indonesian-English pair. Main technique that we describe in this paper is mutual information based on Indonesian-English parallel corpus. Corpus-based CLIR methods are based on bilingual text collections, from which translation knowledge is derived using various methods. Parallel corpus consists of document pairs that are exact translations of each other. Comparable corpus is made of document pairs that are not translations of each other, but share similar topics [16]. A successful works can be found in past CLIR literature in using parallel corpus for translation process. Some methods for CLIR based on corpus that have been reported are relevance model techniques by Victor Lavrenko [10], combination approach for multilingual information retrieval by Martin Braschler [3] and bilingual corpora for translanguing information retrieval by Yiming Yang et al. [17].

Usually the performance of the CLIR result is not as good as that of monolingual retrieval. In order to improve the CLIR performance, query expansion can be applied to the CLIR result. It has been widely known that query expansion techniques can help increase the retrieval performance [2].

Next section gives more explanation about three above previous research in area of CLIR using parallel or comparable corpus and research in Indonesian-English CLIR. Section 3 then describes concept of mutual Information and query expansion that we elaborated in our work. Section 4 presents our experiment and section 5 evaluates the result. Finally section 6 states our conclusions.

## 2 Previous Research of Parallel Corpus

In this section we briefly present some of the research related to CLIR using parallel and comparable corpus and also research CLIR for Indonesian language.

Yiming Yang et al. [17] used parallel corpus for English-Chinese and English-Japanese task. They created a corpus-based term-equivalence matrix extracted automatically from bilingual corpora. The result performed 101% of monolingual IR performance.

Lavrenko et al [10] applied language model for cross language information retrieval using parallel corpus. This model does not rely on a word-by-word translation of the query, instead they attempt to construct an accurate relevance model in the target language, and use that model to rank the documents in the collection. The result showed that the Chinese-English CLIR achieve 95% of monolingual performance.

Martin Braschler presented combination approach for multilingual information system in [3]. He showed that the use of combination several "simple" approaches could substantially increase retrieval effectiveness. One of approaches used for query

translation was similarity thesauri. The similarity thesaurus is an automatically calculated data structure, which is built on training corpus. It links terms to lists of their statistically most similar counterparts [14]. Terms in the source language are then linked to the most similar terms in the target language [15]. Braschler applied this technique for some European languages pair such as English, German, Spanish, Italian and French.

In our earlier study [7], we worked on Indonesian-English CLIR used several language resources in translating the queries and documents. The Indonesian query was translated into English using machine translation, bilingual dictionary and parallel corpus. The parallel corpus was built by translating English documents into Indonesian using bilingual dictionary. The results showed that Indonesian-English CLIR using machine translation achieved 84.82% of monolingual performance, using bilingual dictionary achieved 51.98% of monolingual performance. Although the result from machine translation and bilingual dictionary were good enough, we also did Indonesian-English CLIR using parallel corpus because we thought that parallel corpus could be one of the alternative resources for building cross language information system for Indonesian language. In this experiment the result achieved using parallel corpus was only 8% of monolingual performance.

### 3 Measuring Mutual Information and Query Expansion

#### 3.1 Mutual Information

The use of mutual information in monolingual information retrieval for finding word association has been used by Kenneth Ward Church [4]. Myung-Gil used mutual information in cross-language IR for finding the best translation word from bilingual dictionary [9].

Our research in Indonesia-English cross-language IR based on parallel corpus also used mutual information. We used mutual information for measuring the association degree between Indonesian and English word pair. The word pair that has highest mutual information value is considered to be the best word pair.

Mutual information of two points (words),  $x$  and  $y$ , is defined to be [5]:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \quad (1)$$

Where,  $P(x)$  is the occurrence probability of word  $x$ ;  $P(y)$  is the occurrence probability of word  $y$ ,  $P(x,y)$  is the probability that the words  $x$  and  $y$  occur together.  $I(x,y)$  is their mutual information value.

According to equation (1), the mutual information value is computed based on word co-occurrence statistics [4,9] and, hence, the mutual information can be defined as follows:

$$I(x, y) = \log_2 \frac{N * f(x, y)}{f(x)f(y)} \quad (2)$$

Where  $f(x)$  is the number of documents containing  $x$  in a corpus;  $f(y)$  is the number of documents containing  $y$  in the corpus;  $f(x,y)$  is the number of documents containing both  $x$  and  $y$  in the corpus; and  $N$  is the number of items or words in the corpus.

For this research, we adapted the above formulas (equation 1 and 2) for Indonesian-English parallel corpora. And we estimated the mutual information value as follow (equation 3):

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} = \log_2 \frac{D(x, y)}{(D(x)+1) * (D(y)+1)} * (N_{indo} + N_{inggris}) \quad (3)$$

Where  $D(x)$  is the number of Indonesian documents that contain Indonesian word  $x$  (exclusive);  $D(y)$  is the number of English documents that contain English word  $y$  (exclusive);  $D(x,y)$  is the number of Indonesian-English document pairs that contain Indonesian word  $x$  and English word  $y$ ;  $N_{indo}$  is the number of items or words in the Indonesian corpus;  $N_{inggris}$  is the number of items or words in the English corpus.

### 3.2 Query Expansion

After translating the Indonesian query into English using mutual information, we applied a pseudo relevance feedback technique to expand the query. Query expansion is a process of adding words found in a certain number of top English documents retrieved [2] into the query. We used language model formula in choosing the best words to be added to the query [6].

## 4 Experiment

We evaluated our technique using queries and document collection from the Cross Language Evaluation Form (CLEF) [13]. The Indonesian query sets are taken from the CLIR-CLEF 2006. The document collection contains 169.478 articles from Glasgow Herald (1995) and Los Angeles Time (1994) newspapers.

At beginning, we created the Indonesian version of the English document collection by translating them using Transtool, a commercial machine translation software package to build Indonesian-English parallel corpus. Based on this parallel corpus, we built Indonesian-English word pair list using mutual information. We used this word pair list to translate Indonesian queries into English queries. Then we used Lemur<sup>1</sup> to index English document collections, and retrieve English relevant documents based on the English queries. To improve the result, we also applied query expansion to the translated queries.

Besides using mutual information technique, in this work we also did some experiment in Indonesian-English CLIR. We translated Indonesian queries into English using bilingual dictionary, two machine translations: Toggletext<sup>2</sup> and Transtool, and pseudo-translation based on parallel corpus. In the pseudo-translation technique, the Indonesian queries were run to retrieve Indonesian documents and found their parallel English

<sup>1</sup> <http://www.lemurproject.org/>

<sup>2</sup> <http://www.toggletext.com/>



documents in the top 5 retrieved documents. Then from the top 5 retrieved English documents, the first English word that has the highest weight as the translation of the Indonesian query word was taken. The weight of a word was calculated using the *tf.idf* formula [2].

## 5 Result and Analysis

The performance of the Indonesian query set in retrieving the English documents is shown on Table 1. The best mean average precision is achieved by translating the Indonesian queries using the machine translation tool. Using Toggletext online machine translation achieves 84.92% of the equivalent monolingual performance, and using Transtool machine translation achieves 78% of monolingual performance. Translating the query using bilingual dictionary gives 51.98% of monolingual performance.

**Table 1.** The Mean Average Precision (MAP) of monolingual English queries, the Indonesian queries translated using bilingual dictionary, Toggletext and Transtool machine translation

Technique	MAP
Monolingual	0,3242
Dictionary	0,1685 (-48,02%)
M. Translation ( <i>Toggletext</i> )	0,2750 (-15,18%)
M. Translation ( <i>Transtool</i> )	0,2529 (-22,00%)

The performance of the Indonesian queries translated using a parallel corpus is shown in Table 2. Translating Indonesian queries using pseudo-translation technique based on parallel corpus achieved 69.25% of monolingual performance. Translating Indonesian queries using mutual information based on parallel corpus only achieves 33.47% of monolingual performance. Applying the query expansion technique to the result of using mutual information based on parallel corpus increases the mean average precision by 8.39% of 41.9% of monolingual performance.

**Table 2.** The Mean Average Precision (MAP) of the Indonesian queries translated based on parallel corpus using pseudo-translation and mutual information technique

Technique	MAP
Parallel-Pseudo Translation (PT)	0.2245 (-30.75%)
Parallel-Mutual Information(MI)	0.1085 (-66.53%)
Parallel- Mutual Information with query expansion (MI-QE)	0.1357 (-58.14%)

The retrieval performance of Indonesian queries translated into English depends on the quality of various language tools and resources used. The bilingual dictionary gives the worst result because many Indonesian query words are not in the dictionary. The performance of the Indonesian query translated into English using two machine translation tools also varies. Toggletext machine translation performs better than

Transtool machine translation as the number of Indonesian words that are failed to be translated is less than that of using Transtool machine translation. The mutual information based technique, unfortunately, does not produce a good result. Applying the query expansion technique helps increase the mean average precision slightly.

From the Indonesian-English word pair list that was built using mutual information, we see that the highest value of mutual information does not always give correct English word as translation. For some Indonesian words the highest value of mutual information resulted correct English words, but for others the highest value of mutual information resulted wrong English word. From the result of translation process using mutual information for 260 Indonesian words occurred in queries, we found that there were 110 Indonesian words that had a correct English from the highest value of mutual information (first rank), some example of Indonesian word for this case are shown in Table 3.

**Table 3.** Example of Indonesian words with a correct English word as candidate translation from the highest value of mutual information (first rank)

Indonesian word	English word from mutual information (first rank)
merek	trademark
keadaan	circumstance
pengangguran	unemployment
visa	visa
bahaya	danger

There were 51 Indonesian words had a correct English word from second value to fifth value of mutual information. Table 4 shows some example for these words.

**Table 4.** Example of Indonesian words with a correct English word as candidate translation from second to fifth value of mutual information (second rank to fifth rank)

Indonesian word	English word from mutual information first rank to fifth rank	Correct English word	Rank
africa	afrikan, african, kwazulu, AFRICA, gatsha	AFRICA	4
perawatan	aftercare, surgery, TREATMENT, carefree, arthritis	TREATMENT	3
bijaksana	discreet, PRUDENT, wiser, wisdom, indiscreet	PRUDENT	2
main	romp, overplay, PLAY playground, nut	PLAY	3

There were 44 Indonesian words got wrong English word as translation. Table 5 shows some example of Indonesian words for this case.

**Table 5.** Example of Indonesian words with wrong English words as translation

Indonesian word	English word from mutual information (first to fifth rank)
pengaruh	relentless, unsuspected, mindful, favor, abscond
produksi	latch, lug, lubricant, rapacity, sewage
universitas	conniption, petal, ulna, grantee, packager
pengetatan	unsoundly, bluestocking, sobriety, alfresco, detain
metode	bugle, wallop, malice, seam, trickery

The rest of words (55 words) did not get a best correct English word. From the English word, we can divide the results into four cases below:

1. Some Indonesian words get antonym word in English, for example Indonesian word “adil” that means “justice” in English. Using mutual information, “adil” was translated into “unjustice” (“unjustice” is antonym of “justice”).
2. Some Indonesian words get related word in English. For example Indonesian word “matahari” that means “sun” in English. Using mutual information, “matahari” was translated into “sunlight”, “solar” (“sunlight”, “solar” are related to “sun”).
3. Some Indonesian words get English word but not the best translation. For example Indonesian word “putri”, the best translation for “putri” in English is “princess”, but using mutual information “putri” was translated into “daughter”. This is correct but not the best one.
4. Some Indonesian words get English word that is a translation for other variation of the Indonesian words. For example Indonesian word “acara” means “agenda” in English, but using mutual information “acara” was translated into “lawyer”. “Lawyer” means “pengacara” in Indonesia. “Pengacara” is one of variation of word “acara”.

Second example for this case, Indonesian word “alam”. Word “alam” means “nature” in English, but using mutual information “alam” is translated into “undergo”. Word “undergo” means “mengalami” in Indonesian. Word “mengalami” is one of variation of word “alam”.

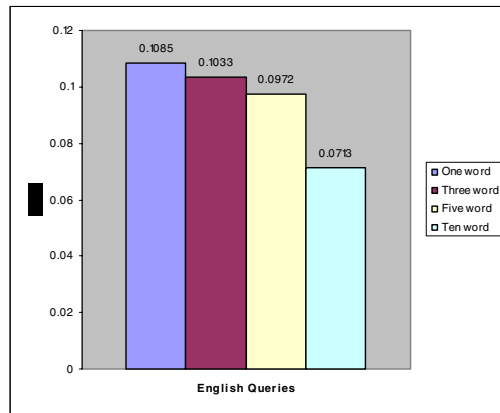
We analyzed this fourth case occurred because we applied stemming process to Indonesian-English parallel corpus when built bilingual word list using mutual information.

In Indonesian queries, there are seven phrases. From the nine phrases we find only three phrases get the correct translation in English using mutual information. Table 6 shows the phrases.

While finding the best result in translating Indonesian queries into English using mutual information, we took English words in the first rank, top three ranks, top five ranks, and top ten ranks that had the highest mutual information values. Figure 1 showed the result of these four variations. The best result was achieved using the English word that appeared in the first rank.

**Table 6.** Phrases occurred in Indonesian queries

Indonesian phrase	In English	Using Mutual Information
bahan bakar	fuel	firewood explosive
energi atom	atomic energy	<b>atomic energy</b>
gerhana matahari	solar eclipse	<b>solar eclipse</b>
gempa bumi	earthquake	quake
lintas alam	cross country	undergo straightaway
perdana menteri	prime ministry	morihiro
sidang pengadilan	trial	unjustice convocation
uang sekolah	tuition	<b>tuition</b>
undang-undang dasar	constitution	basic invitation



**Fig. 1.** The Mean Average Precision (MAP) of Indonesian queries translated into English using mutual information with different number of English words as translation for one Indonesian word

From all the result, we find that mutual information could rank the words in good order based on value of mutual information to get best translation, but because mutual information use property of observing word  $x$  and word  $y$  that occurred together, some time this property resulted in wrong or not best translation.

## 5 Conclusion

Our evaluation on Indonesian resources and tools has provided us with some insight on the issues in Indonesian-English translation. The machine translation techniques are the best translation method at this moment compared to using bilingual dictionaries available on the Internet. However, the result of our evaluation of the parallel corpus that we created using the machine translation techniques indicates that there is room for improvement to explore in our future work. We plan to explore better techniques in finding bilingual word pairs and also applying better query expansion techniques to the result.

## References

1. Aljlal, M., Frieder, O.: Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation. In: Proceedings of the tenth international conference on Information and knowledge management, pp. 295–302. ACM, New York (2001)
2. Baeza-Yates, R., Ribiero-Neto, B.: Modern Information Retrieval. Addison Wesley, New York (1999)
3. Braschler, M.: Combination Approaches for Multilingual Text Retrieval. Information Retrieval 7, 183–204 (2004)
4. Church, K.W., Hanks, P.: Word Association Norms, Mutual Information and Lexicography. Computational Linguistic 1, 22–29 (1990)

5. Fano, R.: *Transmission of Information*. MIT Press, Cambridge (1961)
6. Grossman, D.A., Frieder, O.: *Information Retrieval: Algorithms and Heuristics*, 2nd edn. Springer, Netherland (2004)
7. Hayurani, H., Sari, S., Adriani, M.: Query and Document Translation for English-Indonesian Cross Language IR. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) *CLEF 2006*. LNCS, vol. 4730, pp. 57–61. Springer, Heidelberg (2007)
8. Hedlund, T., et al.: Dictionary-Based Cross-Language Information Retrieval: Learning Experiences from CLEF 2000–2002. *CLEF 2000 7*, 99–119 (2004)
9. Jang, M.G., Myaeng, S.H., Park, S.E.: Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, Morristown, NJ, USA, pp. 223–229 (1999)
10. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-Lingual Relevance Models. In: *The Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 175–182. ACM Press, New York (2002)
11. McCarley, J.S.: Should we translate the documents or the queries in cross-language information retrieval? In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 208–214. Association for Computational Linguistics Morristown, NJ, USA, (1999)
12. Mititelu, V.B., Ion, R.: Cross-Language Transfer Of Syntactic Relations Using Parallel Corpora. In: *Cross-Language Knowledge Induction Workshop, Romania (2005)*
13. Peters, C.: *Cross Language Evaluation Forum (2007)*,  
<http://www.clef-campaign.org/2007/2007agenda.html>
14. Qiu, Y., Frei, H.: Concept Based Query Expansion. In: *Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 160–169. ACM Press, New York (1993)
15. Sheridan, P., Braschler, M., Schäuble, P.: Cross-language information retrieval in a multilingual legal domain. In: *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pp. 253–268 (1997)
16. Talvensaari, T., et al.: Corpus-based cross-language information retrieval in retrieval of highly relevant documents. *JASIST* 58, 322–334 (2007)
17. Yang, Y., et al.: Translingual Information Retrieval: Learning from Bilingual Corpora. *Artificial Intelligence* 103, 323–345 (1998)

# Exploring User Experiences with Digital Library Services: A Focus Group Approach

Kaur Kiran and Diljit Singh

Department of Information Science  
Faculty of Computer Science & Information Technology  
University of Malaya  
Malaysia

kiran@um.edu.my, diljit@um.edu.my

**Abstract.** This study is a description of how university students perceive digital library services, or the integration of digital library elements within the traditional or hybrid library services in academic libraries. These services were identified through a preliminary study that evaluated the digital services offered by twenty university libraries. The top five university libraries with most services offered were selected. Qualitative method of data collection using ten focus groups consisting of 81 postgraduate students was employed. Overall, the focus group discussions reveal that digital services are very well received by users. The most popular being access to online databases, e-journals and e-books. Users almost unanimously request for more online help to assist in search for information and improve their information seeking skills. The results have implications for academic libraries providing digital services in a networked environment and recommendation made include providing better guidance to users, and solicit feedback to make evidence-based decisions on digital library service delivery.

**Keywords:** Digital Library, academic library, web-based services.

## 1 Introduction

The rapid expansion of the Internet and the World Wide Web (WWW) has had a great impact on how libraries are managing their collections and services. In an environment of increasing technological innovations, academic libraries must attempt to measure quality and effectiveness to align themselves with criteria by which higher education institutions are judged (Pritchard, 1996); justify financial investments (Kyrillidou & O'Connor, 2000) and respond to the changes brought about by the digital revolution. Technological progress has changed *how* libraries do their work, not *why* (Kuny & Cleveland, 1996). In the academic library scenario, the digital library has expanded from mere means of offering a collection of digital objects that people can access from their desktop, to a means of offering innovative library services over a networked environment, mainly the web. Henderson (2005) points out that “virtual library”, “digital library”, “electronic library” are all terms used to describe libraries and information services delivered via the Internet. These services

may include access to purchased databases, catalogue databases, electronic theses & dissertations, electronic document delivery, reference service, forums, internet resources, and others (Venkatalakshmi & Sonker, 2002).

As there have been significant efforts in the development of digital libraries, there is need to identify evaluation strategies to assess digital libraries that can yield a variety of data regarding the efficiency, effectiveness and quality of digital libraries (Bertot, 2004). In his review of digital library assessment approaches, Bertot underlines that *user-centered* approaches evaluate the quality of the presentation of resources and services, and the inclusion of user needs. One must observe changes in the information environment to realize that a new breed of users is evolving. This new breed of self-sufficient users do not see the library as the centre of their information environment - they rely on the Internet and WWW. The competition faced by libraries from other information providers such as Google or even Amazon, has been recognized and discussed but very few libraries have taken the challenge seriously and taken steps to actively formulate strategies to counter these developments (Bawden & Vilar, 2006). The availability of digital resources is not enough if not complimented by additional services to support activities that occur during the information seeking process.

This brings up the issue of what users want, when, where and how? Even the most experienced and knowledgeable librarian cannot predict user's needs and expectations. Therefore it is necessary for us to find out what users want, and what more effective way than to ask them directly? This study is part of an ongoing research to develop an approach to measure digital library service quality, specifically in academic libraries. In this initial phase, a focus group methodology is applied to gain insights on how users, particularly post graduate students, perceive digital library services.

## 2 Literature Review

A review of the literature confirms that that there is substantial lack in research relating to the uptake of electronic information services offered via digital libraries. Research has focused on areas such as design, architecture and functionality. Francisco-Revilla et. al. (2001), report that digital libraries are increasingly being defined as ones that collect pointers to WWW-based resources rather than hold the resources themselves. Greenstein (2000) shares this view and states that the digital library is known less for the extent and nature of the collections it owns than for the networked information space it defines through its online services. Jeng (2005) concludes that these views can be seen as calling a library website a portal-type digital library.

The service aspect of a digital library has been noted in the literature (Borgman, 1992; McMillan, 2000; Chowdhury & Chowdhury, 2003; Choi, 2006). According to McMillan, a library is a fusion of resources in a variety of forms, including services and people supporting the entire life cycle of information. Often referred to as the 'virtual', 'electronic' or digital' library, these libraries have several different definitions or explanations attributed to it. Choi (2006), in his study on reference services in digital libraries, revealed that the availability of digital resources is not enough if not complimented by additional services to support other activities that occur during the information seeking process.

Although research has been conducted on digital libraries in the past decade, the majority of the literature concerns design and evaluation of the use of the homepages. Bertot (2004) in his review of the literature, expresses research in digital library services in four broad areas: outputs assessment, performance measures, service quality and outcome assessment, all of which may be library-centered or user-centered. Shropshire (2003) found that the literature commented only on the beginning (design) and the end (evaluation) of the process. Zhou (2003) examined the history of Web portals and their development in libraries and reported that only very few libraries have adopted the portal technology despite wide spread use of my.yahoo.com -type.

In 2000, Lee & Teh evaluated the content and design of 12 Malaysian university library Web sites. Their study revealed that generally university libraries have set up well-designed and useful Web sites, but a very small number offered internal electronic databases, including examination papers, thesis abstract, university publications and public lectures, external electronic databases and some form of electronic reference service. Later in the same year Kiran (2000) examined the content and functionality of 13 public and private Malaysian university libraries on the Web and drew similar conclusions. Suggestions were made for academic libraries to take advantage of the networked access of the Internet and offer resources and services for the virtual user. As the number of databases may be limited by the expenses incurred, libraries could use the World Wide Web to build subject specific links and offer value-added information service to its user community. In 2006, Kiran found that more Malaysian academic libraries had begun to offer services such as web-based OPAC, access to online databases, e-journals, e-books, digitized local content, online forms for document delivery requests, online reserves and requests for purchase, online reference and online information searching skills training, via the library web site. Since a library is not expected to house all that there is available, the librarian should stop and ask the users if this is what they want, or what is it that they want?

### 3 Methodology

An initial examination of university libraries on the Web was carried out to identify libraries that offer digital services or may also be referred to as web-based services. This study used a qualitative data collection approach using focus group discussions. The purpose of the focus groups was to learn about postgraduate students' perception and experiences with digital library services available via the academic library web site. A total of ten groups, comprising of 81 participants were carried out from January 2008 to March 2008. Participants were selected by issuing an e-mail invitation to diverse faculties in all five institutions.

Each focus group lasted about 60-90 minutes and was audio recorded. All responses were separated to individual statements and categorized according to revealing issues. A second phase of categorization further consolidated the issues and allowed for broad categorization and identification of issues in each category.



## 4 Results and Discussion

All participants were first asked to describe their familiarity with digital library services, where and when they accessed these services, what makes their experience memorable or difficult and their overall perception of the services. Overall it is found that a majority of the postgraduate students agreed that they used the digital library services very frequently for preparation of assignments and for literature search for their research activities. The most common usage is the online databases, e-journals and e-books. Since these students are mainly part-timers, they access digital services from their workplace or at home, specially late at night. A small number which consisted of mainly international students stated that they use the digital library services at the library, using either the PCs at the computer labs or own laptops. Among the memorable experiences, though very few responded to this, were ;

*“fantastic! The first time I was shown the online databases by a librarian and I could download latest articles still ‘in press’ – that was so wonderful..I told the whole class about it”*

*“once a badly needed an article which was not full-text..so I went to the library to ask how I can get...the librarian showed me how to ask for the article by filling in a form on the web site...and I could do it from home...I got the article in 2 weeks...its so useful”*

Participants were encouraged to channel their thoughts as their perception of the performance of the digital library services. After reviewing the transcripts, several main issues emerged, which will be discussed in the following categories:

1. The environment in which the digital library service is offered
2. The delivery process of the digital library services
3. The outcome of using the digital library services

### 4.1 The Environment in Which the Digital Library Service Is Offered

One of the major issues discussed among all groups was the environment in which these services are provided. Easy access to the web page containing these digital services was a main concern. Although they were appreciative of the availability of remote access and being able to access services at any time of the day, most participants expressed the difficulties faced to login into the system. Only two of the libraries use a proxy server for remote access that allows users to login in only once to access all the digital content, others need users to know a set of different usernames and password for each online resource that they want to access, may it be a subscribed databases or a institutional database of digitized resources. Consequently users are put off when they don't know the particular password for a database and rather turn to the Internet for quick and easy access.

Besides accessibility issues, users also place importance on the appearance of the site and want all links to be up-dated and working. Since quite a number of students

**Table 1.** Focus group description of the digital library service environment

Focus Group Participant Description of the digital library service environment:	
Accessibility	<ul style="list-style-type: none"> <li>• “User friendly- simple and easy to login”</li> <li>• “Why do we have to login. The system should be able to recognize us from campus. If out of campus I understand- maybe for security but it is troublesome to remember so many passwords.”</li> <li>• “Password is given but different for every database. Why not login once and can search all databases.”</li> </ul>
Site design and layout	<ul style="list-style-type: none"> <li>• “The site is very friendly...I like the simple look”</li> <li>• “There is so much of text in one single page. I have to use my glasses and better layout. Not complicate but crowded”.</li> </ul>
System reliability	<ul style="list-style-type: none"> <li>• “Server must be OK all the time...you finally get to a working PCs [at the library] and then find the system so slow ..takes ages for page to appear”</li> <li>• “It’s so frustrating when the system hang just when you click submit[online form].”</li> </ul>
Links	<ul style="list-style-type: none"> <li>• “I think it is important that all the links are good [working]”</li> </ul>
Equipment	<ul style="list-style-type: none"> <li>• “Wireless not good at the library...also not enough PC so I prefer to use my notebook.”</li> </ul>

use the equipment at the library to access digital services, they want enough PCs and wireless access to be available. As one international student puts it “ *so rich in information ...but so difficult to get a good working PC to access!*”

#### 4.2 The Delivery Process of the Digital Library Services

As participants were probed to discuss the digital services that they used, many begun to bring up the issue of *ease of search*. Mainly there were two broad issues here, one being the ability to recognize what were the services being offered on the site and the other was having the ability to search for relevant information as quickly as possible. Many users prefer to have information resources arranged by research areas or subject discipline. One concern expressed by the users was their lack of skills in searching for information. Many expressed the “need for online help”, let it be technical or to search the resources. Though some libraries provided online help for searching databases, most of the help was in PDF files supplied by the database vendors. Participants are more receptive towards an “online tutorial” or clear interactive guidelines on how to search. Most users were aware that some restrictions on access were placed by online database vendors and library had little control over them, however they hope that the library continually provide access to a comprehensive, up-to-date, relevant, trustworthy and accurate information sources.

Besides digital information resource provision, it is very important that libraries uphold good communication channels with their users. Some participants said that they do not know who exactly to contact as usually only the webmaster’s email is given. Participants look forward to be told what the library has to offer and kept up-to-date about the changes and improvements in the digital library services. Many participants in the focus groups were unaware of online document delivery request and online reference services. They had very negative perception towards librarians. Librarians are generally considered as unapproachable and not willing to help. Troubleshooting or help when users are in difficulty is also asked for. Participants expressed the need for support from the library in the form of a 24 hour online helpdesk that would assist them in technical problems, login and also searching for information.

**Table 2.** Focus Group Participant Description of the digital library service delivery

Focus Group Participant Description of the digital library service delivery:	
Organization of information	<ul style="list-style-type: none"> <li>• “Maybe they could list relevant material according to faculty or field”</li> <li>• “List journal according to subject area, eg; these are the journals for the education”</li> <li>• “Online databases too complicated, too many subjects”</li> </ul>
Ease of searching/use	<ul style="list-style-type: none"> <li>• “he problem is that the search terms you use for different database. Time consuming to use different suitable search terms for different databases. The library can create ontology- very important”</li> <li>• “Not as easy to search as Google”</li> </ul>
Information Content	<ul style="list-style-type: none"> <li>• “So far I am happy with the journals content and find it relevant to my needs...”</li> <li>• “content of the dbases is comprehensive – very good - get all articles full text”</li> <li>• “..with the library databases the information can be trusted... other web sources maybe cannot trust so much.”</li> </ul>
Communication	<ul style="list-style-type: none"> <li>• “Somehow I think I don't use is because I do not see a face – a human. I don't know anybody there. I would love if during the orientation the CL comes and introduce. There must be a relationship. I find myself very remote from the professionals in the library”</li> <li>• “I think the library staff is knowledgeable but we don't have contact with them”</li> </ul>
Customer support	<ul style="list-style-type: none"> <li>• “Students are not alert of this service[online reference]. Librarians should ... create awareness. If we cannot find we will know we can ask for help”</li> <li>• “So much information but we don't know. Nobody tells us.”</li> </ul>

### 4.3 The Outcome of Using the Digital Library Services

Discussion about what they experience after using the digital library services brought about some interesting responses. Basically, library users want digital library services that are dependable and create a feeling of assurance that their needs are taken care of. Participants feel that the library must make sure the fulfill users expectations and make them feel important. Ignored feedback from users is not appreciated and discourages users from returning to the service. What was assuring was that some participants agreed that frequent use of digital services has improved their confidence in searching for information and consolidating the relevant resources from the vast resources.

Focus Group Participant Description of the digital library service outcome:	
Reliability	<ul style="list-style-type: none"> <li>• “If they claim they have it must be there”</li> <li>• “Very risky to e-mail because a lot of virus. Very hard to complain to the library because a lot of students use. I don't think they can control.”</li> </ul>
Assurance	<ul style="list-style-type: none"> <li>• “The question is how fast can we get an answer”</li> <li>• “Won't take any action anyway even though a complaint form is filled.</li> </ul>
Self-reliance	<ul style="list-style-type: none"> <li>• “During my first degree there was a tour by the library so I am quite familiar”</li> <li>• “After a few times I feel confident in using the digital library”</li> </ul>
Functional benefit	<ul style="list-style-type: none"> <li>• “When the article I need is found I am happy. If article hard to find it is frustrating.”</li> <li>• “quickly get what I am looking for”</li> </ul>

Overall the participant felt that as long as the “*get what I want*” and as “quickly” as possible the digital library service is performing well.

#### 4.4 Area of Concern

During the discussion with all ten groups, the reference to Google was inevitable. Participants tend to compare the digital library service with Google. Among the comments were :

*“I use Google first then I use database to search for the article - because it is free, faster and easier.”*

### 5 Conclusion and Recommendations

It is concluded that basically there are three main areas of concern when providing digital library services, the environment in which the service is provided, the delivery of the service and the outcome factor. Within these three categories, main issues of concern are

- Easy access – users want fast and convenient access. There has to be enough equipment (PCs) for access, including wireless access. Username and passwords are an inconvenience to users.
- Need help - these self-reliant Internet savvy users seem to need help in terms of technical help at the point of use and also help to search for relevant information. Within the digital library, with or without a digital librarian, providing reference and user education is not to be compromised.
- Communication – there seems to be a barrier between users and service providers in the digital library. Unlike other services, such as e-retailing, e-commerce, etc., that removes the service provider, digital libraries cannot completely remove the human interaction. Users seem to want someone in the digital environment that can help build a relationship and guide users to be self-reliant information seekers.

What use is a fantastic collection of digital resources if users can't get to or are frustrating when using because the process is so complicated. It is of concern when many of the participants in this study made comparisons of the digital library to the service provided by Google! Their main concern seems to be “easy”, “fast” and “sure to get something”. The revelation that lack of information searching skills is a main factor users are turning to Web search engines to search for information should have digital librarians working towards educating users with skills that would make it easier to search online databases and gain confidence to be able to select and consolidate relevant information.

Current online information services, including digital libraries expect users to know what they want and formulate queries to represent their information needs, which is often not easy (Meyyappan, Foo and Chowdhury, 2004).

The digital library environment requires a new technical and social set of competencies for librarians which were not previously required. Not only do they need to

have the expertise to fully exploit the capabilities of technology to provide value added digital services, they also need learn how best to communicate with users in an online environment. As McMillan (1999) presents it : “ *We have new roles to fill. While the format of our resources may change, while access to information may change, while styles of service may change, the vision of high quality, service-oriented, information centres still fits the library's mission. We will serve our user communities best if we incorporate this into the [digital library].*”

By these insights and recommendations, it is hoped that digital library services will be able to entice more users to discover and efficiently use them.

## References

- Bawden, D., Vilar, P.: Digital libraries: to meet or manage user expectations. *Aslib Proceedings* 58(4), 346–354 (2006)
- Bertot, J.C.: Assessing digital library services: approaches, issues and considerations (2004), <http://www.kc.tsukuba.ac.jp/dlkc/e-proceedings/papers/dlkc04pp72.pdf>
- Borgman, C.L.: The invisible library: Paradox of the global information infrastructure. *Library Trends* 51(4), 652 (2003)
- Choi, Y.: Reference services in digital collections and projects. *Reference Services Review* 34(1), 129–147 (2006)
- Chowdhury, G.G., Chowdhury, S.: Introduction to digital libraries. Facet, London (2003)
- Francisco-Revilla, L., et al.: Managing change on the Web. In: Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 67–76. ACM Press, New York (2001)
- Greenstein, D.: Digital libraries and their challenges. *Library Trends* 49(2), 290–303 (2000)
- Henderson, K.: Marketing strategies for digital library services. *Library Review* 54(6), 342–345 (2005)
- Jeng, J.: What is usability in the context of the digital library and how can it be measured? *Information Technology and Libraries* 24(2), 47–56 (2005)
- Kaur, K.: Malaysian university library Web sites: content and functionality. *Kekal Abadi* 19(2), 1–6 (2000)
- Kaur, K.: Digital libraries in Malaysia: an overview with an E-services perspective. In: TERI (ed.) *Information Management for Global Access: International Conference on Digital Libraries*, New Delhi, India, December 5-8 (2006)
- Kuny, T., Cleveland, G.: Digital libraries: myths and challenges. In: 62nd IFLA General Conference, August 25-31 (1996), <http://www.ifla.org/IV/ifla62/62-kuny.pdf>
- Kyrillidou, M., O'Connor: ARL statistics, 1999-2000. Association of Research Libraries, Washington (2000)
- Lee, K.H., Hai, T.K.: Evaluation of Academic Library Web Sites in Malaysia. *Malaysian Journal of Library and Information Science* 5(2), 95–108 (2000)
- Meyyappan, N., Foo, S., Chowdury, G.: Design and evaluation of a task-based digital library for the academic community. *Journal of Documentation* 60(4), 449–475 (2004)
- McMillan, G.: The digital library: without a soul can it be a library? In: *Books and bytes: Conference Proceedings: 2000 VALA Biennial Conference and Exhibition, VALA, Melbourne* (2000)

- Pritchard, S.M.: Determining quality in academic libraries. *Library Trends* 44, 572–594 (1996)
- Shropshire, S.: Beyond the design and evaluation of library web sites: an analysis and four case studies. *Journal of Academic Librarianship* 29(2), 95–101 (2003)
- Thong, J.Y.L., Hong, W., Kar, Y.T.: What leads to user acceptance of digital libraries? *Communications of the ACM* 47(11), 79–83 (2004)
- Venkatalakshmi, K., Sonker, S.K.: Challenging aspects of Digital Information Services (2002), <http://eprints.rclis.org/archive/00000307/01/CAFDIS.pdf>
- Zhou, J.: A history of Web portals and their development in libraries. *Information Technology and Libraries* 23(3), 119–128 (2003)

# Beyond the Client-Server Model: Self-contained Portable Digital Libraries

David Bainbridge<sup>1</sup>, Steve Jones<sup>1</sup>, Sam McIntosh<sup>1</sup>, Ian H. Witten<sup>1</sup>,  
and Matt Jones<sup>2</sup>

<sup>1</sup> Department of Computer Science  
University of Waikato  
Hamilton, New Zealand  
{davidb,stevej,sjm64,ihw}@cs.waikato.ac.nz

<sup>2</sup> Department of Computer Science  
University of Swansea  
Swansea, U.K.  
matt.jones@swansea.ac.uk

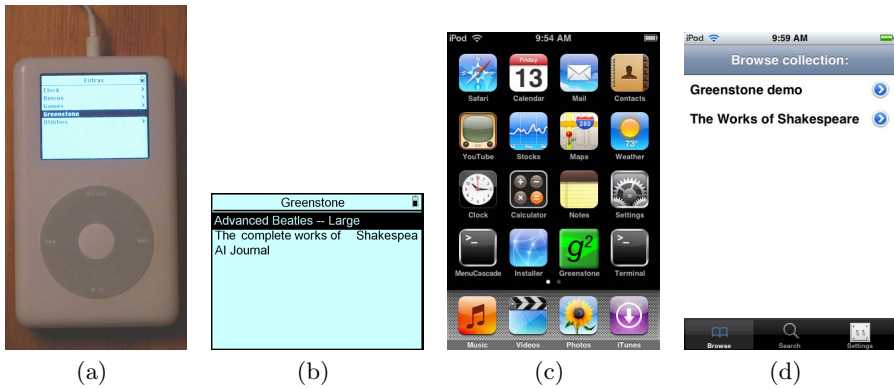
**Abstract.** We have created an experimental prototype that enhances an ordinary personal media player by adding digital library capabilities. It does not enable access to a remote digital library from a user's PDA; rather, it runs a complete, standard digital library server environment on the device. Being optimized for multimedia information, this platform has truly vast storage capacity. It raises the possibility of not just personal collections but entire institutional-scale digital libraries that are fully portable. Our implementation even allows the device to be configured as a web server to provide digital library content over a network, inverting the standard mobile client-server configuration and incidentally providing full-screen access. Anecdotal yet compelling evidence as to the usefulness of being able to access a self-sufficient digital library anytime, anywhere is given through an example built from the PDF files of the Joint ACM/IEEE digital library conference. Other examples include the Complete Works of Shakespeare and collections on Humanitarian Aid. This paper describes the facilities we built, focusing on interface issues that were encountered and solved.

**Keywords:** Portable Digital Libraries, Personal Media Player, Multimedia.

## 1 Introduction

Access to digital libraries is predominantly through the web. Indeed, the assumption that the user is working through a remote web server from a desktop machine or portable device is almost a given, and other modes of access are virtually excluded.

Some researchers have investigated access to digital libraries from portable devices, e.g. [1,3,4]. They envisage a scenario whereby a personal digital assistant (PDA) connects, using wireless networking, to a central digital library server.



**Fig. 1.** Invoking Greenstone on the iPod: (a) & (b) a click-wheel iPod; (c) & (d) an iPod-touch

The interface is web based, and the research involves tailoring it for a small screen (augmented, in some cases, with suitable data caching policies). Such work extends the existing web-based client-server model by adapting it for new kinds of client, a decision that is forced by the fact that most general-purpose PDAs do not begin to approach the storage capacity of centralized servers. To contemplate implementing the entire digital library itself on a portable device is a radical alternative.

There is an obvious incentive for designers of portable devices that are optimized for music and video to include vast amounts of disk storage, and general-purpose PDAs will probably never approach the storage capacity of such devices. Consequently we have considered whether it is possible to subvert personal multimedia devices to store entire digital library collections.

Figure 1 illustrates the result of our work, based on the iPod, and shows the user invoking the top-tier of digital library services (the collections available) on both an click-wheel iPod and iPod-touch. Later examples go in to more detail. But we stress at the outset that this choice of device is just an example, as is the choice of Greenstone [5] as the digital library software used. Our work is exploratory; to stimulate thinking on alternative digital library models in general—specifically, what happens when portable devices are equipped with large-scale storage capabilities.

To take one anecdotal example, one of the authors built a full-text searchable Greenstone collection of the 2008 proceedings of the ACM/IEEE Joint Conference on Digital Libraries, and as an experiment, ran it on an iPod-touch using the prototype presented in this paper. The source documents were provided to delegates on a USB pen-drive as PDF files and in a matter of 15 minutes this had been converted into a Greenstone collection and synchronized with the iPod. During the course of the conference the author—perhaps predictably—found himself looking details up on the iPod during sessions. Natural and convenient to use, a keyword or two was usually all that was necessary to quickly pinpoint the relevant part of an article. Potentially more surprisingly, the author found himself



still referring back to the iPod-based collection after the conference, accessing further information. This urge to turn to the iPod as the information source was instinctive, even though exactly the same information was simultaneously available on the author's laptop.

In this paper we discuss several developments aimed at tailoring the digital library software to a portable, self-contained, environment. For the generation 3–5 iPods, digital library browsing facilities were developed to utilize the unique iPod click-wheel for hierarchical browsing, and the traditional-style web-form based full-text search was altered to ameliorate the problems caused by the lack of keyboard. Also, launcher applications were created to display digital library documents, whether they be text, image, or audio media types. For the iPod-touch (and iPhone) with its touch-screen interface, usability issues were less severe. In both cases we also inverted the traditional PDA client-server model for digital libraries, and installed a web-server on the device that allowed it to serve content to others. Preliminary work on this was presented in [2], which is only about the click-wheel iPod and pin-points the technical issues we faced concerning implementation. The current paper takes a broader view of the topic, with interaction issues and usage scenarios the main focus, along with expanded details including its operation on the iPod-touch.

## 2 Motivation

Imagine being able to carry a library around in your pocket. Full fingertip access through searching and browsing to millions of items: text, image, audio, video, wherever you are. Really, *wherever* you are: no need to be within a wireless hotspot, or mess around with ISP registration (and worry about how much it is costing). No waiting for rich-media content to be transferred to your device, not to mention the accompanying rapid depletion of your battery power that goes with all that communication. Everything travels with you: no flight restrictions on planes, no worrying about connectivity in other countries. Your own personal copy of a large digital library, right there in your pocket.

A relatively low-cost Personal Multimedia Player (PMP) like the iPod can store an astonishing amount of information, and it is now possible to programme these devices and augment them with your own software—such as a digital library application, for instance. In terms of text storage, consider a library with one million books (corresponding to a medium-sized university library). At 80,000 words per book and 6 letters per word (including the inter-word space), each book comes to half a million bytes, or 150,000 bytes when compressed. The whole library amounts to about 150 GB, less than the capacity of a high-end PMP today. This is text only: it does not allow for illustrations. Even so, with the remaining 10 GB of free space you could also comfortably slot in all the articles from Wikipedia.

What about illustrated books? It's hard to generalize about the size of illustrated books because so much depends on their nature, but our experience

with digital library collections of humanitarian information (like the Humanity Development Library<sup>1</sup>) indicates that about 2000 fully-illustrated (and fully-indexed) books fit comfortably on a 650 MB CD-ROM: a 160 GB PMP would store almost half a million of these. Alternatively, over a quarter of a million pages from the New Zealand National Library's Papers Past collection<sup>2</sup> could be comfortably accommodated on such a device. This value includes the necessary files to support full-text indexing of the OCRd images, word highlighting of search terms when the images are viewed, and metadata to allow switching between page and article views. More articles could be uploaded if, for instance, the word highlighting capability (which in disk capacity accounts for roughly half of the supplementary files) was omitted.

And multimedia? Over 2,000 hours of audio or 200 hours of video fits comfortably onto the device. This is particularly attractive if literacy rates of end-users is of concern, as is the case in many developing countries. One doesn't necessarily have to purchase a top of the range device, and there is a buoyant market in second-hand personal media players—even Apple sell reconditioned older models through their website for considerably less than the list price.

### 3 Developing Portable Digital Library Services

The foundation to our work is the combination of open-source operating system and cross-compiler. iPod-Linux<sup>3</sup> is a project with the goal of porting Linux to the various click-wheel generation of iPod. Using the cross-compiler provided through this project, developers can use a host machine to generate executables compatible with the iPod and consequently run their own choice of software on this device—a wide variety of applications have been ported: from a humble text-entry notepad application to first person “shoot-em up” game, Doom. The iPhone and iPod-touch (which already run Unix natively) have also been “jail broken” and a suitable open-source cross-compiler made available for people wanting to have a freer range of control over the software they run on their device.

With the programming environment established, to develop portable self-contained DL services, the Greenstone runtime system—a CGI program intended to be run by a web server in response to each user request/query and produce web pages in raw HTML that are rendered in a web-browser—some alterations were required. Primarily the runtime system needed to be re-cast as a program intended to be run continuously, rather than being re-invoked by the web server for each individual operation. This task was greatly simplified by the internal structure to Greenstone that (even in a standalone CGI program) is cleanly divided in two through a protocol that separates a front-end receptionist concerned with presentation issues, and a back-end collection server responsible for providing full-text indexing and metadata browsing functionality.

---

<sup>1</sup> Available, along with many other humanitarian collections, at [www.nzdl.org](http://www.nzdl.org)

<sup>2</sup> [paperspast.natlib.govt.nz](http://paperspast.natlib.govt.nz)

<sup>3</sup> [www.ipodlinux.org](http://www.ipodlinux.org)

This was the lion's share of the work. Integrating the result into the PMP environment was straightforward. For the click-wheel iPod, the work was packaged as a dynamically loadable module that stipulates where in the main application's hierarchy it should go. For the iPod-touch the software was configured as a bundled application.

The DL reimplementations works directly from the files that are generated by a host machine running the standard Greenstone software. This setup means that any existing Greenstone collections can be transferred to the device, placed in the appropriate location, and displayed there without any intervention or conversion. Bulk file copy is a rather crude mechanism for synchronization, and this is an area that can be improved upon in a version that goes beyond our proof of concept phase; however, we were actually surprised at how well this worked in practice, given the file transfer utilities already available, in particular the Unix *rsync* command that can be used to transfer just the files that have changed between two file hierarchies.

### 3.1 Interactive Browsing

The next step was to map Greenstone's search and browsing functions into the iPod's interactive style, replacing the hyperlink navigation style implemented by the regular CGI version of Greenstone. In principle there is a clear and direct mapping of Greenstone's browsing capabilities (implemented by hyperlinks in a web page) into the iPod style of hierarchical menus (click-wheel or touch-screen). Greenstone collection designers can include a selection of browsing devices in their collection, like alphabetical lists and hierarchical browsers with arbitrary numbers of levels. All these browsing devices map naturally into the kind of hierarchy that users are accustomed to traversing with the iPod.

In order to achieve this technically, Greenstone's access mechanism had to be recast as an iPod interface. The iPod click-wheel software development suite incorporates a modest graphical user interface toolkit specially developed for iPod-Linux, called TTK. This provides a stacked-window system along with a high-level library including event handling and methods for input and graphical output, and forms an abstraction layer between application and lower-level graphics library. It is quite generic, but provides a few specific widgets to get people started: hierarchical menus, text viewer, image viewer, slider, popup windows. TTK is extensible and well documented. For the iPod-touch, a more sophisticated graphical development environment is available through UIKit, the iPod's native interface library. It is modelled very much along the lines of the standard Mac's application interface library AppKit, and similarly is written using Objective-C.

Greenstone's implementation of browsing works by accessing metadata information in a flat-file database. This was easy to port to the iPod. Had we been reliant on a relation database, such as MySQL, this would have been a entirely different proposition. We create menu items on the fly from the contents of the flat-file database as the user traverses the hierarchy, just as the regular CGI version of Greenstone does.

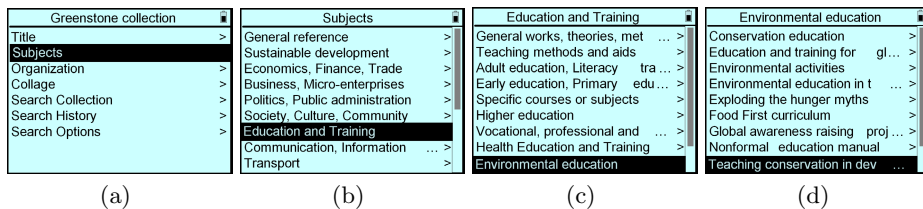


Fig. 2. Browsing a subject hierarchy

The PMP makes navigating hierarchies very natural. Greenstone distinguishes between browsing structures like AZ List for showing alphabetically ordered lists, Date List for showing date selectors, and Hierarchies for showing hierarchical metadata (like classifications). As far as the PMP is concerned, all these are mapped into hierarchies. The AZ List becomes a two-level hierarchy with an alphabetical selector at the first level.

Figure 2a shows the various ways of accessing these hierarchical structures in a particular collection: browsing by title, subject, organization, and collage; as well as searching: these reflect the items that would appear in as a clickable horizontal-bar in the traditional Web-based Greenstone interface. Subject browsing is selected; then Education and Training; then Environmental Education; and finally the specific book Teaching conservation in developing nations. Selecting that would yield a display of pages that can be perused in a similar manner. As with standard iPod, the user can travel back through the series of menus using the iPod's menu button. The whole process takes place very quickly: this kind of interaction is second nature for iPod users.

### 3.2 Viewing Documents

When a leaf of the browsing hierarchy is reached, Greenstone displays the document itself. To emulate this on the PMP, document launcher applications were written for text, image and audio media types, so when a user browsed to a leaf node an appropriate action would be triggered. Text and image launcher applications made use of the TTK toolkit.

Surprisingly, it is not possible to access the iPod's native audio playing functionality from within iPod-Linux. Consequently presentation of audio documents was a challenging task. We were faced with the ironical situation of having a digital library system on a portable music player that could do everything but play audio! Eventually TTK was used in combination with the Linux digital signal processing file-mapped device, enabling the user to play an audio file, pause it, fast-forward and rewind it.

The display of textual documents is probably the weakest part of the Greenstone iPod click-wheel implementation as far as practical deployment is concerned. We have been unable to locate a HTML renderer for TTK, and so such documents cannot be viewed properly. Our interim solution is to convert such documents to plain text before displaying them, but this eliminates all the formatting, hyperlinks, images, and so on. Greenstone collections do not have to

use HTML documents—there is a standard plugin for plain text files—but in practice most of them do. Until a proper HTML renderer can be found, document display will remain unsatisfactory for most collections. The situation for the iPod-touch is much more favourable, through its more sophisticated interface library. See Section 3.4 for an example using this version of the device.

### 3.3 Searching

Full-text search has been a fundamental capability of this digital library software from the very beginning. In the absence of any metadata, searching is the primary access mode for textual documents, and as far as the user is concerned it comes for free.

When implementing full-text search for the generation 3–5 of the iPod, we came up against its minimal user interface. Although perfectly—beautifully!—designed for selecting and playing audio, with click-wheel and five “push” buttons (play/pause, fast-forward to end, rewind to beginning, menu and enter), there is no keyboard.

TTK comes with various text entry widgets: from a full range of characters displayed linearly on the screen (a-z, 0-9, punctuation, ...) through which the click-wheel scrolls forward and backwards, to tapping out characters using Morse code. We did not attempt to come up with any significant innovation in this area, and decided to acquire the text for query terms using the linear selector.

Once acquired, initiating a query and displaying the search results is straightforward. However, text entry is so painful that we decided to make past search results readily accessible. In Figure 3 the user is searching Shakespeare for love. Having selected the collection (Figure 3a) they select Search (Figure 3b) and enter the search term (Figure 3c). Returning to the collection’s main menu, they find that the search results have been added to the access list (Figure 3d). Selecting this item they see the results (Figure 3e) and can then access a particular

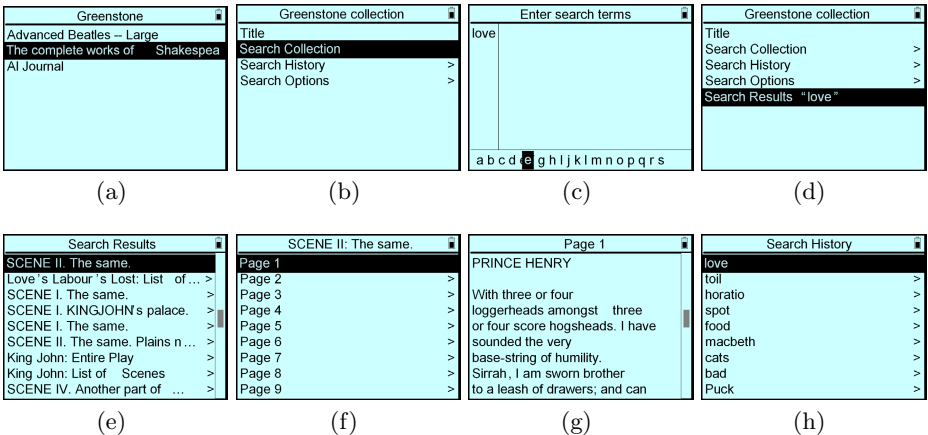


Fig. 3. Searching Shakespeare for love

document. Except for the actual text entry, all this is far easier to do than to describe in print: to iPod users, it feels perfectly natural.

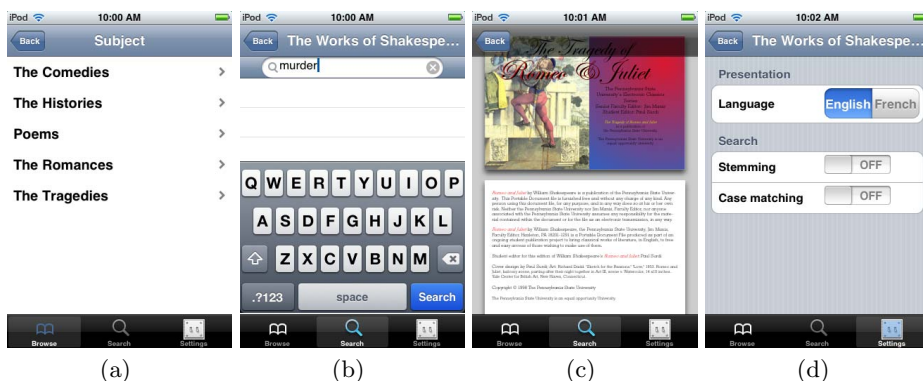
There is a generic search history feature in Greenstone, but it does not associate past queries with particular collections. For the iPod, we decided to associate search history with each collection, and make it persist across sessions (whereas the search results in Figure 3d will disappear if the collection is re-invoked). Figure 3h shows the history list. It is particularly useful because selection from a list incurs a far smaller overhead on the iPod than retyping a query.

### 3.4 Browsing and Searching on the iPod-Touch

Figure 4 shows various snapshots taken from accessing the iPod-touch version of the *Complete Works of Shakespeare*. Figure 4a shows the start of the subject hierarchy to this collection. Touching one of the items causes the screen to scroll left, and the next level be shown. Back is located as a button in the top-left corner.

Along the bottom of the screen are tabs for *Browse*, *Search*, and *Preferences*. In Figure 4b the user has selected search and is in the process of entering the query term *murder* using the virtual keyboard that has panned up from the bottom of the screen. Pressing the search button on the keyboard produces a list of results, from which the user in Figure 4c has selected *Romeo and Juliet* to view the PDF version. Using two fingers touching the screen, starting pinched but then opening out causes the view to be zoomed. Reversing the gesture reduces the zoom level. Should the user tip the iPod on its side, using the sensor information available through the accelerometers on the device, the view automatically changes from portrait to landscape mode, a more convenient format in which to settled down and read the play.

In Figure 4d the user has selected the final tab at the bottom of the screen to review their preferences. From here the language interface can be changed from English to French, and control can be exerted over searching to include stemming and/or case-sensitive matching.



**Fig. 4.** A selection of screenshots from accessing the *Complete Works of Shakespeare* digital library collection on the iPod-touch

### 3.5 Inverting the Client-Server Configuration

While it is attractive to be able to put a large digital library in your pocket, a PMP's miniature screen is a clear disadvantage for many purposes. One solution would be to serve collections either to a single workstation by plugging it in directly (disk access), or else (if possible) serve it over a local area network. The natural way to accomplish this for Greenstone, for either configuration, is through a built-in web server, because this is how it serves its collections in the normal course or events.

Mobility of institutionally sized digital libraries, combined with the handheld device acting as a Web server has some interesting possibilities. For example, natural disasters such as earthquakes or hurricanes, and man-made ones such as terrorist attacks or nuclear accidents, demand immediate and informed response. They present an overwhelming need for information: information that is tailored for the problem at hand, organized so that it can be accessed effectively, and distributed even in the absence of an effective network infrastructure. Digital library technology combined with the self-sufficiency of a PMP offers one potential solution, allowing organized collections of information, graced with comprehensive searching and browsing capabilities, to be created rapidly.

In terms of implementation, supporting a web server configuration came as a *fait accompli* on the iPod-touch which, on a jail-broken device, comes with the Apache web server pre-installed. For the click-wheel the same was possible by porting a version of the Boa web server, a popular choice for embedded systems, served using IP over firewire. But the latter was not easy to set up, which limits its applicability at present—and comes with many caveats such as requiring a particular operating system configuration on the host machine.

A more practical solution for the click-wheel iPod was to make use of the disk mode feature of the device, and again exploit the fact that its disk capacity is substantial, placing pre-compiled web-server binaries (Apache) for popular *host* machine architectures (Linux, Windows, and Mac) on it. Next the iPod was plugged in to the host machine in disk mode, and from the host machine the directory with the web-server corresponding to its architecture started. The nett result is that a PMP that the user was using standalone with the DL capabilities described above just moments before, starts to serve the same content, via the host machine, to anyone on the local area. While deceptively simple, it works surprisingly well in practice, and is a technique equally applicable to the iPod touch, or any PMP device for that matter.

## 4 Conclusions

In motivating our work we talked of imagining a future when a complete digital library could be carried in the pocket. No requirement to access a network to retrieve content from a server; indeed, a future where the pocket device could act as a server of a vast amount of information. Our work has shown that this exciting possibility is now in reach.

A variety of digital library techniques have been trialled and tested on an iPod installed to run Linux (click-wheel) and BSD Unix (iPod-touch and iPhone). The open source Greenstone digital library software forms the basis of the work. The system is not intended as a finished end product, rather at this stage to be experimental—indicative of the sorts of problems that arise when using such a device for digital library use, and the sorts of solutions that can be devised. The choice of using Greenstone, therefore, was somewhat arbitrary, as was the decision to use the iPod.

As we have shown, there are not simply technical issues to be tackled but also questions of how best to design the user interface where the input and output facilities are impoverished, relative to conventional platforms. But, stepping up from the ‘interface’ level to that of ‘interaction’, these devices also challenge us to extend the way future users might perceive and experience digital libraries. Portable devices are often used to fill ‘dead-time’ when commuting, or when bored alone. Conversely they are used socially and collectively—people passing round their phones to show each other the photos and videos they have recorded earlier; listening to each other’s music players. Perhaps, soon, people will begin to similarly embed digital library use into their lives.

## References

1. Smeaton, A.F., Murphy, N., O’Conner, N., Marlow, S., Lee, H., McDonald, K., Browne, P., Ye, J.: The fishlar digital video system: a digital library of broadcast tv programmes. In: Proc. Joint Conf. on Digital Libraries, pp. 312–313. ACM Press, New York (2001)
2. Bainbridge, D., Jones, S., McIntosh, S., Witten, I., Jones, M.: Portable digital libraries on an ipod. In: Proc. of the Joint ACM/IEEE Conf. on Digital Libraries (to appear, 2008)
3. Buring, T., Reiterer, H.: Zuiscat: querying and visualizing information spaces on personal digital assistants. In: Proc. Int. Conf. on Human computer interaction with mobile devices and services, pp. 129–136. ACM Press, New York (2005)
4. Marsden, G., Cherry, R., Haeefe, A.: Small screen access to digital libraries. In: CHI 2002 Extended Abstracts on Human Factors in Computing Systems, pp. 786–787. ACM Press, New York (2002)
5. Witten, I., Bainbridge, D.: How to build a digital library. Morgan Kaufmann, San Francisco (2002)



# New Era New Development: An Overview of Digital Libraries in China

Guohui Li<sup>1</sup> and Michael Bailou Huang<sup>2</sup>

<sup>1</sup>The Graduate School Library, People's Bank of China  
Beijing, China

liguohui@gspbc.edu.cn

<sup>2</sup>Health Sciences Library, Stony Brook University  
Stony Brook, NY 11794-8034, USA

michael.b.huang@stonybrook.edu

**Abstract.** This paper presents an overview of research and development on digital libraries in China since 2002, introduces two completed digital library projects, and analyzes existing problems and countermeasures in Chinese digital library construction.

## 1 Introduction

Entering the twenty-first century, digital library research and construction in China have made considerable progress. A number of digital library projects have been completed, such as the National Digital Library of China and the Chinese National Science Digital Library. Before 2000, digital library research in China was at an elementary stage. Since the new century, the quantity of research has increased greatly and its scope has expanded. Research projects touch upon all aspects of the digital library domain and are more closely related to concrete digital library constructions. Research content is more in-depth and practical. At present, digital library service, technology, information resources, intellectual property, system structure, metadata, comparative studies, standards, and information safety are all hot research topics [1]. Among these topics, research on digital library service tops all other aspects, showing that with gradual maturity of digital library theory and technology, researchers pay more attention to user service. As a result of China's extensive history and culture, ancient books are also prominent in Chinese digital library research and construction.

## 2 Major Achievements of Chinese Digital Library Research and Construction

### 2.1 Strengthening the Protection of Intellectual Property

In recent years, the Chinese government has strengthened the protection of intellectual property. Related laws and regulations were made, for example, the Regulation on Copyright in the Production of Digitized Work, issued by the National Copyright Administration of the P. R. China on March 1, 2000, clearly stipulates the copyright protection in

relation to information resources digitization. Builders of digital libraries are aware of the laws and regulations of copyright and intellectual property. They need to verify intellectual property of related resources while digitizing a collection, sign digital utilization agreements, and limit the use of purchased database products only in their intranet. Copyright administration organizations of digitized work have been established to coordinate legal issues between the holder of copyright and the user [2]. In the aspect of technology, a great number of researches on prevention of utilization of paid information resources by illegal users have been done, such as encryption technology, electronic authentication, digital watermark, and certificate authority [3].

## **2.2 More Emphasis on Digital Library Technology Research**

In recent years, Chinese digital library research transformed from initial research on concepts to concrete implementation of technologies. The achievement of technology research laid a solid foundation for digital library construction.

At present, the content and quality of digital libraries is the most important part of Chinese digital library construction. Therefore, research on digital resource processing and editing technology is the most active. Great achievements have been made in text analyzing technique, classifying technique, network technology, isomeric data integrating technology, and data excavating technology. In addition, a number of breakthroughs were made in text retrieval technique, user recognition technology, Chinese metadata technology, intellectual search engine technology, Web 2.0, storage technique, digital watermark technology, and network safety.

## **2.3 Construction of Resources with Chinese Characteristics**

Chinese culture spans several thousand years of history. The country owns a tremendous amount of art and literature that is an essential part of the world's cultural treasure. In response to the need to preserve these cultural products, resource construction with distinctive national features, particularly in ancient Chinese book digitization, received more emphasis. Some specialized digital library projects were completed and are operational, such as the Oracle-Bone Inscriptions Database, International Dunhuang Project, and Chinese Rubbings Database.

## **2.4 More Emphasis on Resource Sharing**

Digital library construction is a social project. China is a developing country with limited financial resources. From the beginning of digital library construction, researchers and the government paid close attention to issues of how to avoid waste. Researchers not only paid a great deal attention to resource itself but library cooperation in resource development and utilization. Questions of how to make the most of cooperative effects and how to fully explore and utilize every library's resources to attract more and more interest and research were posed. Led by the Chinese government, alliances and consortia of various levels and scopes were established, such as the China Academic Library & Information System (CALIS) and the Beijing Academic Library & Information System (BALIS). These alliances congregated state investment, modern library concepts, advanced technology, and the rich literature resources of academic libraries to achieve the goal of mutual construction and sharing.

## **2.5 International Cooperation**

Since 2001, a number of international conferences on digital libraries have been held in China: the 12<sup>th</sup> International Conference on New Information Technology (2001), the International Forum on Digital Libraries (2002), the 7<sup>th</sup> International Conference of Asian Digital Libraries (2004), the Chinese-European Workshop on Digital Preservation (2004), the 1<sup>st</sup> International Conference on Universal Digital Library (2005), the IFLA Pre-Conference in Hangzhou (2006), and the Shanghai International Library Forum. These conferences provide golden opportunities for international exchange and cooperation. International cooperative projects such as the China-US Million Book Digital Library Project and International Dunhuang Project have all achieved good progress.

## **3 Brief Introduction of Two Notable Digital Library Projects**

### **3.1 Chinese National Science Digital Library (CSDL)**

The CSDL is a national-level digital library project built by the Chinese Academy of Sciences. It began at the end of 2001 and took five years to build with a total cost of 1.4 billion Chinese Yuan. The CSDL maintains a collection of electronic journals in Chinese and many foreign languages, conference proceedings, dissertations, theses, patents, science citation index, subject information portal, and supports access to over 100 primarily web-based electronic science research databases. Additionally, the library offers cross-database searching and browsing, interlibrary loan, document delivery, and online reference service. Using a unified online catalog, the user can search a collection of over 400 libraries.

### **3.2 International Dunhuang Project (IDP)**

The IDP is the result of a successful collaboration between the British Library, the National Library of China, and several other libraries. The IDP was founded in 1994 to create a virtual library bringing all of the Dunhuang documents together in a high-quality digital format. An integrated Chinese website was launched on November 11, 2002 in the National Library of China. So far, over 100,000 manuscripts, paintings, and archaeological artifacts previously hidden for over a thousand years in Silk Road caves are now digitized and made available on the internet for free use.

## **4 Existing Problems and Potential Solutions**

### **4.1 The Scope of Resource Sharing Is Not Wide Enough**

Resources construction is the core of Chinese digital library construction. Effective utilization and sharing information resources is one of the main purposes of digital libraries. However, libraries in China are managed by different administrative units and have different financial sources. The situation has created isolated digital library constructions and formed many different library communities. As a result of this kind of isolation, coordination among different digital library constructions becomes difficult.

At present, library organizations in different communities have begun collaboration in technology, standards, information resource, and service. Their successful results prove that this kind of openness and collaboration are effective. Nevertheless, digital library resource sharing is an integral systemic project. Should the government centralize planning, organization, and collaboration, and break the lines that divide each library circle, resource sharing of a much larger scale would be realized.

## **4.2 Standardization Is Not Popularized Enough**

Many libraries in China own special collections and special databases. During the digitizing process, different technology is adopted in accordance with different resource content. Formatting and processing is also different. Thus, different standards for digitization are created. The difference in standardization hinders digital library resource sharing, literature searching and retrieving by the user, and creates chaos in application of standards. A solution to this problem would be to increase communication and cooperation. To popularize unified standards, the Chinese government, library associations, and consortia should increase such communication and collaboration.

## **4.3 The Gap between Rich and Poor Is Growing Wider and Wider**

The state of Chinese digital library construction is influenced by the strength of the building libraries. Digital library construction is more developed and rapid for those with a bigger budget, advanced technology, and a faster network. There is a gap between the richer Eastern regions and the poorer Western regions. A gap exists between top universities and average universities. The gap has now become wider and wider. Since the majority of universities in China are funded by the government, this problem should be solved in terms of national policy. While supporting a few key universities, medium-to-small schools should not be ignored. Those universities receiving the bulk of government funding should be more responsible in assisting other schools in information resources, technology, and staff training.

## **4.4 A Need for Faster Networks**

At present, most Chinese users log onto the web using 1 MB ADSL. The speed is slow compared to high-speed cable broadband internet with 10 MB or several GB. ADSL is satisfactory for general text browsing but to use multi-media or digital voice, this type of speed is unbearable. Increasing the internet speed has always been a priority.

Both China Netcom and China Telecom, the two largest telecommunication companies in China, are experimenting with Fiber-to-the-Home (FTTH) service in order to increase Internet speed. China Netcom has established fiber networks in several residential areas in Beijing with 2.5 GB broadband serving 15,000 families. China Telecom provides the end-user with digital voice, high-speed cable broadband internet, and CATV. It will take time to spread FTTH service in China due to the cost. The average installation fee is 2,000 Yuan. Most Chinese consumers can accept the price if it is reduced below 1,000 Yuan. However, such a low fee will not sustain FTTH

network operation and maintenance. Increasing network speed depends on new telecommunication technology and a lower service fee.

## 5 Conclusions

The rapid development of Chinese digital library in the new century is closely related to development of China's economy and technology as well as wide-ranging international cooperation. With the support of China's rich culture, the dynamic construction and development of digital libraries in China will contribute greatly to digital library development of the entire world.

## References

1. Zhang, Y.: Collaborative Digital Reference Service of Interregional University Libraries. *Library Work and Study* 3, 97–98, 109 (2008)
2. Sai, S.: Intellectual Property Protection in the Construction of Digital Libraries. *Inner Mongolia Library Work* 3, 36–38 (2006)
3. Xu, M.: Intellectual Property Issues Related to Information Resources Construction of Digital Libraries. *Modern Information* 3, 83–85 (2007)

# Browse&Read Picture Books in a Group on a Digital Table

Jia Liu, Keizo Sato, Makoto Nakashima, and Tetsuro Ito

Oita University, 700 Dannoharu, Oita-shi, Oita 870-1192, Japan  
{wenyao,k-sato,nakasima,ito}@csis.oita-u.ac.jp

**Abstract.** Group reading plays an important role for children. We here formalize a new framework to enhance children's group reading by digitizing the printed picture books and placing a digital table in a physical library. Children around a digital table can easily browse in a large virtual bookshelf to find the desired printed and/or digitized picture books, exchange the opinions about these books while reading, and hand these books over freely among themselves. The benefits of the formulated framework were shown by two usability case studies.

**Keywords:** Group reading, group browsing, picture books, digital table, children's reading room, children's digital library, physical library.

## 1 Introduction

Reading picture books in a group makes children become active members in a reading community. Many ideas exist about children's group reading [1][2][5]. The activity in UK [5], however, pays little attention for children to browse in a bookshelf collaboratively to find the books of their own interests. Also the books that the group finds cannot be read when lent out. Another activity in Baltimore University [2] resolves the lent out problem, but it cannot have the benefit of face-to-face communication while browsing and reading.

For group activities the communication through gesture/chatter and characters/figures is important. Children's group reading would be better performed in an environment where they can exchange opinions and can hand the interesting books over to each other. Here we formalize a new framework that can *enhance children's group reading in a physical library* equipped with a digital table. Sitting around the digital table, children can browse and read (browse&read) both printed and digitized picture books together with their friends and/or families.

In an ideal case, a physical library has all its printed picture books digitized, in addition to the ordinary digital books. The framework here is implemented by placing in the children's reading room our digital table installed with two software packages with the functions: (i) allowing children to browse in a large virtual bookshelf and read the found digitized picture books together with printed ones; (ii) materializing a virtual bookshelf and a digitized picture book so that they can be handled as if they were the real physical objects on the table.

The framework was evaluated through two usability case studies in a children's event and in a reading-to-children program. The digital table was shown to provide a very helpful means for enhancing group reading in a physical library.

## 2 Enhancing Group Reading in a Physical Library

Fig. 1 illustrates the basic concept of our framework. A digital table has a tabletop touch display with a PC installing BrowsReader [3] and the tray system [4]. The digitized picture books are also stored in the PC. The major roles of the digital table are for the children to handle easily the materialized virtual bookshelf and digitized picture books, and to enforce the face-to-face communication among the reading group.

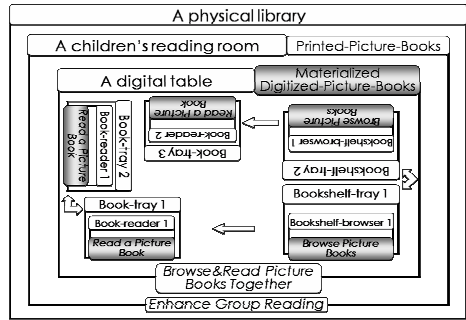


Fig. 1. Enhancing children's group reading

BrowsReader consists of a bookshelf-browser coupled with a book-reader.

The bookshelf-browser allows children to narrow down a large index of the titles/author-names by inputting characters or a large virtual bookshelf by touching a part of the bookshelf, and the book-reader to read the found books. Two types of trays are designed; a bookshelf-tray and a book-tray on which a bookshelf-browser and a book-reader are loaded, respectively. The images of the bookshelf-browser and the book-reader on the trays can be freely moved and rotated as seen in Fig. 1, by which we say the virtual bookshelf and the digitized picture book are materialized. Any tray can have as many clones as needed.

We foresee a new reading environment realized by simply replacing some wooden tables with digital tables would enhance children's group reading. The benefits can be summarized below as:

- (1) Children can browse in a large bookshelf to find the desired picture books.
- (2) Children can exchange information about books through gesture/chatter and hand the materialized or printed picture books over freely to each other.

## 3 Realization and Utilization of a Digital Table

Fig. 2 shows a photo of a digital table around which children can browse&read digitized and printed picture books. This table is the same as an ordinary reading table except the top 30-inch touch display controlled by a personal computer. In this figure two bookshelf-trays and two book-trays loaded with a bookshelf-browser and a book-reader of BrowsReader [3], respectively, are displayed. One of the two bookshelf-trays (and book-trays) is the clone.



Fig. 2. A digital table

The bookshelf-trays and the book-trays are implemented by employing the tray system technologies [4]. We load them with a bookshelf-browser and a book-reader, respectively. Children can touch the bookshelf-browsers and book-readers anywhere, after displacing and rotating, around the digital table as they like. The cloned bookshelf-tray shown in the upper-right in Fig. 2 is reduced due to the limitation of the display size. The original bookshelf-tray is displayed in full-screen, so it can be seen as large as possible. Children can collaboratively input a search character one after the other from either tray. A book-tray is generated every when a child touches a book cover image in the bookshelf.

To see how collaborative reading can be enhanced by the digital table, let's consider the scenario when a young boy inputs ‘さ’ by dragging a character on a bookshelf-tray. His sister can assist him in adding ‘う’ just before ‘さ’ by using her bookshelf-tray. When they find a book about “うさぎ(rabbit),” this can be loaded on a book-tray. He can read it by himself, or with his sister sitting at side and helping him flipping the pages.

## 4 Usability Case Studies

We conducted two usability case studies to evaluate the formulated framework.

(*Case 1: In the children's event*) A temporary children's reading room was established in the playground of the 2008 Children's Event at Oita University. Two digital tables with 30-inch touch displays were set. BrowsReader was installed on PCs (The tray system was not employed). The collection consisted of the digitized versions of 232 printed picture books and 228 web picture books. Total 146 children (aged 2-13) together with their friends and/or families took part in the study from 10:00 to 15:00. A questionnaire for each group was prepared.

We had conducted a similar event in 2007, where the number of participated children (aged 2-13) was 111. The major differences were that much more digitized picture books were prepared in event 2008, and a 30-inch display was laid down horizontally in 2008 rather than kept standing vertically. We expected the group reading could be enhanced by laying down the display.

**Table 1.** Percentages of the positive answers in the events 2007 and 2008

No.	Question	2008			2007		
		kinder	lower	higher	kinder	lower	higher
Q1	Are you happy when you used it?	96.3	99.0	100	100	98.5	100
Q2	Did you read a book when you used it?	77.8	88.9	95.0	81.3	89.7	96.3
Q3	Did you use it with others together?	81.5	61.6	45.0	81.3	32.4	29.6
Q4	Would you like to use it again?	88.9	96.0	100	93.8	94.1	100

The statistics for the questionnaires in 2008 and those in 2007 are summarized in Table.1 (kindergarteners for ages 2-5, lower graders for ages 6-9, and higher graders for ages 10-13). The results were almost the same except for Q3 in lower and higher graders. The percentage of the positive answers was highly increased. This agrees with our expectation. The answers for Q1, Q2, and Q4 were pretty stable, indicating



the overall usability remains the same. During this study some pairs often dragged characters or flipped pages in turn, and many gathered together.

(*Case 2: In the reading-to-children program*) The same digital table used in event 2008 was settled in the children's reading room in the Oita University library. This time a bookshelf-browser and a book-reader of BrowsReader were loaded on the trays. We expected that this table setting can further enhance group reading in a reading room. During the 90-minute long reading-to-children program, 5 children (aged 4-7) with their parents and two volunteer librarians participated in the 45-minute study. Every child showed great interest in the digital table, and browsed&read digitized books together with others. A video recorder was set to log the children's reading activities. The parents and volunteer librarians were also interviewed after the study.

By skimming the video, we found that total 9 picture books were read by pairs of the children together with the parents or the volunteer librarians, and 6 books were handed over among the children. There was a scene in which three children and a volunteer librarian gathering around a corner of the digital table read a picture book about “うさぎ(rabbit),” after cloning a book-tray. This book was found, using a bookshelf-tray, by one of the children with the assistance of the librarian for inputting characters. Between this child and the volunteer librarian a younger child joined in the group reading. While the book was read out by the volunteer librarian, two children followed the pages and sometimes flipped pages by themselves. The taped interview includes such voices: “This time my children read more books together than usual,” and “This is almost perfect for group reading, especially for our reading-to-children program.”

## 5 Conclusions

A framework featuring a digital table placed in a physical library to better assist children in browsing&reading picture books in a group is formulated. The ultimate goal of our research is to have digital tables set in the places with real printed picture books, such as pre-schools, community centres and public libraries, and become an indispensable part of these places. We are planning to carry out usability studies in these places.

## References

1. Children's Reading Group Ideas summarized by EncompassCulture, <http://www.encompassculture.com/readinggroups/readinggroupideas/childrensrggideas/>
2. Colorado, B.: Reading Alone Together: Creating Sociable Digital Library Books. In: Proc. of IDC 2005, pp. 88–94. ACM Press, New York (2005)
3. Liu, J., Nakashima, M., Ito, T.: BrowsReader: A System for Realizing a New Children's Reading Environment in a Library. In: Goh, D.H.-L., Cao, T.H., Sølvyberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 361–371. Springer, Heidelberg (2007)
4. Nakashima, M., Sato, K., Ito, T., et al.: Strengthening Interaction on a Tabletop Display with Two-Module Trays by Considering the Levels of Collaborative Work (in Japanese). In: Proc. of IPSJ Interaction 2008, Tokyo, vol. 4, pp. 145–146 (2008)
5. The CILIP Carnegie and Kate Greenaway Children's Book Awards, <http://www.carnegiegreenaway.org.uk/home/index.php>

# Towards a Webpage-Based Bibliographic Manager

Dinh-Trung Dang, Yee Fan Tan, and Min-Yen Kan

Department of Computer Science, School of Computing,  
National University of Singapore, Singapore, 117590  
{dangdinh, tanyeeefa, kanmy}@comp.nus.edu.sg

**Abstract.** We present ForeCiteNote, an application that organizes personal digital collections of research articles. It is architected as a single HTML page with embedded Javascript that runs within a web browser. On top of standard annotation and tagging functionality, it also supports both online and offline usage patterns, including local storage of the paper collection.

## 1 Introduction

Scholarly information has greatly expanded in recent years, and while the exact balance of causes responsible for this are difficult to pinpoint, certainly open-access digital libraries and timely dissemination play a part [1]. While such free flow of information presents an unparalleled learning opportunity, it also presents challenges in managing these information sources, especially for beginning researchers.

To focus on the needs of researchers, we performed a focus group study at two academic institutions, involving 12 participants from the French National Institute for Research in Computer Science and Control (INRIA) and 7 from the National University of Singapore (NUS). Each participant completed ten questions on how they performed their own bibliographic management. The findings showed that around 60% store their papers both on- and off-line and have some difficulty organizing them. 43% percent reported using offline applications to take notes on research papers. Over 60% usually access their data when they are away, and do not view their own method as convenient nor systematic. 83% of those surveyed manually enter all the metadata (e.g., author, date, conference) when annotating papers. A good *bibliographic management application* (BMA) is thus needed for organizing documents: allowing researchers to collect, retrieve, annotate and organize citations and digital copies of scholarly articles.

## 2 Related Work

BMAs are not a new topic; in fact there are many commercial systems to cover such needs. A comprehensive survey of all solutions is beyond the scope of this paper. We focus on a few current major systems that represent the spectrum available. End-Note (<http://www.endnote.com/>) is the canonical example of the standalone BMA. It features comprehensive import and export functionality, organization of digital files, and word processing program integration. On the other hand, web-based applications such as CiteULike (<http://www.citeulike.org/>) [2] and BibSonomy

(<http://www.bibsonomy.org/>) [3] have central servers housing users' bibliographies, thereby promoting sharing and collaborative annotation and tagging among users. Connotea (<http://www.connotea.org/>) [4] also facilitates offsite web use through browser extensions that can send annotations and references back to the central server. Zotero (<http://www.zotero.org/>) takes this idea further by embedding the whole BMA within a Firefox extension, eschewing a server based solution entirely.

Each architecture has distinct advantages and compromises. A standalone application offers the maximum flexibility but requires the user to run a separate application and is prone to data loss (if the user's hard drive crashes). Server based BMAs make the provider responsible for data integrity and immediately provide collaborative possibilities. However, this architecture cannot manage a user's local digital collection and requires access to the web at all times. Browser extension based BMAs play well both on- and off-line, but increase the memory footprint of the browser and also have the problem of data loss. Also, as the system is tied to a specific web browser instance, their library does not port with them if the user changes browser or computer.

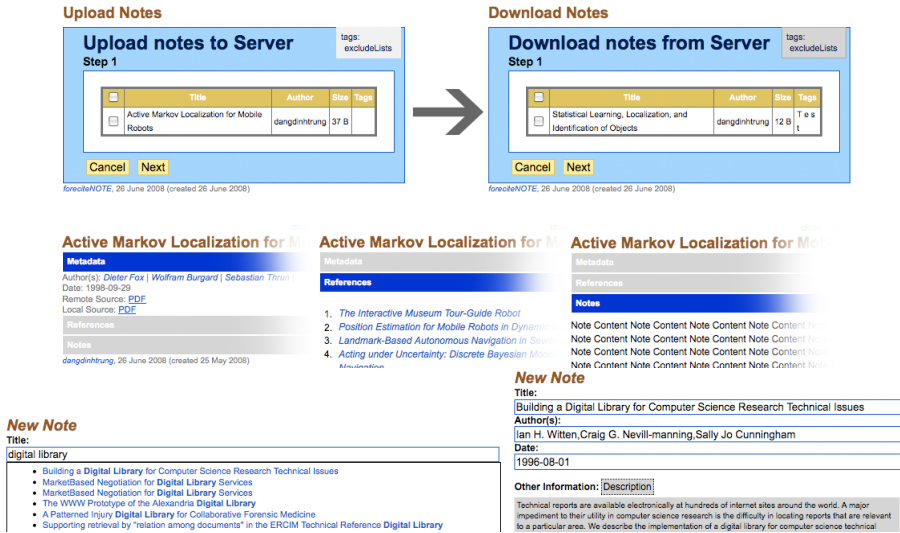
From the above architectural discussions, four central issues emerge: 1) accessibility both on- and off-line, 2) portability across machines, 3) preventing data loss, and 4) local file management. ForeCiteNote aims to solve these issues, by basing its bibliographic management on a different technology: the Javascript-enabled local web page.

### 3 ForeCiteNote: A TiddlyWiki-Based Bibliographic Manager

*ForeCiteNote* is a client-side solution, with a supporting server component, that addresses the above issues. The server component provides data integrity by allowing users to synchronize or backup their annotations and references. It also features online access to the user's data and is capable of adding references and annotations, similar to other server-based BMAs. The key difference is in *ForeCiteNote*'s client-side component, which is a modified version of a TiddlyWiki.

TiddlyWiki (<http://www.tiddlywiki.com/>) is a self-contained HTML page which uses embedded Javascript to give the page the functionality to act as a wiki or a blog. A user can add entries to keep notes or progress logs to the TiddlyWiki in entries known as *tiddlers*. As it is a single, standard HTML file, it can be uploaded to a web server so that it can be accessed anywhere with an internet connection, and downloaded to other PCs, resolving Issue 1. When used off-line (locally on the user's computer), the permissions afforded by the `file:` protocol allows new and modified tiddlers to be saved by overwriting the TiddlyWiki itself on the user's disk. Since TiddlyWiki is just a single, small (~400KB) web page and requires no installation, users can store their own TiddlyWiki on a local USB drive and carry it around to view on any offline computer equipped with web browser. This makes TiddlyWiki eminently portable and thus resolves Issue 2. To resolve the remaining two issues, we had to modify the basic TiddlyWiki, resulting in the final *ForeCiteNote* client, which we describe next.

To solve Issue 3, *ForeCiteNote* features a synchronization feature that allows the client *ForeCiteNote* to send its annotations and entries to a *ForeCiteNote* server. The synchronization engine first allows the upload any entries to the server and then allows



**Fig. 1.** Top row: Synchronization wizard. Middle row: New display format for tiddlers. Bottom row: Suggestion lists (left) and metadata autocompletion (right).

the user to download any missing entries (Figure 1, top row) illustrates this. In the drastic case where the user has lost their client-side web page, one obtains an empty client and synchronizes it to download all of their entries.

To solve Issue 4, the ForeCiteNote client also manages a directory where it stores local documents (named *paperlib*). As users may want to browse the local collection outside of the ForeCiteNote webpage application, care was taken to organize local documents in a useful and domain-independent way. According to [5], researchers identified the author, year and title as the three most prominent metadata that identify a scholarly document. ForeCiteNote structures its directory accordingly, in an iTunes-like directory structure: where the author and year are the first and second respective sub-directory level structures within the local document directory, and where the document itself is renamed after the title of the document.

For documents that the ForeCiteNote server knows the metadata and a location of an open access copy, the user can use the client to download and store a copy into the local document directory. The user can similarly ask the ForeCiteNote client to store a copy of other documents that the user has a copy, but which is unknown to the server. Sharing a document collection is as simple as sending the directory tree to a collaborator, which can be done by using archiving utilities and email. Also, since there are many desktop search applications, we intentionally left out local collection search functionality from the client.

We also customized ForeCiteNote to improve upon its usability for recording research notes. First, we modified ForeCiteNote's tiddler format, to better display and alternate between displaying a work's metadata or its annotations in one place, as in Figure 1 (middle row).

Second, in our focus groups, metadata entry was identified as a severe bottleneck in annotation. Users clearly wanted references chunked into specific fields (e.g., author, title, year) but did not want to manually enter all of these fields. When the client has internet access, it can query the server to retrieve pertinent metadata to perform metadata autocompletion. When the user first creates a new entry for a paper, they start by typing in the document title. After three letters have been typed in, the client queries the server for any autocompletion matches based on papers known to the system. These are shown to the user, as in Figure 11 (bottom row, left). The user can then either ignore the suggested matching list or pick the document's title if shown. If a suggested known paper is chosen, the client retrieves the metadata from the server, including the abstract and any open-access location, and fills in the appropriate fields in the entry, as in Figure 12 (bottom row, right).

## 4 Conclusion

We have introduced ForeCiteNote, a webpage-based bibliographic manager. As an adaptation of the revolutionary TiddlyWiki software, ForeCiteNote not only allows researchers to collect, annotate and organize citations and digital copies of scholarly articles and also a means to retrieve them, but also guarantees the accessibility both online and offline and portability across machines.

ForeCiteNote has undergone several rounds of informal user design testing and has incrementally been improved to its current state, reported here. Current work focuses on a long-term, longitudinal assessment of ForeCiteNote's usability. We expect to release the application for beta testing soon. Future work includes extending the ForeCiteNote server functionality to support the collaborative work of distributed groups.

## References

1. Prosser, D.C.: Scholarly communication in the 21st century - the impact of new technologies and models. *J. Serials Community* 16, 163–167 (2003)
2. Emamy, K., Cameron, R.: CiteULike: A researcher's social bookmarking service. *Ariadne* (2007)
3. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Bibsonomy: A social bookmark and publication sharing system. In: *Conceptual Structures Tool Interoperability Workshop at 14th ICCS Conf.*, pp. 87–102 (2006)
4. Anonymous: Join a social revolution. *Nature* 436, 1066 (2005)
5. Day, M.Y., Tsai, T.H., Sung, C.L., Lee, C.W., Wu, S.H., Ong, C.S., Hsu, W.L.: A knowledge-based approach to citation extraction. In: *IRI*, pp. 50–55 (2005)

# Spacio-Temporal Analysis Using the Web Archive System Based on Ajax

Suguru Yoshioka, Masumi Morii, Shintaro Matsushima, and Seiichi Tani

College of Humanities and Sciences  
Nihon University

{s-yoshio,m-morii}@chs.nihon-u.ac.jp  
{matsushima,sei-ichi}@tani.cs.chs.nihon-u.ac.jp

**Abstract.** Ajax is a new technology based on asynchronous communication between a client Web browser and a back-end server, allowing Web applications to request data without Web application to request and receive data ever reloading the page. We designed and implemented the Ajax Web archive application which supports to discover and analyze spacio-temporal structure of a sequence of sentences.

**Keywords:** Ajax, Digital library, Human computer interaction, Theatrical analysis, Web archive.

## 1 Introduction

In recent years, research for digital library has been widely applied to various disciplines [5]. Within the development and spread of information technology and computer networks, we need to note effective usage of analysis resources and discover advanced forms of utilization. For such occasions, Ajax technology is now attracting a lot of attention [1,2,7]. Ajax is an abbreviation of “Asynchronous JavaScript and XML.” Thus, Ajax is an asynchronous communicating system in Web browser, and it provides a Web interface application to extend a visual interaction. For example, the Google Maps application downloads new data in accordance with the movement of the mouse in background, i.e. the web page is not reloaded. Such display method without a page transition is effective to encourage the interest of general users. So, we designed and implemented the Web application system<sup>1</sup> with Google Maps API to dynamically operate a geographical contents. As the target of this analysis, we consider Kitamura Rokuro’s diary. His diary is a typical example that prescribes the spacio-temporal relations of affairs. Our system includes visually-enhanced interface control module; general users can visually check the data of the geophysical makeup of descriptions of Kitamura Rokuro’s diary. We dared to display the description of the forepassed diary on the current map.

---

<sup>1</sup> <http://133.43.160.36/Kitamura/Public/public.html>

## 2 Kitamura Rokuro's Diary and Geographical Information

Many researchers have developed digital library systems for theatrical materials that replace paper-based information database systems<sup>[5]</sup>. This paper focuses on Kitamura-Rokuro library<sup>[2]</sup>. Kitamura Rokuro is one of most popular actors from the end of the Meiji era to the beginning of the Showa era viz. from the late 1800s to the early 1900s<sup>[4]</sup>. He is well known as a contributor to the development of Shinpa, often translated as “New School Theater” refers to a theatrical genre born in Meiji Japan (1868-1912). So, his diary serves as a useful reference for theatrical research and Japanese cultural study. But today, in recent decades after the diary was written, the face of the street has certainly changed. And so, it is difficult for us to locate some store, some building and so on in his diary.

So, we implemented the interface system to archive the Kitamura Rokuro's diary. We show our system helps us to create a digital graphical library.

## 3 System Configuration

In this section, we show our Web application system. Our purpose is to raise the visualization environment for general users and the operability for administrators. We show the public Web pages and the administrator Web pages for general users and experts, respectively.

### 3.1 The Public Web Pages

Our system build multilayered representation for a temporal axis and a spatial awareness. For example, for the diary of May 29th, 1931, we focus the following descriptions (landmarks): Alasuka, Kunieda, Yamanaka, Mitsukoshi, Takashimaya, Nihonbashi, Sennichi and Shinmachi. Our system draws the moving route from Alasuka to Shinmachi as lines on Google Maps. We show the web page for the above situation in the left-hand side of Figure 1. In the left-hand side of Figure 1, the text of the diary is displayed on the upper right box of this page. And, the Google Maps appears at the lower of this page; our system puts reference markers on the map, which correspond the distinctive descriptions in the above body text. Users will find on each marker a number corresponding to the appearance sequence of the distinctive descriptions. And, our system draws red lines in order of increasing maker number. Thus, users can look see his gait by the text body and map. In the text of the diary, some descriptions are linked to a web page or a marker on the map. For example, in the left-hand side of Figure 1, Alasuka, Kunieda, Yamanaka and so on are established a link. If an user click on the link, the corresponding marker is focused in the center of the map. In addition, because of markers on the map are added a link to some information and pictures, users can click on a marker to enlarge an image of the

---

<sup>2</sup> “Academic Frontier” Project for Private Universities: matching fund subsidy from Ministry of Education, Culture, Sports, Science and Technology, 2002-2007.

face of the old street. In the right-hand side of Figure 1, the picture of Alasuka in those days was enlarged. Thus, we can recognize the spatial relationships between the current location and the objects of the past by preparing precious pictures and information in advance.

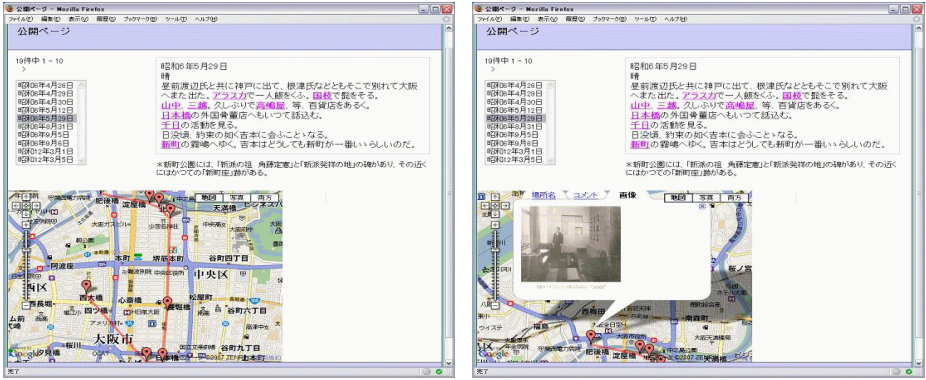


Fig. 1. Map for Kitamura Rokuro’s diary

### 3.2 The Administration Web Page for Data Management

In this section, we show the creation and the management of database. Our system utilizes the internal format of data named *diary data* and manages it. The *diary data* consists of the tuple  $\langle \text{body text, reference information} \rangle$ . And the *reference information* consists of the tuple  $\langle \text{expository text, terminological list, geographical information} \rangle$ .

Here, we show the procedure for a new registration of the geographical information. On the administrator page, we can add new data for the location, comments and pictures to our database. On the page, we can pick out registered pictures from a preview window and configure these pictures as enlarged pictures on the Google Maps without the screen transition. Hence, users need to specify the location of marker by clicking on the Maps or dragging the marker with the mouse over the map. But then, the address retrieval function is available to the user. And, if users wish to link to the reference information in text body of diary, they enclose link items in caret ^ for the documents in the diary text box or the expository text box, and then they select the registered reference information for these items.

### 3.3 Logical Foundation of Our System

We can view landmarks in the diary mapping onto the Google Maps. That is, we regard the description in the diary as a semantical interpretation, because of a marker corresponds to a description in the diary. So, for a set of markers  $M (= \{m_1, m_2, \dots, m_n\})$ ,  $m_1 \models d_1, m_2 \models d_2, \dots, m_n \models d_n$ , where



$D (= \{d_1, d_2, \dots, d_n\})$  denotes a set of landmarks (words) in the diary. And, for  $m_i, m_j \in M, \forall i \exists j. C(m_i, m_j)$ , where  $C$  is a predicate which denotes a connection with line. This  $C$  does not expressly appear in the diary, however, we can regard it as action, information flow, or temporal aspect. If we define  $M$  and  $C$  as the situation and the temporal aspect, respectively, this model has completeness [3]. And  $C$  has transitive reflexive closure [6]. In fact, temporal relation in the diary is consistent because the diary is a typical example that prescribes the spacio-temporal relations of affairs. Here, we infer this relation is introduced into  $D$ . We show this relation as follow:

$$\begin{array}{ccc} d_l & \xrightarrow{(-C)} & d_m \\ \parallel & & \parallel \\ m_i & \xrightarrow{C} & m_j \end{array}$$

Where  $d_l, d_m \in D$  and  $m_i, m_j \in M$ . Thus, our system could clarify the implicit cognition in the diary as an explicit line on the map.

## 4 Conclusion

In our system, users can successively handle an addition, a change and a deletion of data by using Google Maps API and Ajax. Especially, we attempted to draw descriptions in the diary as graphical representation. As a result, a spatial aspect was appended to the text data, and readers were interested in the diary as a diversity-carrying image. For a specific example, we could confirm the moving route on the map. We anticipate the development of our system as an user participation system, i.e. Web 2.0 system.

As a future subject, we need to apply our system to other documents. And we anticipate to find out the new information by comparing the several geographical information.

## References

1. Hanakawa, N.: An intelligent web browser plug-in for automatic translation to Ajax approach. The International Journal The IPSI BgD Transactions on Internet Research 3(2), 23–31 (2007)
2. Paulson, L.D.: Building rich web application with Ajax. IEEE Computer 38(10), 14–17 (2005)
3. Shoham, Y.: Reasoning about Change. MIT Press, Cambridge (1988)
4. Yanagi, E.: Shinpa no 60nen, Kawade shobo (in Japanese) (1948)
5. Yoshioka, S., Morii, M., Tani, S., Kensuke, K., Toda, S.: Temporal Reasoning System for the Digital Theater Library. In: Proceeding of the IMSA 2007, vol. 577, pp. 165–171 (2007)
6. Yoshioka, S., Tojo, S.: Many-dimensional Modal Logic of Tense and Temporal Interval and its Decidability. Journal of the Japanese Society for Artificial Intelligence 21(3), 257–265 (2006) (in Japanese)
7. <http://maps.google.com/>

# Mining a Web2.0 Service for the Discovery of Semantically Similar Terms: A Case Study with Del.icio.us

Kwan Yi

School of Library and Information Science, University of Kentucky  
Lexington, Kentucky 40506 USA  
kwan.yi@uky.edu

**Abstract.** This study develops and implements methods of identifying similar terms using collaboratively constructed folksonomies. In this study, two folksonomy-based methods are proposed with an aim of demonstrating the usefulness of folksonomy as a source for the discovery of similar terms, especially for 'non-in-the-dictionary' terms: co-occurrence-based and correlation-based methods. The experimental results show that the co-occurrence-based method performs comparatively better and that the folksonomies have a potential as a source for the discovery of similar or near-similar terms. The result implies that as the web2.0 service for the folksonomies evolves, the potential of folksonomy for the task will be increased.

## 1 Introduction and Previous Research

People tend to use different terms to describe the same or similar concept or object. This term variation issue has been an inherent barrier in human-computer interaction [1]. In the broad areas of natural language processing and information retrieval, term discrepancies have long been a primary challenge.

A conventional approach for alleviating the issue is the automatic identification of semantically similar or related terms. Two different kinds of data sources have been primarily utilized in the previous studies: machine-readable dictionaries or thesauri [2, 3] and corpora of documents such as web documents [4, 5]. In the dictionary-based study, it is assumed that synonyms or similar terms of a term co-occur in its definition in dictionaries. The relationships between terms and definitions are represented by a graph (dictionary graph), and various methods applied to the graph (web graph) of a web (Kleinberg's hubs and authorities method [6]; a variation of PageRank [7]) are employed [8, 9]. Similarly, the underlying assumption of the corpora-based study lies in that similar terms co-occur in similar documents. Traditional information retrieval techniques such as vector space model, cosine similarity measure, and term discrimination values such as mutual information were applied in this context [10]. In Web document corpora, to assess the similarity of terms, term presence and absence, co-occurrence of terms and distance between terms is measured, in conjunction with a combination of classic Boolean operators (AND, OR, NOT) and advanced operators such as NEAR [11, 12].

In this study, a popular Web 2.0 social bookmarking website, del.icio.us, is investigated and tested as a new type of potential data source for the automatic similar term discovery. Social bookmarking systems have quickly become popular digital information repositories built with the collaboration of all participants. Particularly, del.icio.us had 3 million registered users and 100 million unique URLs bookmarked as of September 2007 that had been more than tripled in the past 12 months [13]. Participants of social bookmarking systems annotate digital resources relating to their interests, primarily webpages, by assigning keywords freely chosen by the participants to the resources for the purpose of accessing them later with those assigned keywords. We define folksonomy as a collective set of keywords used by participants in the systems. With the systems being used, more keywords can be assigned by more people to the same digital resource. The characteristics of folksonomy can be summarized: (1) Folksonomy is a user-created vocabulary – by information users or consumers, not by information professionals or providers; (2) folksonomy is created in a collaborative manner – the size of the folksonomy quickly grows as more participate; (3) A list of terms (folksonomy) assigned to digital resources is known. The aforementioned features were unique to the Web 2.0 driven data sources, as opposed to the conventional data sets. The Web2.0 features motivated the hypothesis of this study that a group of semantically related or similar terms can be found in folksonomy. With the objective of this study to mine similar terms from folksonomy, del.icio.us (<http://del.icio.us>) is selected as the Web2.0 service for this study, primarily due to its high popularity in both practice and research. The following research question will be addressed in this study: Can the folksonomy be an effective source for the automatic discovery of similar terms, especially for technical terms, relatively new terms, and ‘not-in-a-dictionary’ terms?

## 2 Folksonomy in Del.icio.us

A tagging activity can be more formally defined as a tuple of (R, P, T), where a participant P annotates a resource R and assigns tag(s) T to the resource. A participant is allowed to assign multiple tags to a single web resource, and different people are involved in tagging a same resource. Two different types of folksonomy is available in del.icio.us: Global-level and resource-level folksonomies. At a global level folksonomy shown in figure 1, a number of the most popularly used tags on Del.icio.us are listed. The tags in the folksonomy are displayed in the order of popularity which is depicted by the tag size – the more larger the more popular. The *design*, *blog*, and *webdesign*, were the most popular three tags at the time of the screenshot. At the resource level shown in figure 2, the entire set of tags assigned to a resource is ranked according to its frequency, at the time the screenshot was taken. The figure shows that the resource titled ‘Folksonomy – Wikipedia, the free encyclopedia’ is tagged by 1741 participants, and the word ‘Folksonomy’ was the most commonly assigned tag for the resource by 291 people, and then ‘tagging’ and ‘web2.0’ by 181 and 175 people, respectively.

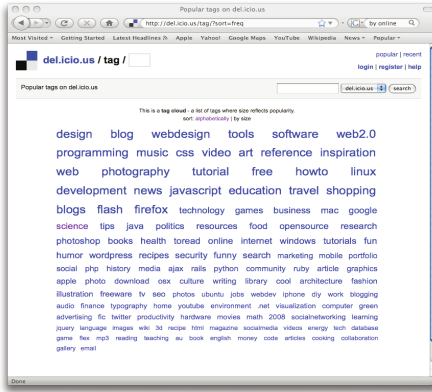


Fig. 1. Example of a global level of folksonomy

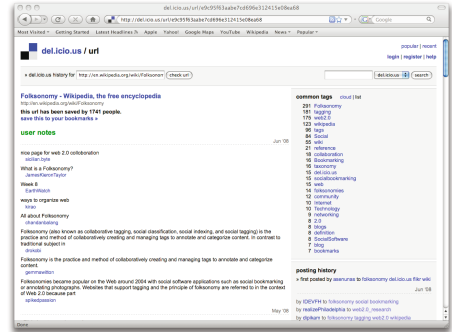


Fig. 2. Example of a resource level of folksonomy

### 3 Finding Similar Terms from Folksonomy

We propose two methods of identifying similar terms using folksonomies in del.icio.us.

#### 3.1 Co-occurrence-Based Similarity Algorithm

**Assumption 1:** Words appearing in a ‘common tags’ list of a resource-level folksonomy are higher chances of being similar to the each other than words not on the list.

**Assumption 2:** The more often a pair of words appears in a ‘common tags’ list the greater chance they are similar in semantics.

Based on the Assumption 1 and 2, the proposed co-occurrence-based method is: Given a target word  $w$ , collect 10 resource-level of folksonomies (we can obtain 10 ‘common tags’ lists from the folksonomies) in which the target word  $w$  is placed on the top of each of the folksonomies (or corresponding ‘common tags’ list). For example, the ‘common tags’ list appeared in the Figure 2 can be used for the target word ‘Folksonomy’ because the word is top ranked in the list. Then, the co-occurrence frequency of a pair of the term  $w$  and  $y$  is counted, where  $w$  is not equal to  $y$  and  $y$  is on the ‘common tags’ list. For example, with the ‘common tags’ list shown in the Figure 2 (there are 25 terms on the list, including ‘Folksonomy’), each of the twenty-four pairs of ‘Folksonomy’ and each of the remaining 24 terms on the list is counted as one, regardless of the frequencies in the list. To eliminate the likelihood of a happening-by-chance term, we considered only the terms on the ‘common tags’ lists created by at least a certain number of people. In our experiment, we set 20 to be the threshold value of minimum number of people, which appears to be feasible as a result of our previous empirical work.

The co-occurrence frequency will fall between 1 (at minimum) and 10 (at maximum) in this case. In our experiment, we set various levels of co-occurrence ratio to control the degree of similarity of words: 0.5, 0.7, and 0.9. That is, the 0.9 ratio of co-occurrence frequency of a target word  $w$  to a word  $y$  means that the word  $y$  was co-appeared 9 times among the 10 ‘common tags’ lists each of which the word  $x$  is top ranked.

**Definition 1:** Given a certain level of similarity  $A$ , a word  $y$  is co-occurrence-based similar to a term  $x$  when the term  $y$  appears at the minimum  $A$  ratio on ‘common tags’ lists with  $x$  being top ranked, which is denoted by  $\text{CoSA}(A: y \rightarrow x)$ .

### 3.2 Correlation-Based Similarity Algorithm

**Assumption 3:** Terms  $w_1$  and  $w_2$  with both  $\text{CoSA}(A: w_1 \rightarrow w_2)$  and  $\text{CoSA}(A: w_2 \rightarrow w_1)$  are considered to be more strongly semantically similar than they are in either  $\text{CoSA}(A: w_1 \rightarrow w_2)$  or  $\text{CoSA}(A: w_2 \rightarrow w_1)$ .

Given the Assumption 3, we propose the correlation-based similarity algorithm that is described as follows: Let  $O_{x,p} = \{x_1, x_2, \dots, x_n\}$  be a set of co-occurrence-based similar words of a word  $x$  with  $p$  co-occurrence threshold value. Then,  $x_i \in O_{x,p}$  is considered as a correlation-based similar word for the word  $x$  only if both  $\text{CoSA}(p: x_1 \rightarrow x_2)$  and  $\text{CoSA}(A: x_2 \rightarrow x_1)$  are achieved at the same time. The correlation-based similarity imposes a more strong term relationship than the co-occurrence-based. Note that the co-occurrence-based similarity is a directional relation, but the correlation-based similarity is a non-directional relation.

**Definition 2:** Given a certain similarity level  $A$ , a word  $x$  is correlation-based similar to a word  $y$  only if both  $\text{CoSA}(A: w_1 \rightarrow w_2)$  and  $\text{CoSA}(A: w_2 \rightarrow w_1)$  are satisfied, denoted by  $\text{CrSA}(A: w_1 \leftrightarrow w_2)$ . Note that  $\text{CrSA}(A: w_1 \leftrightarrow w_2)$  is equivalent to  $\text{CrSA}(A: w_2 \leftrightarrow w_1)$ .

## 4 Experiment and Its Results

Using the two proposed methods (CoSA referring to the co-occurrence-based algorithm and CrSA referring to the correlation-based algorithm), similar terms of the most popular tags on del.icio.us posted on the fifteenth of May 2008 (see Figure 2 for the full list of the 140 tags) were collected. We are unable to collect similar terms for 19 out of the 140 tags: We failed to collect as many as 10 folksonomies with 17 tags, and there was no co-occurring word appearing at the minimum ( $p * 10$ ) times from the collected 10 ‘common tags’ lists for 2 tags, where  $p$  (0.5, 0.7, or 0.9) is the pre-set threshold value of co-occurrence frequency.

The average numbers of similar terms found for the 121 (140 – 19) tags are 10.1 ( $p=0.5$ ), 5.1 ( $p=0.7$ ), and 2.6 ( $p=0.9$ ) with CoSA, and 2.6 ( $p=0.5$ ), 1.6 ( $p=0.7$ ), and 0.9 ( $p=0.9$ ) with CrSA. It is easily seen that higher threshold value ( $p$ ) means applying more selective criteria. Also, CrSA provides more selective condition in discovering similar terms that CoSA does.

The 121 terms are checked to see if they are listed in the online version of Merriam-Webster Dictionary (<http://www.merriam-webster.com/dictionary>). Out of the 121, 32 terms are not listed on the dictionary with the following reasons: (1) they does not appear as a main entry or a variant entry in the dictionary; (2) meanings of the considering terms used in Del.icio.us is not in the dictionary (for example, ‘ruby’ is defined in the dictionary as a type of precious stone, but it is mostly used in Del.icio.us as a programming language). There are 22 tags<sup>1</sup> for the first reason and 10 tags<sup>2</sup> for the second reason. The average numbers of similar terms for the 32 not-on-the-dictionary terms are 10.1 ( $p=0.5$ ), 5.9 ( $p=0.7$ ), and 3.3 ( $p=0.9$ ) with CoSA, and 2.4 ( $p=0.5$ ), 1.7 ( $p=0.7$ ), and 1 ( $p=0.9$ ) with CrSA. The average numbers of similar terms for the 32 tags are very close to the ones for the 121 tags we reported earlier. We do not formally conduct an assessment of the quality of the collected similar words for not-on-the-dictionary tags. Instead, we observe that either CoSA with  $p=0.9$  or CrSA with  $p=0.7$  looks to be the best choice, but we recommend CoSA with  $p=7$  for the application interesting having a larger pool of similar terms.

## 5 Conclusion

In this paper, we have proposed two folksonomy-based methods for generating similar terms. Our approach is new in that it utilized the collaboratively collected tags from a web2.0 tool in discovering similar terms of ‘not-in-the-dictionary’ terms particularly. The value of this study lies in the application of collaboratively created indexing terms and the structure of a web2.0 service in mining similar terms. Experimental results demonstrate that our proposed methods identify somewhat similar or near-similar terms for terms that are not yet listed on or not available in a dictionary, whether it is too new, technology-related, or etc. In future work, we plan to work at accumulating knowledge from multiple web2.0 services to enhance the performance of our methods.

## References

1. Chen, H., Ng, T.D., Martinez, J., Schatz, B.R.: A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. *Journal of the American Society for Information Science* 48(1), 17–31 (1997)
2. Wu, H., Zhou, M.: Optimizing synonym extraction using monolingual and bilingual resources. In: *Proceedings of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications*, Sapporo, Japan (2003)
3. Lin, D.: Automatic retrieval and clustering of similar words. In: *COLING-ACL*, pp. 768–774 (1998)
4. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of ACL 2002*, pp. 417–424 (2002)

---

<sup>1</sup> Net, actionscript, css, firefox, iphone, javascript, jquery, lifehacks, linux, microsoft, open-source, osx, php, rails, seo, socialnetworking, ubuntu, webdesign, web2.0, wiki, wordpress, youtube.

<sup>2</sup> Ajax, apple, flash, flex, java, mac, python, ruby, twitter, windows.

5. Chen, H., Lynch, K.J.: Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics* 22(5), 885–902 (1992)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5), 604–632 (1999)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 39(1-7), 107–117 (1998)
8. Senellart, P.P.: Extraction of information in large graphs: Automatic search for synonyms. Technical Report 90, Universite catholique de Louvain, Louvain-la-neuve, Belgium (2001)
9. Jannink, J., Wiederhold, G.: Thesaurus entry extraction from an on-line dictionary. In: *Proceedings of Fusion*, Sunnyvale CA (1999)
10. Crouch, C.J.: An approach to the automatic construction of global thesauri. *Information Processing and Management* 26, 629–640 (1990)
11. Turney, P.D.: Mining the Web for synonyms: PMI\_IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) *ECML 2001. LNCS (LNAI)*, vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
12. Lin, D., Zhao, S., Qin, L., Zhou, M.: Identifying synonyms among distributionally similar words. In: *IJCAI*, pp. 1492–1493 (2003)
13. Arrington, M.: Exclusive: Screen shots and feature overview of Delicious 2.0 preview. *TechCrunch* (September 6, 2007) (Retrieved September 10, 2008), <http://www.techcrunch.com/2007/09/06/exclusive-screen-shots-and-feature-overview-of-delicious-20-preview/>

# Looking for Entities in Bibliographic Records

Trond Aalberg<sup>1</sup> and Maja Žumer<sup>2</sup>

<sup>1</sup> Norwegian University of Science and Technology, Norway

<sup>2</sup> University of Ljubljana, Slovenia

**Abstract.** In this paper we summarize some of the problems and possibilities for frbrising existing bibliographic records based on experiments with 3 different databases.

## 1 Introduction

The Entity-Relationship model published in the IFLA Functional Requirements for Bibliographic Records (FRBR) [7] has generated much interest in library circles and other domains, not only for its theoretical contributions to the field but also through its implementation in practice [2,3]. The core of the model is a set of "product" entities: *Work*, *Expression*, *Manifestation* and *Item* and the relationships that exist between these, together with entities for *Persons* and *Corporate Bodies* and the relationships that can be used for associating these to the appropriate product entity they are related to.

The processing of existing information in order to convert it or interpret it according to the FRBR model is commonly referred to as *frbrisation* [5,6,9], but results so far have shown that it is difficult to achieve high quality. Typical problems include insufficient or erroneously identified entities and relationships and insufficient identification of equivalent entities. Furthermore, many current initiatives only apply parts of the FRBR model or only process parts of the information found in a record. A more complete frbrisation would introduce additional problems that further could reduce the quality of the results.

The main objective for our work in this area is to acquire knowledge about frbrisation of existing bibliographic information and the level of frbrisation that can be achieved in terms of completeness and correctness. Our longer term objective is to develop and explore different strategies and solutions that can be used to improve the frbrisation process of existing bibliographic information. In our work we have analyzed MARC records from 3 different bibliographic databases; the Norwegian BIBSYS database, the Swedish BURK database from BTJ and the Slovenian National bibliography. To analyze these collections we have used a frbrisation tool developed at NTNU [1] that reads a collection of MARC records and produces a collection of FRBR records for the entities and relationships that the tool is able to identify from the input. The tool is rule-based which means that the developers can specify rules for the interpretation and processing of the MARC records and for the above mentioned databases we have developed three independent rule sets. The results have been evaluated by manually inspecting examples.



## 2 Findings

A MARC record is a datastructure based on the ISO 2709 standard [8] and each record is logically divided into fields. The variable datafields typically reflect a logical grouping of information with the subfields as the leaf nodes that carry the information. All MARC formats follow this common structure, but substantial differences between formats occur due to different use of tags and subfield codes and different cataloguing rules and practice. Our experiments have shown that it is possible to use the existing field and subfield schema as the guide for identifying entities in the records. In some cases, such as for persons and corporate bodies, there are specific fields corresponding to the occurrence of entities. In other cases the occurrence of an entity has to be identified by looking for specific subfields or sets of subfields such as for those fields that contain information about both person and title. The MARC structure alone is, however, not sufficient for the identification of entities. Typically there can be multiple fields that in one case should generate two different entities, whereas in other cases this should be interpreted as one and the same entity. Another problem is introduced by the use of subfields containing structured text based on the use of ISBD separators and/or prefix notations. In some cases there is a need to tokenize the text in a subfield when the same subfield value actually is a listing of titles such as in notes and in some uses of the title statement field. Even though the text often is based on strict rules for the use of prefixes and/or separator this typically leads to many errors caused by inconsistent syntax.

Persons and Corporate Bodies are by far the easiest entities to extract from existing MARC records. The fields and subfields for these entries can directly be used to identify and extract attributes of such entities from existing records. Libraries typically maintain authority files to ensure that persons and corporate bodies are catalogued consistently and uniquely and for this reason there is a high level of identification of equivalent entities.

The work entity is the cornerstone of the FRBR model and any frbrisation process needs to include a way of identifying the work entities that are described implicitly or explicitly in the records. In our tool we basically look for titles and the creator of the work. The Norwegian and Swedish databases have specific repeatable fields for original titles that frequently are used for works that have expressions in different languages. For the Slovenian bibliography some original titles were found in uniform title fields, others were only found in note fields. For all databases, however, the number of records with a specific work-level title was low. In app. 80% of the records the title statement is the only source for discovering what work(s) it contains. Dealing with publications that contain multiple works is another challenge. In particular for translated works, where both original and translated titles are listed, it is in most cases impossible to determine whether an added entry should be interpreted as an additional work or not. In general, MARC records have a structure that is tailored to publications that contain a single work/expression. The practice for cataloguing publications that contain multiple expressions is typically different between catalogues and even within one and the same catalogue we found a variety of solutions in use.

Expressions are on one hand often perceived as the more vague entity in the FRBR product model but on the other hand they are less problematic to identify. An expression is always related to a single work and if we are able to identify the work contained in a publication, we implicitly also can identify that the publication contains an expression of this work. Expressions are basically identified by the works they are expressions of, and in addition we need the unique characteristics of each expression to distinguish it from other expressions of the same work. If a work is expressed in different translations we typically should look for the language of the expression, the translator and other features that can be associated with the expression. If no person or corporate body can be associated with an expression during the process, expressions can only be identified at a categorical level which typically was the case for both the Norwegian and Swedish databases. In the Slovenian database it was, however, possible to use persons/corporate bodies related to the expressions in the identification due to the extensive use of relator codes. Since the identification of expressions are directly related to the works we identify, errors introduced in the work identification process are inherited in the expression identification process.

Manifestations are quite straightforward to identify due to the correspondence between a record and a manifestation. Each record typically describes a single publication and the record will in this case only result in a single manifestation. None of the examined catalogues maintain information about holdings. In other catalogues there are fields used for recording individual items in libraries by call number, shelf list etc. which can be used for the identification of items.

Being able to deduce a set of entities from a record is merely the first step in a frbrisation process. The more difficult part of the process is to determine the relationships between the entities that are identified. One problem we have encountered in these databases is caused by insufficient use or lack of relator codes. Without knowing the role of a person or corporate body it is difficult to determine whether the person or corporate body should be related to the work identified, the expression or the manifestation. In both the Norwegian database and the Swedish database there are only a few relator codes in use, whereas in the Slovenian bibliography there was an extensive set of relator codes. Ambiguous or lacking relator codes typically lead to large number of persons that in the end remain unrelated to any entity. Another problem we encountered in these databases was the problem of establishing proper relationships between works and expressions when there are multiple persons and multiple works/expressions identified within a record. With multiple uniform or original titles and separate entries for multiple persons there is essentially no way of telling which person should be related to which work/expression. In UNIMARC, entries for titles and persons are separated, which leads to a problem for all records that contain descriptions of multiple works/expressions. Ideally this should have been solved by having multiple records and using linking, but in practice this was not the case. In the Norwegian and Swedish databases this was a mainly a problem for translated works, where the cataloguers typically would prefer using the original title fields in combination with person-only entries in the added entry fields.

### 3 Conclusion and Further Work

Our experiments have shown that it is possible to frbrise existing MARC based records, but there are many problems yet to be solved before we can achieve results of acceptable quality. Records that describe a publication with a single expression of a single work, with the use of detailed relator codes and if there is a uniform title or original title present, can easily be frbrised, but for more complex records the MARC structure simply contains insufficient information for the generation of FRBR relationships. Possible solutions that can be used to further improve the results is on one hand to apply more advanced text/data processing such as proposed by [4]. On the other hand we need to build authority services that can be used to identify, look up and retrieve information about such entities as works and expressions. Such services can be populated by mining several catalogues for entities and relationships with a high level of reliability and in this way enable the reuse of information across records in the frbrisation process.

### References

1. Aalberg, T.: A process and tool for the conversion of MARC records to a normalized FRBR implementation. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 283–292. Springer, Heidelberg (2006)
2. Le Boeuf, P.: FRBR and further. *Cataloging & Classification Quarterly* 32(4) (2001)
3. Le Boeuf, P.: *Functional Requirements for Bibliographic Records (FRBR): Hype, or Cure-All?* Haworth Press, Birmingham (2005). Published simultaneously as *Cataloging & Classification Quarterly* 39(3-4) (2005) ISSN 0163-9374
4. Freire, N., Borbiha, J., Clavado, P.: Identification of FRBR works within bibliographic databases: An experiment with UNIMARC and duplicate detection. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 267–276. Springer, Heidelberg (2007)
5. Hegna, K., Muromaa, E.: Muromaa. Data mining MARC to find: FRBR? In: 68th IFLA General Conference and Council, Glasgow, Scotland (2002)
6. Hickey, T.B., O'Neill, E.T., Toves, J.: Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). *D-Lib Magazine* 8(9) (September 2002) <http://www.dlib.org/dlib/september02/hickey/09hickey.html>
7. IFLA study group on the functional requirements for bibliographic records. *Functional Requirements for Bibliographic Records: Final report*. Saur (1998)
8. International Organization for Standardization. *Information and Documentation: Format for Information Interchange*. International standard ISO 2709:1996 Third edition, ISO (1996)
9. O'Neill, E.: FRBR (Functional Requirements for Bibliographic Records): application of the entity-relationship model to Humphry Clinker. *Library Resources and Technical Services* 46(4) (2002)

# Protecting Digital Library Collections with Collaborative Web Image Copy Detection

Jenq-Haur Wang<sup>1</sup>, Hung-Chi Chang<sup>2</sup>, and Jen-Hao Hsiao<sup>2</sup>

<sup>1</sup> National Taipei University of Technology, Taiwan

<sup>2</sup> Academia Sinica, Taiwan

jhwang@csie.ntut.edu.tw, {hungchi, jenhao}@iis.sinica.edu.tw

**Abstract.** There's a pressing need for protecting digital library (DL) collections since digital objects can be easily copied, edited, and redistributed. Digital content protection is thus attracting attention. In this paper, we propose a peer-to-peer approach to collaborative Web image copy detection for protecting DL collections, which is helpful to public libraries with limited resources. We design a collaborative framework that incorporates multiple peers to share the loads of crawling and copy detection tasks. It facilitates better utilization of limited network and computing resources. Our experiment results demonstrate the efficacy and potential of the proposed approach. Further investigation is needed to verify its scalability.

## 1 Introduction

Copyrighted content in digital library (DL) collections can be copied and distributed very easily that threatens their intellectual property rights. How to protect valuable digital assets is now a pressing issue in the field of digital rights management (DRM). DRM systems usually incorporate information security techniques such as cryptography and access control for protection. They discourage copyright infringement by enforcing a uniform policy in rights enforcement environments [6] through the lifecycle of contents.

Instead of restricting access to the content, we adopt copy detection as a second line of defense after the content has been released to the public. Multimedia contents typically use digital watermarking, but embedded watermarks degrade the quality of content, and cannot survive some image processing attacks. Therefore, content-based copy detection (CBCD) has attracted increasing attention. Without hiding additional information in an image for copy detection, the image itself can serve the same purpose. The popular ordinal measures [1] use a single global feature vector to describe an image. Here we extract representative local features from images, which is more accurate and resistant to image attacks.

The goal is to utilize CBCD techniques to find Web images that could potentially be near-duplicates of protected content in DLs. One challenge is to efficiently collect the huge number of Web images given the limited resources in public libraries. Also, comparing against the huge collections is time-consuming. Therefore, we propose a peer-to-peer (P2P) approach to collaborative Web image copy detection for protecting DL collections. It's a collaborative framework for DLs with common needs of content

protection to contribute part of their computing and network resources and share the loads of image sampling, processing, and querying tasks. First, we adapt a hybrid P2P architecture [8] for effective distribution of tasks among DLs. Second, we apply SIFT descriptors [5] and approximate nearest-neighbor search to efficiently find near-duplicate images. Experiments on Web images and the National Digital Archives Program (NDAP) archival sites in Taiwan were conducted. The collaborative crawlers showed excellent efficiency in image sampling, and the image copy detection method achieved good accuracy with reasonable response time. Further investigation is needed to verify this approach in larger scales.

## 2 The Proposed Approach

The proposed approach for DLs to collaborate on Web image sampling and copy detection consists of four main modules: Domain Partitioning (DP), Image Sampling (IS), Query Processing (QP), and Image Copy Detection (ICD). They are divided into two classes: *global* coordinating functions across DLs, and *local* functions on each DL. Therefore, we adapt a hybrid P2P architecture [8] where each domain contains a *super-peer* coordinating some *general peers*. The super-peer maintains global coordinating functions, while general peers implement local processing functions.

**Domain Partitioning (DP).** The framework is initiated by DP module that dynamically partitions the Web and assigns the partitions represented by the corresponding Uniform Resource Locators (URLs) to each general peer. *URL dispatching* [8] is employed to find the peer with the minimum load and network proximity for a given URL.

**Image Sampling (IS).** Once assigned URLs from DP module, IS module on each peer crawls the images in responsible domains. The basic operations are similar to those of a distributed crawler [2] with dynamic assignment of hierarchical partitions in an exchange mode. New images are directly stored into the partial Web image DB on the same DL to avoid the huge overhead for moving image files. A distributed Web image DB is thus collectively formed by individual partial DBs.

**Query Processing (QP).** When DLs want to check if the protected images could possibly be copied, the images will be sent as queries to the super-peer. QP module will forward the query to selected peers. The simplest way is to forward to all peers or randomly selected peers. If we want to further minimize the number of queries to be forwarded, a compact feature summary is needed on each peer for quick estimation of peers that might have similar content features as the query.

**Image Copy Detection (ICD).** Given an image query, ICD module estimates the similarity between the query and the sampled images in the Web image DB.

- (1) *Feature Extraction.* Local features called *SIFT descriptors* [5] are extracted from images. It employs transformations to identify potential interest points in images, and *keypoints* are located after removing unstable candidates.
- (2) *Bags of Keypoints.* Since the number of SIFT descriptors in an image is quite large, a bag of keypoints method [3] is then used to group them into more concrete visual

object components for efficient comparison. The  $k$ -Means clustering algorithm is applied to SIFT descriptors to construct *visual vocabularies* from the cluster centroids.

- (3) *Approximate Nearest-Neighbor Search*. For each visual vocabulary in the query, we find the  $k$  nearest neighbors in the distributed Web image DB. The images that are similar to more visual vocabularies get higher ranks as potential copies.

### 3 Experiments

**Performance of Distributed Crawling.** To compare the performance of distributed crawlers with ordinary ones, we designed two modes of operation: *standalone* and *collaborative* modes. Within a fixed period of time, identical seeds from selected NDAP archival sites were used in both modes, and the performance was evaluated using metrics such as the numbers of pages crawled, links extracted, images found, and the size of files downloaded. These metrics reflect the amount of data processed, which implies the coverage of crawling and the chance to find copies. In a 60-hour crawl for the simplified setting of a single P2P domain with one super-peer and 5 general peers, 109,778 pages (4.75 GB) were fetched in standalone mode and 409,671 pages (4.77 GB) in collaborative mode.

**Table 1.** A comparison of the relative performance for collaborative and standalone modes

Metric	Crawled pages	Extracted links	Found images	Size of downloaded files
Ratio	18.66	19.39	18.34	5.02

As shown in Table 1, collaborative peers outperformed standalone crawlers in terms of the amount of data processed. Such outstanding performance could be due to our good design in infrastructure and coordinating functions that effectively balance communication and system loads among peers. This shows the potential of our proposed approach for efficiently sampling Web data.

**Performance of Image Copy Detection.** The images extracted from the 409,671 pages in the first experiment were filtered where 20,000 images were used as our *base set*. Then 100 images were randomly selected from the base set as the copyrighted images, and also as the *test queries*. To test the resistance to image attacks, we use some categories of attacks from Stirmark benchmark [7] to generate 125 near-duplicate images for each of the 100 copyrighted images, which forms an *answer set* of 12,600 images. Thus a total image collection of 32,500 images was built. The popular DCT ordinal measure [4] was adopted as the baseline for comparison. As shown in Table 2, the detection accuracy of the proposed approach outperforms the baseline in both recall and precision rates. The  $F$ -Measure shows an increase of 26%.

**Table 2.** A comparison of average precisions/recalls and average response time per query

	Recall	Precision	F-Measure	Standalone	Collaborative
Proposed Approach	0.945	0.987	0.966	3.2 s	1.9 s
Baseline	0.584	0.892	0.706	2.8 s	1.6 s

The efficiencies of both approaches are reasonable, but the proposed approach requires more computation time for extracting local features from images. The average response time per query improved in collaborative mode, and the ratio between two modes was about 1.6:1, which could be due to the approximate search.

## 4 Discussions and Conclusion

The experiments on collaborative crawlers showed efficient image sampling, and the image copy detection method achieved high accuracy within reasonable time. For further improvement, query processing is critical, especially when the number of Web images is huge. As the number of peers or domains grows, more computation and network overhead will be introduced for more complicated message exchanges. Designing an effective feature summary could be an important future work. Many applications of the proposed framework are possible: cross-archive applications such as distributed archiving, video copy detection, and Web mining for relevant information from images. Our approach can also work as an infrastructure for sampling the content usage statistics on the Web. Further investigation is needed to verify its efficacy in large scales.

## References

1. Bhat, D., Nayar, S.: Ordinal measures for image correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(4), 415–423 (1998)
2. Cho, J., Garcia-Molina, H.: Parallel crawlers. In: *Proceedings of the 11th World Wide Web conference (WWW 2002)*, pp. 124–135 (2002)
3. Csurka, C., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoint. In: *Workshop on Statistical Learning in Computer Vision. ECCV 2004* (2004)
4. Kim, C.: Content-based image copy detection. *Signal Processing: Image Communication* 18(3), 169–184 (2003)
5. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
6. Nicolakis, T., Pizano, C., Prumo, B., Webb, M.: Protecting digital archives at the Greek Orthodox Archdiocese of America. In: *Proceedings of the 3rd ACM Workshop on Digital Rights Management*, pp. 13–26 (2003)
7. Stirmark benchmark 4.0, <http://www.petitcolas.net/fabien/watermarking/stirmark/>
8. Wang, J.H., Chang, H.C., Lin, C.Y., Chien, L.F.: A peer-to-peer approach to collaborative repository for digital libraries. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) *ICADL 2006. LNCS*, vol. 4312, pp. 511–514. Springer, Heidelberg (2006)

# Enhancing the Literature Review Using Author-Topic Profiling

Alisa Kongthon, Choochart Haruechaiyasak, and Santipong Thaiprayoon

Human Language Technology Laboratory (HLT),  
National Electronics and Computer Technology Center (NECTEC)  
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand  
{alisa.kon,choochart.har,santipong.tha}@nectec.or.th

**Abstract.** In this paper, we utilize bibliographic data for identifying author-topic relations which can be used to enhance the traditional literature review. When writing a research paper, researchers often cite on the order of tens of references which do not provide the complete coverage of the research context especially when the targeted research is multidisciplinary. Author-topic profiling can help researchers discover a broader picture of their topic of interest including topical relationships and research community. We apply the Latent Dirichlet Allocation (LDA) to generate multinomial distributions over words and topics to discover author-topic relations from text collections. As an illustration, we apply the methodology to bibliographic abstracts related to Emerging Infectious Diseases (EIDs) research topic.

**Keywords:** Bibliographic data, text mining, Latent Dirichlet Allocation (LDA), author-topic profiling, literature review.

## 1 Introduction

The conventional literature review process usually starts by identifying a few state-of-the-art papers. The information on authors and citations of these papers will be used extensively to identify other related literature. Researchers will then digest content of these articles on a one-by-one basis. Hence, such a process can often limit researchers' perspective to only particular pieces of literature.

With the current advancement in information technology, digital libraries can now be used to make documents more easily accessible. Bibliographic databases are becoming widely known as a starting point for the unconventional literature review process. These databases usually provide tools for researchers to search for their articles of interest. In addition to a typical basic or advanced search capability, modern bibliographic databases provide features to list and rank search results by fields such as author, affiliation, subject area, and publication year. Researchers can comprehend the data set by observing the list of a particular field. For instance, the list of leading subject areas represents the prominent sub-topics within that research topic. However, subject areas are usually controlled and indexed by database providers. Using such subjects may not fully reflect the



real intention of the authors due to bias and error introduced by human editors. Hence, an approach to automatically extract topics from content of documents is needed.

There have been many studies on discovery latent topics from text collections [1]. Latent Semantic Analysis (LSA) uses singular value decomposition (SVD) to map high-dimensional term-by-document matrix to a lower dimensional representation called latent semantic space [2]. However SVD is actually designed for normally-distributed data. Such a distribution is inappropriate for count data which is what a term-by-document matrix consists of. As an alternative to standard LSA, Probabilistic Latent Semantic Analysis (pLSA) assumes each word in a document as a sample from a mixture model, where the mixture decompositions are multinomial random variables that can be viewed as representations of topics [3]. Hence each word is generated from a single topic, and different words in a document may be generated from different topics. However, the pLSA model encounters overfitting problem because the number of parameters grows linearly with the number of documents. Latent Dirichlet Allocation (LDA) is then introduced to correct such problem. LDA is a generative probabilistic model for a set of documents [4]. The basic idea behind this approach is that documents are represented as random mixtures over latent topics, where each topic is represented by a probability distribution over words.

Steyvers et. al. extended the LDA to include authorship information so that authors are linked to terms in documents via latent topics [5]. This model not only discovers what topics are expressed in a document, but also which authors are associated with each topic. Instead of associating each document with a distribution over topics, the author-topic model associates each author with a distribution over topics. However, one common problem associated with the topic model is how to effectively label the discovered topics. Typically, the topic is labeled numerically. Another common way to assign topic name is by appending the terms appear in that topic together. In this paper, we modify the author-topic model to capture the relationship between authors and topics from a set of bibliographic data. Each author is represented by a probability distribution over topics, and each topic is a probability distribution over words extracted from abstracts and controlled keywords prepared by the database provider. We utilize these two types of terms in our model so that we could have the most comprehensive information. For each derived topic, the controlled keyword is then used as a representative for all words within that topic. With such informative representation, researchers can better understand the concepts within their research of interest. Our approach promises to improve traditional literature review process by helping researchers depict the “forest” (broader patterns of research activity) before looking at the “trees” (important prior art).

## 2 The Author-Topic Profiling

In our author-topic profiling model, the input data consists of a set of  $m$  bibliographic documents denoted by  $\mathcal{D} = \{D_0, \dots, D_{m-1}\}$ . Given a document

collection, the author-topic identification problem becomes the model fitting that finds the best estimate of the topic-word distributions and the author-topic distributions. Gibbs sampling is used to solve this model fitting problem. As a result, the LDA algorithm generates a set of  $n$  topics denoted by  $\mathcal{T} = \{T_0, \dots, T_{n-1}\}$ . Each topic is a probability distribution over  $p$  words denoted by  $T_i = [w_0^i, \dots, w_{p-1}^i]$ , where  $w_j^i$  is an estimated probability value of word  $j$  assigned to topic  $i$ . Based on this model, each author can be represented as a probability distribution over the topic set  $\mathcal{T}$ , i.e.,  $A_i = [t_0^i, \dots, t_{n-1}^i]$ , where  $t_j^i$  is an estimated probability value of topic  $j$  assigned to author  $i$ . For each generated topic, we also calculate the binomial Z-score, which measures the degree of independence of the term from the topic. The higher Z-score means that the term is more dependent on the topic. Hence we select a controlled keyword with the highest Z-score as the representative for each topic.

### 3 A Case Study of Emerging Infectious Diseases (EIDs)

We illustrate the application of our proposed method through a case study of Emerging Infectious Diseases (EIDs). This case study aims to explore the possibility of using converging technologies to combat EIDs<sup>1</sup>. The literature review is conducted by searching for related publications from *Compendex* database. To come up with an appropriate search terminology, we experimented with various Boolean search operators to cover “emerging infectious diseases.”<sup>2</sup> The result data set contains 5,046 records. Obviously with the traditional literature review process, one cannot digest content by reading all of these papers. We apply the proposed author-topic profiling to enhance the conventional literature review process. Figure 1 shows six different derived topics (out of 50 topics). Each table in Figure 1 presents the top-10 words that are most likely to be generated when that topic is created, and the top-5 authors who are most likely to write a word if it has come from that topic. The topic name is associated with a controlled keyword with the highest Z-score.

Figure 1 shows quite representative results. Topics related to different research areas such as ecosystem, algorithms, biosensors, among others, can be derived from our data set. The words associated with each topic are also quite precise in a semantic sense of a particular area of research. The topic name also best describes words within that topic. Such analytical results can help researchers depict related topics to EIDs such as viruses and immunology. Moreover, previously unknown relevant topics such as ecosystem and biosensors can be discovered as well. Also if researchers need to focus on a biosensors topic, they can start searching for articles written by S.S. Iqbal or other leading authors.

<sup>1</sup> EID:Roadmapping Converging Technologies for Combat Emerging Infectious Diseases, <http://www.apecforesight.org>

<sup>2</sup> Search terminology - [(infectious disease) OR (infectious diseases) OR pandemic OR epidemic OR outbreak OR outbreaks OR flu OR influenza] NOT [(computer viruses) OR (computer worms) OR (network protocols)].

Topic: Viruses			Topic: Ecosystem			Topic: Healthcare		
Extracted terms		Top-5 Authors	Extracted terms		Top-5 Authors	Extracted terms		Top-5 Authors
virus	0.048	Hirschman L.	species	0.015	Martikainen P.	health	0.020	Nichter L.S.
influenza	0.041	Jin M.	forests	0.010	Sitonen J.	risk	0.010	Cox M.J.
transmission	0.014	Subba V.	population	0.009	Punttila P.	public	0.008	Chou D.
pandemic	0.011	Spiro D. J.	outbreaks	0.008	Corbett D.	disease	0.007	Liu B.
human	0.010	St. George K.	beetles	0.005	Erlandson J.M.	management	0.007	Weng L.
avian	0.010		bark	0.004		epidemic	0.006	
flu	0.009		host	0.004		system	0.005	
sars	0.007		insect	0.004		outbreak	0.005	
h5n1	0.007		climate	0.004		cancer	0.004	
disease	0.006		moth	0.003		prevention	0.004	

Topic: Immunology			Topic: Algorithms			Topic: Biosensors		
Extracted terms		Top-5 Authors	Extracted terms		Top-5 Authors	Extracted terms		Top-5 Authors
hiv-1	0.016	Ahuja S.K.	data	0.017	Edmonds C.B.	detection	0.030	Iqbal S.S.
cells	0.015	Begum K.	system	0.012	Basham D.	pcr	0.014	Bruno J.G.
infection	0.012	Jimenez F.	information	0.011	Cronin A.	dna	0.012	Batt C.A.
hiv	0.011	Telles V.	detection	0.007	Jagels K.	rapid	0.010	Mayo M.W.
immune	0.010	Stahl-Hennig C.	disease	0.006	Simmonds M.	system	0.009	Hsieh T.-M.
virus	0.010		algorithm	0.006		assay	0.009	
cd8	0.008		network	0.006		molecular	0.006	
aids	0.007		genome	0.005		real-time	0.005	
t-cell	0.006		plague	0.005		infectious	0.005	
immunodeficiency	0.006		outbreaks	0.005		sensitive	0.005	

**Fig. 1.** Sample topics related to EIDs. Each topic is illustrated with the top-10 words and the top-5 authors.

## 4 Conclusions

We proposed an approach called author-topic profiling to augment, not to replace, the conventional literature review. This proposed method is based on a probabilistic topic model which can automatically extract information about authors and topics from a large set of documents. In addition, our model can effectively label the discovered topics by utilizing controlled keywords. From the illustrative case study, our model was able to extract “hidden” information from our sample data set.

## References

1. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (eds.) *Latent Semantic Analysis: A Road to Meaning*. Lawrence Erlbaum, Mahwah (2006)
2. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6), 391–407 (1990)
3. Hofmann, T.: Probabilistic latent semantic indexing. In: *Proc. of the 22nd annual international ACM SIGIR conference*, pp. 50–57 (1999)
4. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 993–1022 (2003)
5. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery. In: *Proc. of the 10th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pp. 306–315 (2004)

# Article Recommendation Based on a Topic Model for Wikipedia Selection for Schools

Choochart Haruechaiyasak and Chaianun Damrongrat

Human Language Technology Laboratory (HLT)  
National Electronics and Computer Technology Center (NECTEC)  
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand  
{choochart.haruechaiyasak, chaianun.damrongrat}@nectec.or.th

**Abstract.** The 2007 Wikipedia Selection for Schools is a collection of 4,625 selected articles from Wikipedia as educational for children. Users can currently access articles within the collection via two different methods: (1) by browsing on either a subject index or a title index sorted alphabetically, and (2) by following hyperlinks embedded within article pages. These two retrieval methods are considered static and subjected to human editors. In this paper, we apply the Latent Dirichlet Allocation (LDA) algorithm to generate a topic model from articles in the collection. Each article can be expressed by a probability distribution on the topic model. We can recommend related articles by calculating the similarity measures among the articles' topic distribution profiles. Our initial experimental results showed that the proposed approach could generate many highly relevant articles, some of which are not covered by the hyperlinks in a given article.

**Keywords:** Latent Dirichlet Allocation (LDA), content-based filtering, recommender system, educational Web contents, Wikipedia.

## 1 Introduction

Today there are numerous Web sites which contain excellent and freely available educational contents. One of the most well-recognized resources is the Wikipedia encyclopedia. As of July 2008, there are over *2,400,000* articles available in English and many in other languages. The full volume of Wikipedia contents, however, contains some articles which are unsuitable for children. In May 2007, the SOS Children's Villages, the world's largest orphan charity, launched the Wikipedia Selection for Schools [1]. The collection contains *4,625* selected articles based on the UK National Curriculum and similar curricula elsewhere in the world. All articles in the collection have been cleaned up and checked for suitability for children.

The content of Wikipedia for Schools can be navigated by browsing on a pictorial subject index or a title word index of all topics. Figure 1 lists the first-level subject categories provided with the collection. Organizing articles into the

---

<sup>1</sup> <http://schools-wikipedia.org>

<b>Art</b>	<b>Business Studies</b>	<b>Citizenship</b>	<b>Countries</b>
<b>Design and Technology</b>	<b>Everyday life</b>	<b>Geography</b>	<b>History</b>
<b>IT</b>	<b>Language and literature</b>	<b>Mathematics</b>	<b>Music</b>
<b>People</b>	<b>Religion</b>	<b>Science</b>	

**Fig. 1.** The subject categories under the Wikipedia Selection for Schools

subject category set provides users a convenient way to access articles on the same subject. However, articles across different subject categories may also be related. For example, the article *Great Wall of China* under the subject category *Design and Technology: Architecture* is considered related to the article *Beijing* which is categorized under *Geography: Geography of Asia* as well as the article *Qin Shi Huang* under *People: Historical figures*.

In addition to the browsing method, users may access articles by following hyperlinks embedded within the article content. This method is, however, subjective to human editors. Some links to related articles could be neglected unintentionally. Also, a related article could not be inserted as a hyperlink if there is no term describing it within the current article.

Due to the drawbacks of the above retrieval methods, we propose a content-based filtering method for recommending related articles based on a topic model. We apply the Latent Dirichlet Allocation (LDA) algorithm on the article collection to generate a topic model. An article can be expressed by a distribution over a set of topics. A topic can be represented as a distribution over a set of terms. Therefore, given an article, we could recommend related articles by calculating the similarity over the topic distribution profiles. Article recommendation based on the topic model could discover relevant articles which may come from different subjects and may not be reachable by the hyperlinks.

## 2 A Topic-Model Based Article Recommendation

There have been many studies on discovering latent topics from text collections [2]. Recently, the Latent Dirichlet Allocation (LDA) has been introduced as a generative probabilistic model for a set of documents [1]. The basic idea behind this approach is that documents are represented as random mixtures over latent topics. Each topic is represented by a probability distribution over the terms. Each document is represented by a probability distribution over the topics.

Figure 2 illustrates the process of recommending articles based on a topic model. The input data for the LDA algorithm consists of an article collection which is a set of  $m$  documents denoted by  $\mathcal{D} = \{D_0, \dots, D_{m-1}\}$ . The LDA algorithm generates a set of  $n$  topics denoted by  $\mathcal{T} = \{T_0, \dots, T_{n-1}\}$ . Each topic is a probability distribution over  $p$  words denoted by  $T_i = [w_0^i, \dots, w_{p-1}^i]$ , where  $w_j^i$  is a probability value of word  $j$  assigned to topic  $i$ . Based on this topic model, each document can be represented as a probability distribution over the topic set  $\mathcal{T}$ , i.e.,  $D_i = [t_0^i, \dots, t_{n-1}^i]$ , where  $t_j^i$  is a probability value of topic  $j$  assigned to

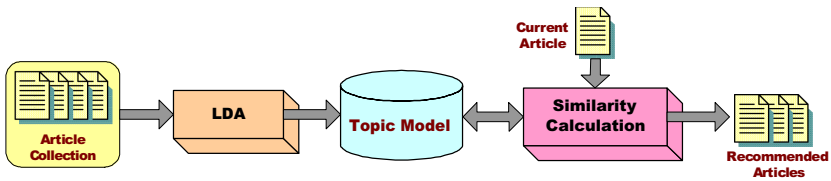


Fig. 2. The process of article recommendation based on a topic model

document  $i$ . To recommend relevant articles, we calculate the similarity between the topic distribution of a given article to all other articles's distributions and select the ones with the highest similarity values.

### 3 Experiments and Discussion

The Wikipedia Selection for Schools is available from the SOS Children's Villages Web site [2](http://www.soschildrensvillages.org.uk/charity-news/wikipedia-for-schools.htm). We used the LDA algorithm provided by the linguistic analysis tool called LingPipe [3](http://alias-i.com/lingpipe) to run our experiments.

Topic #11		Topic #29		Topic #34		Topic #44	
Terms	Prob.	Terms	Prob.	Terms	Prob.	Terms	Prob.
storm	0.017	art	0.015	species	0.016	computer	0.016
hurricane	0.011	painting	0.011	dna	0.010	windows	0.010
tropical	0.010	italy	0.010	cell	0.009	system	0.009
florida	0.010	style	0.010	plant	0.006	software	0.006
damage	0.010	artist	0.008	organism	0.006	data	0.006
wind	0.008	architecture	0.007	genetic	0.005	internet	0.005
cause	0.007	rome	0.007	life	0.005	user	0.005
atlantic	0.007	renaissance	0.005	darwin	0.005	version	0.005
season	0.006	baroque	0.005	protein	0.005	microsoft	0.005
august	0.006	sculpture	0.005	animal	0.004	programming	0.004

Fig. 3. Examples of topics generated by using the LDA algorithm

We applied the article recommendation approach described in the previous section on the article collection. Figure 3 shows some topic examples generated by the LDA algorithm. Each table lists the top-10 terms ranked by the probabilistic values. It can be observed that the LDA could conceptually cluster highly similar terms into the same topics.

Figure 4 shows the top-10 article recommendation given specific article titles. The recommended articles which do not appear as hyperlinks on the given articles are highlighted using boldface. It can be observed that most of recommended articles are not included as the hyperlinks. Therefore our proposed method could help discover previously missing articles but are considered relevant.

<sup>2</sup> <http://www.soschildrensvillages.org.uk/charity-news/wikipedia-for-schools.htm>

<sup>3</sup> <http://alias-i.com/lingpipe>

**Article: Gravitation**

Related articles	Score
(1) Black hole	0.9920
(2) Redshift	0.9915
(3) Hubble's law	0.9894
(4) Big Bang	0.9837
(5) Metric expansion of space	0.9681
(6) Cosmic microwave background radiation	0.9671
(7) Universe	0.9608
(8) Speed of light	0.9590
(9) Plasma (physics)	0.9509
(10) Cosmic inflation	0.9393

**Article: Tsunami**

Related articles	Score
(1) Tropical cyclone	0.9638
(2) 2004 Indian Ocean earthquake	0.9461
(3) Eye (cyclone)	0.9431
(4) Tornado	0.9401
(5) Flood	0.9100
(6) 2005 Sumatra earthquake	0.8893
(7) Hurricane Floyd	0.8859
(8) Storm of October 1804	0.8833
(9) Cyclone Rosita	0.8780
(10) Tropical Storm Vamei	0.8775

**Article: Isaac Newton**

Related articles	Score
(1) Leonhard Euler	0.9429
(2) Georg Cantor	0.9274
(3) Carl Friedrich Gauss	0.9230
(4) Albert Einstein	0.9179
(5) Niels Bohr	0.8813
(6) Paul Dirac	0.8728
(7) David Hilbert	0.8576
(8) Max Planck	0.8549
(9) William Thomson, 1st Baron Kelvin	0.8471
(10) John von Neumann	0.8320

**Article: Open source**

Related articles	Score
(1) Internet	0.9677
(2) Functional programming	0.9568
(3) Markup language	0.9559
(4) Computer programming	0.9557
(5) World Wide Web	0.9529
(6) Python (programming language)	0.9518
(7) BASIC	0.9474
(8) C++	0.9473
(9) Perl	0.9473
(10) Cryptography	0.9455

**Fig. 4.** Examples of article recommendation based on the topic-model approach

## 4 Conclusions

We proposed a topic-model based method for recommending related articles in Wikipedia Selection for Schools. The topic model is generated by using the Latent Dirichlet Allocation (LDA) algorithm. The experimental results showed that the proposed method could help discover additional related articles, some of which are not listed as hyperlinks within a given article.

Our future works include the construction of an evaluation corpus. A set of random articles will be selected and all related articles will be judged by human experts. The corpus is useful in performing the empirical analysis of adjusting the LDA parameters to achieve the best recommendation results.

## References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *The Journal of Machine Learning Research* 3, 993–1022 (2003)
2. Steyvers, M., Griffiths, T.: Probabilistic topic models. In: Landauer, T., McNamara, D., Dennis, S., Kintsch, W. (eds.) *Latent Semantic Analysis: A Road to Meaning*. Lawrence Erlbaum, Mahwah (2006)

# On Developing Government Official Appointment and Dismissal Databank

Jyi-Shane Liu

University Libraries, Natioanl Chengchi University  
64 Sec. 2 Zhih-Nan Rd., Taipei, Taiwan  
jsliu@cs.nccu.edu.tw

**Abstract.** In this paper, we report a databank development project in which structured textual data from historical documents are extracted to provide information access of higher data granularity. The availability of the databank opens up tremendous opportunities for research topics in government personnel systems that were limited by data acquisition difficulty in the past. The project demonstrates the potential of library as data producer in exploiting primary textual resources and developing value-added digital collection.

**Keywords:** Databank development, Value-added digital collection.

## 1 Introduction

Over the last twenty years, advances in digital libraries have re-shaped how users expect to use libraries and how libraries respond to changing information needs. While maintaining traditional functions, libraries have become pro-active in creating new values and expanding new services [1][2]. We consider a new role of digital library as data producer that compiles primary research data from textual resources. The notion of databank development has been mostly associated with recording and organizing scientific data [3] and social surveys [4]. As libraries become more active in selecting source materials for digital archiving, the reviewing process presents an opportunity for important subject data to be identified. Many subject data of research values are difficult to access as they are buried in voluminous text of historical documents. Thus, libraries are in a unique position to recognize the opportunity and initiate a data production process to extract subject data from textual resources. By exercising the capability to produce valuable textual data from primary source collection, libraries provide yet another vital support to research communities.

## 2 Subject Domain

Government gazettes are printed publications available in most major libraries. As bulletins of official announcements, records, codes, and orders, government gazettes provide authoritative government information and are persistent with the existence of governments. Appointment/dismissal orders are authoritative directions issued by the reigning President to appoint/dismiss a named official to/from a named government



post. Every instance of appointment/dismissal of a government official is authorized through a written order and is publicly announced.

Given the textual nature of the appointment and dismissal orders, there are a number of ways to access the government personnel changes information. First, paper-based documents can be converted to electronic texts so as to enable full-text search. Words and phrases can be used with Boolean logics to retrieve a subset of flat documents where constraints of linguistic form are satisfied. However, as in common full-text document retrieval, the results require considerable human processing and filtering. Second, electronic texts of appointment/dismissal orders can be further annotated (encoded) by adding semantic information to text pieces with selected tag set. Text encoding enables semantic search/retrieval and allows automatic mapping to entity relation data model. Personnel information as revealed in the appointment/dismissal orders can be fully specified and accessed. Prerequisite of text encoding includes acquiring convenient annotation tools and training capable annotators.

It is observed that appointment/dismissal orders are compact written forms of personnel changes. Only a very small portion of text pieces does not correspond to entities and relations of personnel changes instances. Text encoding is usually intended to distinguish a small part of informational text pieces from the remaining text. In the case of appointment/dismissal orders, text encoding seems to be over-annotating the entire text. Another problem also arises from the compact nature of the appointment/dismissal orders. Several persons may share the same association of government unit, rank, or title in an order, while only one set of textual pieces are present. These omitted information needs to be inferred and appended so that job change information of each named person is as complete as semantically revealed in the order. Given the purpose of compiling personnel changes information for direct data access and analysis, we take the approach of straightforward extraction and conversion into structured data. In other words, informational textual pieces from the documents are manually identified and keyed into a relational database. Omitted information is also concurrently inferred and the missing textual pieces are supplied during data entry process.

### **3 Databank Development**

The development of government official appointment and dismissal databank has been an on-going project for three years. A full-time librarian was appointed to run the routine work of retrospective data acquisition and inspection, and has been supported by several part-time students as database clerks. The ultimate goal of the project is to complete the retrospective process to include several political regimes related in people or land. Fig. 1 shows the evolution of five political regimes to be included in the databank. Currently, the project has concluded the second stage with complete coverage for the Taipei government (from present back to the year of 1949) on the island of Taiwan and has entered the third stage of retrospective data acquisition for the Nationalist government (from the year of 1949 back to 1925) in mainland China.

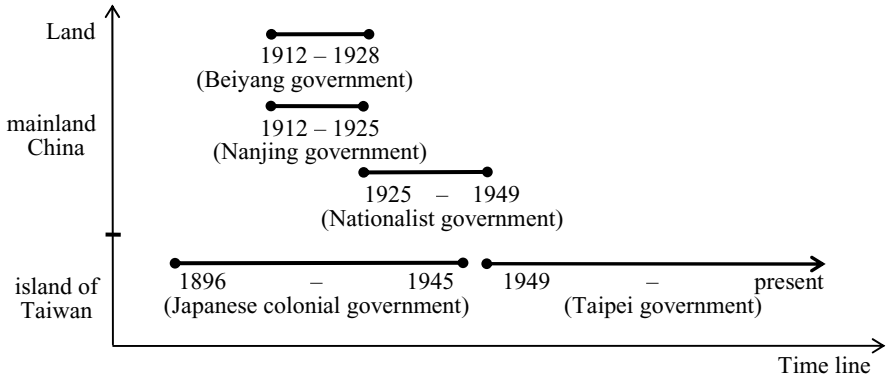


Fig. 1. Evolution of political regimes related in people or land

Three other political regimes are planned to be included in future extension. Nanjing government and Beiyang government both ruled a part of China in the warlord era. Nanjing government took control of the southern China between 1912-25, while the northern China was reigned by Beiyang government between 1912-28. The island of Taiwan was once a Japanese colony between 1896 and 1945. Government gazettes of these political regimes are all available for data acquisition. A further retrospective coverage of the island of Taiwan governance under Qing Dynasty China before 1896 is also considered when relevant documents are identified.

(姓名: 蕭萬長)  
 (person name: Vicent Siew) (Total: 32 items, 1-10 shown, 0.0018.. sec)  
 共 32 筆, 本頁顯示第 1 到 10 筆 (0.00183916091919秒)

清除 搜索歷史 (0) (姓名: 蕭萬長) | 每頁顯示 10 筆 (自第 1 筆) 顯示 | 依 公告日期, 遞增 | 排序

(prev. page)(next page)  
 << 上一頁 >> 下一頁 本頁全選 清除勾選 顯示勾選 詳目顯示 簡目輸出 名人錄

(seq. no.)	(person name)	公告日期	出刊日期	(change type)	(unit name+title)	(doc. image links)
序號	姓名			異動	單位職等	影像 圖點
<input type="checkbox"/> 1	蕭萬長	53/5/22	53/4/29	任命方式: 試用	外交部科員	📄 ⬆️ 🗑️
<input type="checkbox"/> 2	蕭萬長	54/1/5	53/12/22	任命方式: 任命	行政院外交部科員	📄 🗑️
<input type="checkbox"/> 3	蕭萬長	55/5/13	55/4/18	免職原因: 另有任用	行政院外交部科員	📄 🗑️
<input type="checkbox"/> 4	蕭萬長	55/7/12	55/6/15	任命方式: 任命	中華民國駐吉隆坡領事館副領事	📄 🗑️
<input type="checkbox"/> 5	蕭萬長	58/8/27	58/5/15	免職原因: 另有任用	駐吉隆坡領事館副領事	📄 🗑️
<input type="checkbox"/> 6	蕭萬長	58/8/29	58/8/29	任命方式: 任命	駐吉隆坡領事館副領事	📄 🗑️
<input type="checkbox"/> 7	蕭萬長	59/10/2	59/10/2	免職原因: 另有任用	駐吉隆坡領事館副領事	📄 🗑️
<input type="checkbox"/> 8	蕭萬長	60/01/01	59/12/31	任命方式: 任命	駐吉隆坡領事館領事	📄 🗑️
<input type="checkbox"/> 9	蕭萬長	61/4/14	61/4/14	免職原因: 另有任用	駐吉隆坡領事館領事	📄 🗑️
<input type="checkbox"/> 10	蕭萬長	61/9/6	61/9/6	任命方式: 任命	外交部科員	📄 🗑️ ⬇️

(order date: yy/mm/dd) (error report)

Fig. 2. Career path of a named official

The databank currently contains data for official appointment and dismissal dated from 2008 back to 1936. The total number of entity instances is approximately six hundred thousands, which includes two hundred thousands persons, twenty thousands

government units, and two thousands job titles. A query interface has been developed for the databank that supports query with Boolean combination of attribute values. As mentioned earlier, a named official's career path can be easily retrieved by a query specifying the official's name as the value of the attribute "person name". Fig. 2 reveals the career path of current Vice President, Vincent Siew, who started his government work as a staff of the ministry of foreign affairs in 1964. Each appointment/dismissal instance is also linked to the document image where the order appeared and provides a chance for users to detect and report errors.

An essential implication of recording personnel change information in a structured collection of data sets is the ability to provide higher data granularity for more effective information use. In other words, the databank provides fine-grained information that can be integrated and analyzed as needed to reveal unknown aspects and trends of subject matters. Such a data analysis utility is not possible with document retrieval.

## 4 Conclusion

This paper reports the development of government official appointment and dismissal databank. The databank provides higher data granularity of government personnel information and enables dynamic data exploration that helps discover facts and knowledge with multi-faceted analytical investigation. The databank development work provides an example of how digital collection of document images can be extended to create new values for users with more effective information use. In particular, structured textual data of great value to special subject domains can be extracted from historical documents and other primary source collections to facilitate creative research in humanities and social sciences. With direct access to the collected primary source materials and the expertise of information organization, libraries are in a unique position to play the role of data producer and develop textual databanks that help reveal information and knowledge hidden within volumes of documents. The proposed direction not only upholds libraries' core value of information service but also creates a valuable niche for libraries in the shifting information landscape.

## References

1. Brogan, M.: A Survey of Digital Library Aggregation Services. Technical report, The Digital Library Federation (2003)
2. ApSimon, J., and NDAC Working Group: Building Infrastructure for Access to and Preservation of Research Data. Technical report, Social Sciences and Humanities Research Council of Canada (2002)
3. Benson, D.A., Karsch-Mizrachi, I., et al.: GenBank. *Nucleic Acids Research* 28(1), 15–18 (2000)
4. Kavaliunas, J.: Census 2000 Data Products. *Government Information Quarterly* 17(2), 209–222 (2000)

# An Integrated Approach for Smart Digital Preservation System Based on Web Service

Chao Li<sup>1,\*</sup>, Ningning Ma<sup>2</sup>, Chun-Xiao Xing<sup>1</sup>, and Airong Jiang<sup>2</sup>

<sup>1</sup> Research Institute of Information Technology, Tsinghua University, Beijing, 100084

<sup>2</sup> Tsinghua University Library, Beijing 100084

lichao00@tsinghua.org.cn

**Abstract.** There are massive digital resources in digital libraries and other organizations. While hardware and software used to read digital data become obsolescent soon, and a lot of information in these unreadable data is so valuable that it is urgent to preserve digital resources for long-term utilization. A variety of tools or systems can solve part of the problem, but most of them are isolated. This paper describes an integrated and flexible digital preservation system AOMS which leverages the existing tools and services. AOMS considers not only file formats and versions, but also storage media, hardware, software, OS, and so forth as digital preservation risks, so the corresponding risks may be interlaced. Consequently, AOMS also helps collection managers schedule the preservation actions effectively and efficiently.

**Keywords:** digital preservation, XML, integration, Web Service.

## 1 Introduction

Nowadays, most digital libraries collect massive digital objects of all kinds, and store them in diverse hardware with different software for editing and reading. With the rapid development of information technology, these organizations face a problem of how to deal with the obsolescence of formats, software and hardware.

Since the 1980s, a good many work has been under way, however, most of them offer only one piece in the complete preservation puzzle. For example, the PRONOM [1] project and VersionTracker [2] are software and format registries. OCLC's INFORM [3] project and Cornell's VRC project [4] provide risk measurement and notification services.

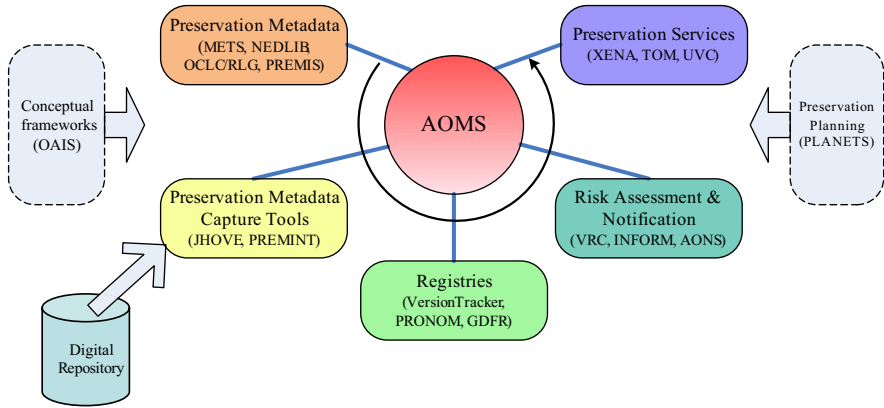
This paper proposes an integrated and flexible automatic obsolescence management system based on Web Service—AOMS (a subsystem of our THDP-Tsinghua Digital Preservation system), which makes good use of existing tools and services to help collection managers monitor the obsolescence status related to digital preservation.

---

\* This work is supported by the Key Technology R&D Program of China under Grant No. 2006BAH02A12 and the National High-tech R&D Program of China under Grant No. 2006AA010101.

## 2 Objectives

Our aim is to leverage the efforts of the different preservation initiatives by integrating the growing range of methods, tools and services being developed into a flexible, extensible Web-based framework.



**Fig. 1.** A flexible and extensible Web-based framework

PANIC [5] and AONSII [6] have the similar ideas with us, but they only focus on the risk of file format/version, and are based on Semantic (Ontology) Web Service.

We call AOMS “an integrated approach for smart digital preservation system based on Web Service” owing to the following features:

- 1) Practical preservation service discovery and description mechanism: AOMS is based on web service rather than semantic web service.
- 2) Comprehensive risk notification: AOMS considers and manages the risks of not only file formats/versions obsolescence, but also storage media, hardware, software, OS, etc.
- 3) Aid in scheduling on preservation actions: Consequently, the corresponding risks may be interlaced, and AOMS helps collection managers schedule the preservation actions effectively and efficiently. Of course, fewer steps and less cost would be preferred.

## 3 System Architecture

The AOMS system comprises nine main software components that are developed to support the following four steps in the overall preservation process:

**Preservation Metadata Capture:** This comprises tools that enable the generation of preservation metadata. Details of the metadata input tools are provided in the next section. The preservation metadata is saved following an extended METS schema.

**Risk Detection and Notification:** This component periodically compares preservation metadata about software and formats with corresponding registries that store information about the latest available authoring, rendering or viewing software and recommended formats. When any of the risks are detected, a notification will be sent to the relevant agent (human or software).

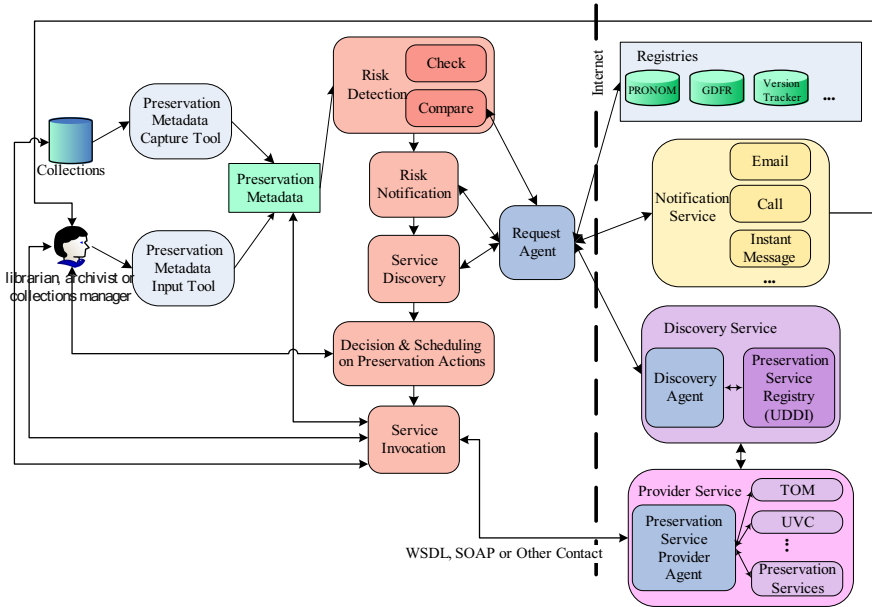


Fig. 2. Architecture of AOMS

**Decision and Scheduling on Preservation Actions:** When all kinds of risk factors are considered, the corresponding risks may be interlaced and aids are necessary for collection managers to decide the order of preservation actions. Optimum action orders should be economical and efficient.

**Preservation Service Discovery and Invocation:** When preservation action is needed, the collection managers specify the preservation service requirements. Discovery Agent then discovers appropriate preservation services by matching the requirements against descriptions of available services. Collection managers can choose from the ranked list of services retrieved. Service Invocation Agents then invoke the optimum preservation services and update the preservation metadata.

## 4 System Implementation

According to the architecture above, the implementation plan of AOMS is as follows:

**Table 1.** The implementation plan of AOMS

	manually input	PREMINT [8]
Risk Detection and Notification	registries	Formats: GDFR or PRONOM.
		Software: VersionTracker Web site
		Hardware/OS: to be built by ourselves
	Recommended Formats: RLG or the Library of Congress.	
	check	to be developed by ourselves
	comparing and notification	Quantitative risk assessment methodologies such as VRC , INFORM or AONSII could easily be incorporated to help triggering this notification.
Decision and Scheduling on Preservation Actions	In the case of several different risks interlace, all services for all risks should be listed at the same time (self-developed)	
	Aids (e.g. the ranked list of risks and corresponding services) are necessary for collection managers to decide the order of preservation actions. (self-developed)	
Preservation Service Discovery and Invocation	Make preservation software available as Web services and describe them using WSDL, and other preservation services available as Yellow Page services and describe them using PSD XML schema designed for AOMS.	
	Discovery agent discovers proper services and service providers. UDDI [9] registries are used to advertise available services.	
	Choose and invoke a particular preservation action automatically or manually.	

## 5 Conclusion

This paper proposes an integrated and flexible automatic obsolescence management system based on Web Service—AOMS. In this area, for the first time, AOMS comprehensively considers not only file formats and versions, but also storage media, hardware, software, OS, etc. as digital preservation risks, so the corresponding risks may be interlaced. Consequently, AOMS also helps collection managers schedule the preservation actions effectively and efficiently.

## References

1. PRONOM, <http://www.nationalarchives.gov.uk/pronom>
2. VersionTracker: <http://www.versiontracker.com>
3. Stanescu, A.: Assessing the durability of formats in a digital preservation environment: the INFORM methodology. *D-Lib Mag.* 10(11) (2004), <http://www.dlib.org/dlib/november04/stanescu/11stanescu.html>
4. Virtual Remote Control (VRC), <http://irisresearch.library.cornell.edu/VRC/>
5. PANIC: <http://www.metadata.net/panic/>
6. AONS, <http://www.apsr.edu.au/aons>
7. JHOVE, <http://hul.harvard.edu/jhove/jhove.html>
8. PREMINT Preservation Metadata Input Tool: <http://maenad.dstc.edu.au:8080/premint/index.jsp>
9. UDDI: <http://www.uddi.org/>

# Personalized Digital Library Framework Based on Service Oriented Architecture

Li Dong<sup>1,2</sup>, Chun-Xiao Xing<sup>3</sup>, Jin Lin<sup>4</sup>, and Kehong Wang<sup>1</sup>

<sup>1</sup> Department of Computer Science and Technology, Tsinghua University,  
100084 Beijing, P.R. China  
dongli@lib.tsinghua.edu.cn, wkh-dcs@tsinghua.edu.cn

<sup>2</sup> System Division of Library, Tsinghua University,  
100084 Beijing, P.R. China  
dongli@lib.tsinghua.edu.cn

<sup>3</sup> Research Institute of Information Technology, Tsinghua University,  
100084 Beijing, P.R. China  
xingcx@tsinghua.edu.cn

<sup>4</sup> Department of Electrical Engineering, Tsinghua University,  
100084 Beijing, P.R. China  
linjin03@mails.tsinghua.edu.cn

**Abstract.** In order to solve the conformity of personalized services with legacy systems, we design and implement a personalized digital library framework based on Service Oriented Architecture (SOA). In this paper, we introduce the framework development as following: Analyzing current systems and their workflow with SOA design pattern, and extracting needed personalized service interface; implement the Enterprise components and data interface of those service interface according to SOA standard; composing the service components according to workflow, so that to build up the whole framework.

**Keywords:** Service Oriented Architecture (SOA), personalized service.

## 1 Introduction

As digital libraries become commonplace, as their contents and services become more varied, and as their patrons become more experienced with computer technology, people expect more sophisticated services from their digital libraries. Thus personalized information service is gaining more and more focus in digital library development. Many applications of personalization in digital libraries has been built up such as MyLibrary, the ACM Digital Library, and the SpringerLink[1]. In a practical digital library, providing personalized services should cope with the distributed systems problems, such as: digital documents exist in distributed databases; legacy systems have different user information databases. In order to solve the conformity of personalized services with legacy systems, we design and implement a personalized digital library framework based on Service Oriented Architecture (SOA).



## 2 Functions of the Framework

The framework named THU-MyLibrary includes five main function modules, which can be described as figure 1.

(1) User information module

User information module stores the registration, accessing histories, resource evaluations and preferences for each user, so the system can use those information above to implement personalized search and recommendation calculations.

(2) Document model module

Document model module is a group of documents. Full-text and classified index is built up in this module to improve the effectiveness and meet different requirements.

(3) User interests management module

User interest management module processes 3 parts of information: user search histories, resources evaluation and tag data. All the information is stored in database which can be utilized to maintain the interest model for each user.

(4) Personalized search module

Personalized search module uses Nutch and Lucene to capture web documents then index them. After system sorts the search result by collaborative filtering algorithm, different search result will be achieved for different user.

(5) Personalized recommendation module

Personalized recommendation module can recommend both tags and web documents. Tag recommendation module manages users' tags and allow user to tag search results or resources browsed. User tag has been proved to be an excellent resource management method and paid more and more attention to. Web documents recommendation module uses collaborative filtering algorithm and content-based filters to recommend resources to users.

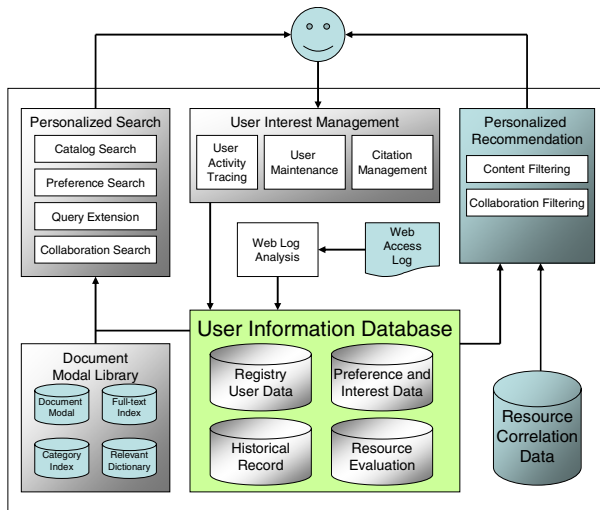


Fig. 1. Function modules of THU-MyLibrary

### 3 Building Up the Framework

Service Oriented Modeling and Architecture(SOMA), which is recommended by IBM to model service and design architecture, is used in this paper to extract personalized services of this system. According to function design of THU-MyLibrary, service candidates need to be developed as table 1:

**Table 1.** Personalized service to be developed

Service candidate	Function design
User Information service	<ul style="list-style-type: none"> <li>* Acquiring user identifications.</li> <li>* Acquiring user authorizations.</li> </ul>
Document information service	<ul style="list-style-type: none"> <li>* Acquiring document text.</li> <li>* Analyzing document text, classifying and clustering the document.</li> </ul>
User interests acquisition service	<ul style="list-style-type: none"> <li>* Acquiring the users' information.</li> <li>* Analyze the identity information of user automatically.</li> <li>* Calculate the users' interest degree for a special document, based on user's browsing history, searching history.</li> <li>* Clustering users with similar interest.</li> </ul>
Personalized search service	<ul style="list-style-type: none"> <li>* Analyses on the user's interest and the history of search results use, then automatically reorders search results.</li> <li>* Analyze on search results, and then automatic classifies search results.</li> <li>* Analyses on similar search entry, then according to the user's interest and the history of search results use, automatically expand the search entry.</li> <li>* Analyses on similar user's search history, then automatically expand the entry.</li> </ul>
Personalized recommendation service	<ul style="list-style-type: none"> <li>* According to user interests, system automatically recommends the hot documents for the corresponding categories.</li> <li>* According to user interests, system automatically recommends the hottest documents read by others with the same interests.</li> <li>* Actively recommends documents according to user's favorite way.</li> </ul>

All key services are published by Web Service. According to the business framework references of WebSphere, all the services will be categorized into different business components which are placed on Enterprise Service Bus(ESB), ESB will provide directory routing service for all the components and services[2].

Next, we can implement the Enterprise components and data interface of those service interface according to SOA standard, all the components used in personalized system locate on service and business process choreography layers[2].

Finally, compose the service components according to workflow, so that to build up the whole framework.

## 4 Conclusions and Future Work

We have built up a prototype of a personalized digital library framework based on SOA, and have apply some service component into some existing self-developed systems in our university library. In the future, we plan to integrate those personalized services into more existing systems in our university library to verify our framework's effectiveness.

**Acknowledgements.** This work is supported by the Key Technology R&D Program of China under Grant No. 2006BAH02A12, and the National High-tech R&D Program of China under Grant No. 2006AA010101.

## References

1. Smeaton, A.F., Callan, J.: Personalisation and recommender systems in digital libraries. *International Journal of Digital Library* 5, 299–308 (2005)
2. IBM redbooks series: Getting Started with WebSphere Enterprise Service Bus V6, <http://www.redbooks.ibm.com/abstracts/sg247212.html?Open>

# Automatic Document Mapping and Relations Building Using Domain Ontology-Based Lexical Chains

Angrosh M.A. and Shalini R. Urs

International School of Information Management, University of Mysore  
{angrosh,shalini}@isim.ac.in

**Abstract.** The paper describes ontology-based information retrieval system developed from titles of Electronic Theses and Dissertations (ETDs) for Vidyanidhi Digital Library. The paper also advances the notion of domain ontology-based lexical chain for automatic classification and mapping of documents to the ontology.

## 1 Introduction

Ontology-based systems facilitate more powerful and interoperable information retrieval systems for Digital Libraries through declarative models that define and represent concepts, attributes and relationships. While ontology development plays a crucial role in ontology-based information systems, automatic mapping of digital objects to the ontology is essential for intelligent systems. Novel methods that extract domain knowledge from digital documents are important for automatic mapping of documents to domain ontologies. Lexical chain techniques can be used for mapping instance data to domain ontologies. In this context, the present paper is an effort to explore the use of lexical chain analysis for mapping documents to domain ontologies. To this end, the study uses the Indian Agricultural Research Ontology for deriving domain related lexical chain results.

The paper is organized as follows. Section 2 highlights few related works. Section 3 and 4 details the process of developing Ontology-based Information Retrieval System and its search features. Section 5 and 6 illustrate the use of lexical chains for automatic mapping of ETDs to the ontology and derivation of semantic relations for connecting terms of different classes in the ontology.

## 2 Related Work

Efforts are underway for developing ontology-based information retrieval systems for agriculture domain. The Food and Agriculture Organization of the United Nations (FAO) has made commendable contribution through the development of AGROVOC, a multilingual, structured and controlled vocabulary, which covers the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains [1]. Concomitantly, lexical chain techniques have been extensively used in the areas of Information Retrieval. For instance, Yu et al use lexical chain and structural features for

automatic text summarization [2] and Chen et al have performed multi-document summarization based on lexical chains [3]. The present study uses these techniques for automatic classification and mapping of documents to the domain ontology.

### **3 Ontology-Based Information Retrieval System for Vidyanidhi Digital Library**

The primary task in developing ontology based information system for agricultural ETDs is to develop a shared understanding of the agricultural domain, which captures knowledge represented in them. Focusing on the limited semantic information available, the study adopted the following methodology for developing the ontology: (1) Identify all keywords in the titles of agricultural ETDs. (2) Identify and define classes and subclasses to which these terms belong to and (3) Define relations binding individuals (terms) of different classes.

Each of the 2800 agricultural ETD titles was carefully analyzed to identify the possible keywords present in these titles. Upon identification of keywords, we identified the different classes and subclasses to which these keywords would belong. The classes and subclasses relationships were identified using agricultural handbooks and subject classification systems [4]. The methodology followed above resulted in development of Agricultural Research Ontology with finer semantic relations connecting individuals (terms) belonging to different classes, which was used for developing ontology-based information retrieval system for Vidyanidhi Digital Library (VDL). The technical details of the developed system are discussed in our earlier work [5].

### **4 Search Features of Ontology-Based Information Retrieval System**

The prominent search features of the ontology-based information retrieval system included the following:

#### **4.1 Simple Search**

The simple search facilitated performing search on terms present in the ontology, metadata of agricultural records and AGROVOC vocabulary.

#### **4.2 Taxonomic View**

The domain ontology is used to provide a taxonomic view and navigation search.

#### **4.3 Query Building Mechanisms**

The system also facilitates in extending the taxonomy based information search and retrieval for query building mechanisms. The Pellet Reasoner is employed for developing such query mechanisms, which primarily queries for relations binding individuals of different classes in the ontology [6].

An automatic method for identification of key individual terms of different classes in a document would not only facilitate mapping of documents to the ontology, but

would also aid in defining relations between the terms of different classes. With this objective, the technique of lexical chain indexing was examined for achieving automatic classification and mapping of documents to the ontology. The following section details the experiments carried out with lexical chain techniques.

## 5 Lexical Chain Indexing for Document Mapping

Lexical resources such as WordNet [7] and Roget's Thesaurus based Electronic Lexical Knowledge Base (ELKB) [8] have been extensively used for Text Summarization. Relying on rich data repositories, these models define explicit semantic relationships between word and word groups, which is used to decide on the semantic proximity of words. Lexical chains are primarily sequences of related words in a text that represent the same topic [9]. The process of lexical chain building involves (1) Selection of set of candidate words; (2) Identification of an appropriate chain based on a criterion among the members of the chain and (3) inclusion and updation of the term.

The process of selecting a set of candidate words involves removal of stop words and high frequency words that appear often in the text. The Standard stop lists are used for removing the stop words and high frequency words. For each of the candidate word, an appropriate chain based on a criterion among the members of the chain is identified. One of the prominent thesaural relations as identified by Morris and Hirst is the presence of semantic relatedness due to the words falling in the same group [9]. With the availability of domain ontologies, this relation can be used as a criterion for inclusion of a candidate word in a chain. For instance, the knowledge representation structure of the Agricultural Research Ontology is used for deriving lexical chains for ETDs in the agricultural domain.

The next step involves inclusion of the term in the chain and updates the chain accordingly. The identification of semantic relatedness due to the grouping of words under the same category facilitates addition of terms in the chain to result in a lexical chain grouping all related terms of a category in a single chain. The derived lexical chain primarily represents key individual terms and classes defined in the ontology. The resulting lexical chains are used to identify key individual terms of different classes for mapping the Agricultural ETDs. Further the chain is used for defining relations between terms of different classes.

## 6 Results of Ontology-Based Lexical Chain Indexing

National Bank for Agriculture and Rural Development (NABARD) is a development bank with a mandate to promote agriculture and rural development. Forestry is one of the activities in rural areas and NABARD has promoted a number of projects for rehabilitation of degraded forests through people's participation and also through Forest Development Corporations (FDCs). Out of India's 63.73 million ha recorded forests, 31 mha are suffering from degradation. The Xth Plan approach paper of GOI envisages bringing 33% of the land under tree cover by 2012. Hence a massive programme of rehabilitation of degraded lands has been planned.

Source: Food and Agricultural Organization.

<http://www.fao.org/DOCREP/ARTICLE/WFC/XII/0088-C1.HTM>

For the sample text above, the lexical chain technique, based on the ontology representation results in the following sequences of terms

```
C:\LCI>java LexicalChain -f sampledata.txt
Building lexical chains: -f Resources/sampledata.txt -s Elkb
degraded, forests, Forest, forests, degraded, degraded, forests,
forests, Forest, Management, degraded, forests, JFM, JFM, forests,
degraded, forests [score: 17.0, class: 4]
NABARD, NABARD, NABARD [score: 3.0, class: 6]
Sustainable, Development, Sustainable, Development [score: 4.0,
class: 5]
```

The score of the lexical chain primarily measures the number of times the related term of a particular topic appear in the text and the class number represent the particular class these terms belongs to. These results are effectively used to achieve (1) Automatic Classification of Documents in the Ontology (2) Automatic mapping of documents to related individuals and (3) Derive relations between different classes.

## 7 Conclusions

The present study successfully developed ontology-based information retrieval system based on Indian Agricultural Research ontology. The paper extends domain ontologies for deriving ontology-based lexical chains. We also illustrated how domain ontology based lexical chains can be used for automatic classification and mapping of documents to ontologies. Further the challenging task of defining semantic relations for connecting terms of different classes in the ontology can also be achieved through lexical chains. We conclude that the present line of research is promising and can be pursued further for building robust systems on the principles discussed above.

## References

1. AGROVOC Thesaurus, FAO (2007), [http://www.fao.org/aims/ag\\_intro.htm](http://www.fao.org/aims/ag_intro.htm)
2. Yu, L., Ma, J., Ren, F., Kuroiwa, S.: Automatic Text Summarization based on Lexical Chains and Structural Features, pp. 574–578. IEEE Computer Society, Los Alamitos (2007)
3. Chen, Y., Wang, X., Liu, B.: Multi-Document Summarization based on Lexical Chains, pp. 1937–1942. IEEE Computer Society, Los Alamitos (2005)
4. Indian Council for Agricultural Research: Handbook of Agriculture. Indian Council of Agricultural Research, New Delhi (2006)
5. Angrosh, M.A., Urs, S.R.: Development of Indian Agricultural Research Ontology: Semantic rich relations based information retrieval system for Vidyanidhi Digital Library. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 400–409. Springer, Heidelberg (2007)
6. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL Reasoner. Web Semantics: Science, Services and Agents on the World Wide Web 5, 51–53 (2007)
7. Wordnet: A lexical database for the English language (2006), <http://wordnet.princeton.edu/>
8. Roget's Thesaurus – Electronic Lexical Knowledge Base (ELKB), <http://www.nzdl.org/ELKB/>
9. Morris, J., Hirst, G.: The Subjectivity of Lexical Cohesion in Text. In: Shanahan, J., et al. (eds.) Computing Attitude and Affect in Text, pp. 41–48. Springer, Berlin (2004)

# A Paper Recommender for Scientific Literatures Based on Semantic Concept Similarity

Ming Zhang, Weichun Wang, and Xiaoming Li

School of Electronics Engineering and Computer Science  
Peking University, P.R. China

mzhang@net.pku.edu.cn, wangwchpku@gmail.com, lxm@pku.edu.cn

**Abstract.** Recently, collaborative tagging has become more and more popular in the Web2.0 community, since tags in these Web2.0 systems reflect the specific content features of the resources. This paper presents a recommender for scientific literatures based on semantic concept similarity computed from the collaborative tags. User profiles and item profiles are presented by these semantic concepts, and neighbor users are selected using collaborative filtering. Then, content-based filtering approach is used to generate recommendation list from the papers these neighbor users tagged. The evaluation is carried out on a dataset crawled from CiteULike, with satisfied experiment results.

**Keywords:** Recommender, Semantic concept, Web2.0, Collaborative tag.

## 1 Introduction

The paper recommender systems are emerging with the explosive growth of the WWW. McNee et al. mapped the Web of citations between papers into the user-item rating matrix where the paper “votes” for the citations it references [1]. Pennock et al. proposed a collaborative approach for recommending articles in CiteSeer [2].

In this paper, we designed and implemented a paper recommender based on semantic concept similarity computed from the collaborative tags. Section 2 describes the modeling of the concept graph; Section 3 introduces the key steps of our semantic-based hybrid recommendation system; Section 4 experimentally evaluates the algorithm on a dataset crawled from CiteULike, and summarizes this paper.

## 2 Profile Modeling Based on Concept Graph

First, we build a concept graph with the semantic concepts derived from the collaborative tags to semantically extend these profiles [3]. User and paper profiles are then represented by the semantic concepts. The concept graph is a quintuple:

$$(C, V, R, rel, map). \quad (1)$$

where: (1)  $C$  is a set of semantic concepts;

(2)  $V$  is a set of terms assigned to semantic concepts in  $C$ , which is a set of strings;



(3)  $R$  is the relevant relationship between semantic concepts in  $C$ ;

(4)  $rel : R \rightarrow C \times C \times VAL$  is a function, which means each pair of semantic concepts has a relevant relationship, with  $VAL$  as its degree of relevance (similarity);

(5)  $\forall V_i \in V, C_i \in C, map : C_i \rightarrow V_i$  means  $V_i$  is one of terms used to represent  $C_i$ .

An  $m \times n$  matrix  $M$  is built as an association matrix, where  $m = |V|$  is the total number of tags and  $n$  the number of papers, respectively. Here  $M_{ij}$  denotes the association degree between the  $i^{\text{th}}$  tag and the  $j^{\text{th}}$  paper, defined as follows:

$$M_{ij} = C_{ij} \times \log(n |DOC(t_i)| / l). \quad (2)$$

where  $C_{ij}$  denotes the number of users who tagged the  $j^{\text{th}}$  paper with the  $i^{\text{th}}$  tag, and  $|DOC(t_i)|$  represents the number of papers tagged by the  $i^{\text{th}}$  tag  $t_i$ . Given the matrix  $M$ , the  $i^{\text{th}}$  tag can be represented as a row vector  $T_i = (M_{i1}, M_{i2}, \dots, M_{in})$  of  $M$ .

The semantic similarity between two tags are measured with cosine similarity:

$$Sim_{tag}(t_i, t_j) = \cos(T_i, T_j). \quad (3)$$

where  $T_i$  and  $T_j$  are row vectors corresponding to tag  $t_i$  and tag  $t_j$ , respectively.

Given the selected tag  $t_i$ , we choose a synonymic set of tags most related to  $t_i$ :

$$ST_{t_i} = \{t_j \mid t_j \text{ is similar to } t_i\}. \quad (4)$$

This tag set is generated using the following rules:

1.  $t_i$  should be among the  $N$  most similar tags related to  $t_i$ ;
2. The similarity should be larger than a threshold  $\theta$ .

According to the experiments, the  $N$  and  $\theta$  in the rules are set to 4 and 0.7, respectively. Then, a semantic concept  $C_{t_i}$  for  $t_i$  is represented by the following set:

$$C_{t_i} = ST_{t_i} \cup \{t_i\}. \quad (5)$$

Given two semantic concepts  $C_i = \{t_{i1}, t_{i2}, \dots, t_{iu}\}$  and  $C_j = \{t_{j1}, t_{j2}, \dots, t_{jv}\}$ , we use the following algorithm to calculate the similarity between them.

Algorithm 1. Computation of Concept Similarity

```

input  $C_i \bullet \{t_{i1}, t_{i2}, \dots, t_{iu}\}$  and  $C_j \bullet \{t_{j1}, t_{j2}, \dots, t_{jv}\}$ 
int sum = 0.0;
int count = 0;
do
  search tags  $t_{ix}$  and  $t_{jy}$  from  $C_i$  and  $C_j$ , respectively,
  which maximize  $Sim_{tag}(t_{ix}, t_{jy})$ 
  sum = sum +  $Sim_{tag}(t_{ix}, t_{jy})$ 
  count = count + 1
   $C_i = C_i - \{t_{ix}\}$ ;
   $C_j = C_j - \{t_{jy}\}$ ;
while  $|C_i| > 0$  and  $|C_j| > 0$ ;
output sum / count.
```

To simplify the graph, we ignore the edges with similarity lower than a threshold. We define user profile and paper profile uniformly as a set of semantic concepts:

$$C_{t_i} = ST_{t_i} \cup \{t_i\} . \tag{6}$$

For the user profile,  $\{ t_1, t_2, \dots, t_L \}$  is the set of tags the user tagged; while for the paper profile,  $\{ t_1, t_2, \dots, L \}$  is the set of tags annotated on this paper. For measuring the similarity between two profiles, we apply an algorithm similar to Algorithm 1.

### 3 The Hybrid Recommender Based on Semantic Concepts

Given a target user, collaborative filtering is employed to find neighbor users for this user based on the similarity between user and paper profiles as follows:

$$Sim_{user}(u_1, u_2) = \frac{1}{N_S - N_C + 1} \times \frac{\sum_{j \in U_C} \min(Sim_{profile}(u_1, j), Sim_{profile}(u_2, j))}{\sum_{j \in U} \max(Sim_{profile}(u_1, j), Sim_{profile}(u_2, j))} . \tag{7}$$

- where: (1)  $U_C$  denotes the intersection of paper sets user  $u_1$  and  $u_2$  tagged;
- (2)  $U_S$  denotes the union of paper sets user  $u_1$  and  $u_2$  tagged;
- (3)  $N_S = |U_S|$ ,  $N_C = |U_C|$ ;
- (4)  $Sim_{profile}(u_i, j)$  denoted the similarity between the profiles of  $u_i$ ' and the  $j^{th}$  paper. User  $u_i$ 's neighbor users are with similarities larger than a threshold as follows:

$$S_{u_i} = \{u_j \mid Sim_{user}(u_i, u_j) > \delta\} . \tag{8}$$

The candidate papers for user  $u_i$  are the papers the neighbor users tagged, eliminating the papers that user  $u_i$  tagged, denoted as follows:

$$R_{u_i} = \bigcup_{u_j \in S_{u_i}} K_{u_j} - K_{u_i} . \tag{9}$$

where  $K_{u_i}$  denotes the set of papers which user  $u_i$  tagged.

We define the similarity between two papers as a combination of the weighted mean of their text similarity and their semantic concept-based profile similarity:

$$Sim_{doc}(j, j') = \alpha \cos(V_j, V_{j'}) + \beta \cdot Sim_{profile}(j, j') . \tag{10}$$

where  $V_j$  denotes the row vector of the  $j$ th document. According to our experiments, the value of  $\alpha$  and  $\beta$  are set to 0.6 and 0.4, respectively.

### 4 Experiment and Discussion

We created a paper dataset extracted from CiteULike (<http://www.citeulike.org/>), which contains 220,723 papers, 6,800 users who tagged 70,796 tags on these papers. We divided the dataset into training/test datasets at a 90%/10% ratio.

The experiment evaluated the influence caused by the size of neighbor user set, shown in Figure 1. Result shows that the hit percentage [1] increased when the size of neighbor user set was larger. Our approach gave the removed papers high ranks because the hit percentages from top-20 to top-50 have little difference.

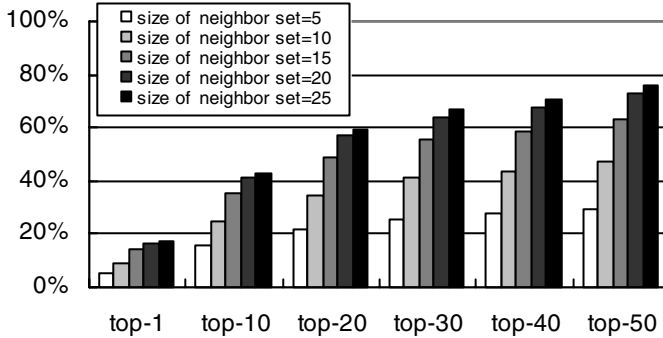


Fig. 1. Hit percentages of different sizes of neighbor user set

In the future, we will use clustering technique to group the users before to improve the quality of neighbor users in PKUSpace (<http://fusion.grids.cn/PKUSpace>) [4].

## Acknowledgement

This study is supported by the Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20070001073), HP Labs China, and the National Natural Science Foundation of China (Grant Nos. 90412010 and 60773162).

## References

1. McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A., Riedl, J.: On the Recommending of Citations for Research Papers. In: Proceedings of the ACM 2002 Conference on Computer Supported Cooperative Work (CSCW 2002), New Orleans, LA, pp. 116–125 (2002)
2. Pennock, D., Horvitz, E., Lawrence, S., Giles, C.: Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach. In: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, San Francisco, June 30 - July 3, pp. 473–480. Morgan Kaufmann, San Francisco (2002)
3. Li, R., Bao, S., Fei, B., Su, Z., Yu, Y.: Towards effective browsing of large scale social annotations. In: WWW 2007: Proceedings of the 16th international conference on World Wide Web, pp. 943–952. ACM Press, New York (2007)
4. Yin, P., Zhang, M., Li, X.M.: Recommending Scientific Literatures in a Collaborative Tagging Environment. In: Goh, D.H.-L., Cao, T.H., Sølvsberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 478–481. Springer, Heidelberg (2007)

# Network of Scholarship: Uncovering the Structure of Digital Library Author Community

Monica Sharma and Shalini R. Urs

University of Mysore, Mysore, India  
{monica,shalini}@isim.ac.in

**Abstract.** The present study is a part of research, studying the structure and composition of Digital Library (DL) community as expressed in different communication platforms. This study analyzes co authorship network of DL research community using social network analysis metrics – degree and betweenness centrality. D-Lib magazine and International Journal on Digital Libraries form the basis of data source. Our results are compared with that of the Liu et al to understand network structure of collaboration in journals and conferences.

## 1 Introduction

Research efforts have consistently shown that it is possible to unravel various structural features of research communities by studying the network of communication patterns using social network analysis (SNA). The patterns and structural features of academic communities may be analyzed by using various SNA measures such as centrality, geodesic, components etc. Academic networks have been examined extensively to determine the structure of scientific collaboration and scholarship [1, 2, 3]. Co-authorship networks are of great interest to scientists as authors as individuals and as a group and their relationship get influenced by the network they form [1].

## 2 Motivation for the Study

Present study is part of a larger research focusing on understanding the structure and composition of Digital Library (DL) community. Studies on authors of journals and conferences and an online database have been carried out by the researchers to understand the composition of DL domain [4,5]. Liu et al studied the co- authorship network in DL as in Joint Conference on Digital Libraries (JCDL) [6]. The objective of the present work is to expand our previous study by examining the network characteristics of DL as reflected in author collaboration in journals and also to look for similarities and dissimilarities of co authorship patterns of conferences and journals. We identified two highly popular journals in DL i.e. D-Lib Magazine and International Journal on Digital Libraries (JODL). D-Lib is an electronic journal, which tops the magazines and journals in “digital Libraries” in terms of publication count, as in Thomson Reuters Web of Knowledge. JODL is a top ranked peer reviewed journal covering research on DL domain.

### 3 DL Community: Who and How Big?

Stories (1995-Aug2000) and articles (Sep2000-2008) amounting to 599 in D-Lib and articles amounting to 196 from volume 1-8 (1997-2008) in the case of JODL formed the dataset for the present study.

The number of papers per author ranges from 1-18 with a highly significant number of authors contributing a single paper. Two authors have published more than 15 papers exhibiting power law tail. Liu et al report that 78% of the authors have published only one paper. Our study also shows that 81% of authors had contributed a single paper. We also analyzed the publication count of authors to identify the stars of DL community and compared with Liu et al study. Five authors - Edward A. Fox, Ian H. Witten, Carl Lagoze, Michael L. Nelson and Terence R. Smith find a place in both lists.

### 4 Network of Authorship

Based on the data from D-Lib and JODL, three co authorship networks were constructed, one each for D-Lib scientific community; JODL scientific community; and combined (D-Lib plus JODL) scientific community.

#### 4.1 Collaboration Network of D-Lib and JODL Community

As can be observed in figure 1(b) the D-Lib co authorship network is dense and connected. There are 1031 distinct authors who have contributed to the D-Lib. Unlike D-Lib network, JODL authorship network is more fragmented (Fig. 1 (a)). There are 591 distinct authors, almost half of the number of D-lib authors, who have contributed to the journal.

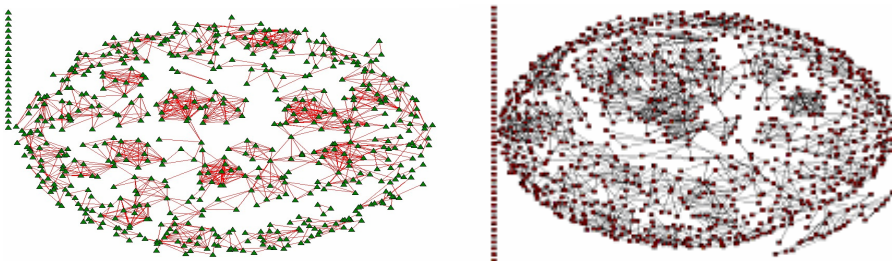
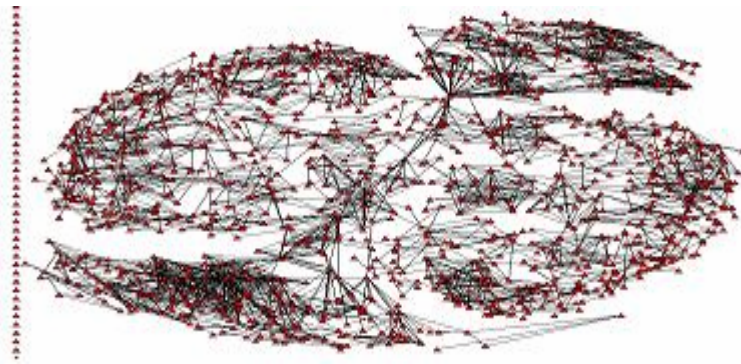


Fig. 1. Co authorship network of (a) JODL and (b) D-Lib research community

#### 4.2 Collaboration Network of Digital Library Community

To study the DL community, the data was combined, deduped and analyzed and a network of 1527 distinct authors was created. Our findings are similar to the observations made by Liu et al, that the DL network does not form a single connected graph with the largest component consisting of 38% of authors. The present study shows that the largest component consists of 32.4% of authors (Fig. 2).



**Fig. 2.** Co authorship network of D-Lib plus JODL research community

The JCDL network yielded average distance of 3.6[6]. Our previous study on co authorship network of journals and conferences yielded a geodesic of 3.1. Present study shows that average geodesic between the reachable pairs is 6.1 coinciding with the six degrees of separation concept. DL community of conferences is a well-connected network whereas that of journals is a fragmented network.

### 4.3 Degree Centrality and betweenness Centrality

Centrality describes the locations of individuals in terms of how close they are to the "centre" of the action in a network [7]. Centrality in terms of degree is the measure of direct connections a node has.

**Table 1.** Authors ranked according to the value of degree and betweenness centrality

Degree centrality		Betweenness centrality					
JODL plus DLIB	Liu etal JCDL	JODL and DLIB	Liu etal JCDL				
E. A. Fox	42	H. Chen	59	E. A. Fox	43211	H. Chen	89251
H. V. Sompel	32	E. A. Fox	55	R. W. Moore	41174	E. A. Fox	83164
C. Lagoze	31	T. R. Smith	31	S. L. Weibel	33522	J. Klavans	57423
C. Peters	28	C. Lagoze	31	T. Koch	29872	W. Y.Arms	52242
T. R. Smith	27	J. Klavans	27	C. Lagoze	26828	N. Wacholder	39226
P. Pagano	24	Z. Huang	26	H.-J. Schek	20143	C. N. Manning	38808
M. L. Nelson	24	G Marchionini	25	G. Janee	18602	D. M.Levy	35769
D. Castelli	24	W. Y.Arms	21	C. Peters	12859	A. P.Bishop	32280
G. Janee	23	R. Furuta	21	H. Chen	12817	T. D.Ng	30197
M. Freeston	22	L. Gravano	20	H. V. Sompel	12608	G. Marchionini	29594
S. Abiteboul	20	M. Freeston	19	T. R. Smith	9538	A. Hauptmann	29142
Y. E.Ioannidis	19	I. H.Witten	18	Y.E. Ioannidis	8878	C. C.Marshall	28587
C. Nikolaou	18	H.G. Molina	18	B. Ludaescher	8674	T. R. Smith	23692
S. Kapidakis	18	M. G.Christel	18	D. Castelli	8598	C. Lagoze	22193
H.-J. Schek	17	D. Millman	18	P. Pagano	8598	D. Bainbridge	21168

Four authors, Edward A. Fox, Carl Lagoze, Michael Freeston and Terence R. Smith are among top 15 authors with highest degree centrality in both lists reiterating their popularity in the network (Table 1).

Betweenness centrality can be regarded as a measure of the extent to which a node has control over information flowing between others. The node or author with highest "Betweenness" acts as a gatekeeper controlling the flow of resources between the nodes, which it connects. Edward A. Fox dominates the network in terms of betweenness centrality as well. Edward A. Fox, Terence R. Smith, Carl Lagoze and Hsinchun Chen are among the top 15 authors (Table 1). Our previous study on author network and that by Liu et al report that Hsinchun Chen and Edward A. Fox are the top 2 authors [4]. Fair amount of corroboration is observed in the results of our study and that of Liu et al work.

## 5 Conclusion

The individual co authorship network of D-Lib and JODL were analyzed. In addition, network of combined dataset was analyzed using various parameters. We have also tried to compare results of our study with the study carried out by Liu et al . The network of D-Lib and JODL is not dense and connected but to some extent exhibit same characteristics as of JCDL network. Even though there was significant difference in the metric values between our study and that of Liu, there are few authors who are consistent in both the studies and these authors dominate the collaboration network of digital library world. Edward A. Fox is one who plays a very important role as also shown in our earlier study [4]. We also found that the network is divided into many components, the giant component comprises of the majority of nodes or the authors. The average geodesic distance between any pair is 6.1, suggesting that six degrees of separation theory holds true for digital library world.

## References

1. Carolan, B.V., Natriello, G.: Data-mining journals and books: Using the science of networks to uncover the structure of the educational research community. *Educational Researcher* 34(3), 25–33 (2005)
2. Elmacioglu, E., Lee, D.: On six degrees of separation in dblp-db and more. *ACM SIGMOD Record* 34(2), 33–40 (2005)
3. Barabasi, A.L., et al.: Evolution of the social network of scientific collaborations. *Physica A* 311, 590–614 (2002)
4. Sharma, M., Urs, S.R.: Small world phenomenon and author collaboration: How small and connected is the digital library world? In: *Proceedings of the 10th International Conference on Asian Digital Libraries*, pp. 510–511. Springer, Vietnam (2007)
5. Sharma, M., Urs, S.R.: Mapping network structure of digital library research of CiteSeer database. In: *Proceedings of the Workshop on Mining Social Data, 18th European Conference on Artificial Intelligence, Greece*, pp. 6–10 (2008)
6. Liu, X., et al.: Co-authorship networks in the digital library research community. *Information Processing & Management* 41(6), 1462–1480 (2005)
7. Hanneman, R.A., Riddle, M.: *Introduction to Social Network Methods*. University of California, Riverside (2005)

# Understanding Collection Understanding with Collage

Sally Jo Cunningham and Erin Bennett

Dept. of Computer Science, Waikato University, Hamilton, New Zealand  
{sallyjo, ekb2}@cs.waikato.ac.nz

**Abstract.** A streaming collage visualization of images extracted from the collection's documents has been proposed as an effective tool for gaining a comprehensive overview of a digital library collection ("collection understanding"). A qualitative study provides insight into what users actually understand about a collection from viewing such a streaming collage.

**Keywords:** Streaming collage, qualitative research, collection visualization.

## 1 Introduction

To make an informed decision on whether to search/explore a given information source, the user must grasp the gist of a collection—have an impression for its focus, coverage, size, and so forth. This gut feeling is referred to as 'collection understanding'. The tools to support collection understanding include textual metadata summaries (eg, [5]), but research interest has been strongest in visualizations. One such visualization, streaming collage of images drawn from a collection, has been suggested to be particularly effective in developing a sense of the contents of a digital library ([1] [2] [3]). However, analysis of its effectiveness has been limited; the study reported in [2] focuses on usability of a specific collaging interface rather than the more general question of what sorts of insights users glean from viewing a collage. Further, earlier collection visualization studies ([1] [2] [3] [5]) examine collections in which the documents are primarily images, and the implementation of the collage interface allowed the user to explore the collection by search as well as by collage. Little is known about whether collage visualizations are useful for document collections containing both text and images, and what collection facets the 'understanding' afforded by streaming collage includes.

## 2 Description of Study

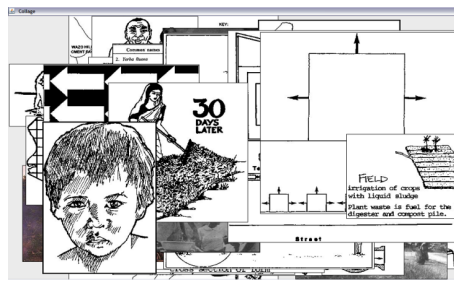
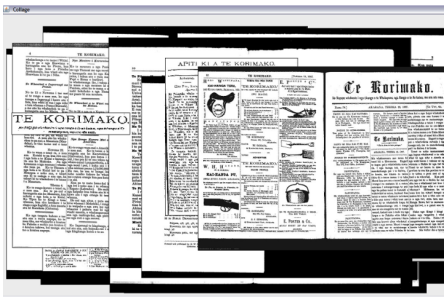
We investigate collection understanding of digital libraries whose documents include images, but the documents are primarily text. To this end, we conducted a qualitative study in which participants viewed streaming collages and described their 'understanding' of the collections from which the collages were drawn. As our goal is to explore the inferences about a document collection that can be reached solely through the lens of a streaming collage, no additional collection or document metadata was visible, and no search or additional browsing facilities were available. We follow the



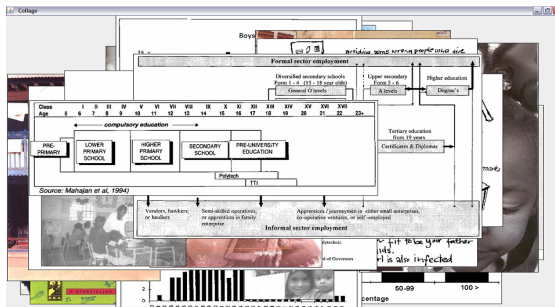
advice of Nielsen [4] that including more than a handful of participants (here, seven) does not necessarily gain greater insight into system usability issues.

**Participant demographics:** Seven subjects participated in the experiment: four male and three female, with four participants aged in their 20s, two in their 30s, and one over 40. All were fluent in English (four as native speakers).

**Collaging display:** Three document collections were visualized in this experiment (Figure 1). For each of the three test collections, all images were extracted from each document in the collection. Most images were left at their original size; the largest were resized to 800x800 so that the full image could be viewed without scrolling. A collage tool placed the images one by one randomly on the collage canvas, one image every five seconds.



**Fig. 1a.** Niuepepa (NIU): historic Maori news-papers; issues from 34 serials; 17,000 pages; 14K PDF page images **Fig. 1b.** East African Development (EAD): UNESCO human development documents for lay readers; 592 documents; 55K pages; 12K images



**Fig. 1c.** Research Education Development (RED): education and training in developing countries; 56 documents; 50K pages; 17K images

**Study procedure:** Sessions were audio recorded. Each participant viewed the three collection collages in the same order: NIU, EAD, RED. The participant viewed a collection collage for up to 10 minutes, or until the participant indicated that s/he had a good understanding of the collection.

At that point, the participant was asked two deliberately broad questions to elicit their initial understanding of the collection: *Tell me about any discoveries you've made about this collection; Was there was anything significant about this collection that caught your attention?* More specific follow-up questions probed their opinions about the topics of the collection, its probable size, and document format/media.

The researcher then described the collection. The participant was then asked how well they thought the collage had represented the collection, and how difficult they felt it to be to get a feeling for this collection through the collage visualization.

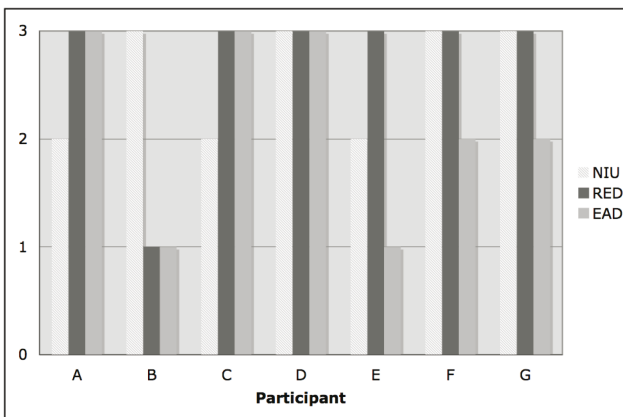
**Table 1.** Sample inferences about collection topics

NIU	– historic journals of New Zealand + large number of newspapers covering a long period
RED	+ education and third world countries – demography of third world countries; South America
EAD	+ the circle of life in South Africa, animals & plants + horticulture and agriculture research in Africa – agriculture and horticulture, worldwide

### 3 Results

The participants viewed the collages for a minimum of two minutes (participant E, RED) and a maximum of 10 minutes. Average viewing times per collection were: NIU, 5.9 minutes; RED, 5.29 minutes; EAD, 6 minutes.

Table 1 presents sample correct inferences (+) and incorrect or partially correct inferences (-) from participants when describing the collections. A set of keywords describing the major themes of each collection was drawn from each collection's description. A three-point scale was used to measure the extent to which a participant understood a collection's topic: *Success* (mentioning all of the keywords or closely related concepts); *Moderate Success* (mentioning most of the keywords or closely related concepts); *Failure* (mentioning few or none of the keywords or closely related concepts). Topic understanding results are presented in Figure 2.



**Fig. 2.** Topic understanding (3= Success; 2 = Moderate Success; 1 = Failure)

The participants were best able to come to an accurate understanding of a collection when they could draw upon previous exposure to document formats (eg, research articles, printed newspapers) and imagery (eg cultural and geographic references in photos) similar to those in the collection. Clearly collaging would be markedly less useful in understanding a collection covering a domain that is novel and unfamiliar to the user. Interestingly, many of the inferences were based on a familiarity with conventions of print publishing—recognizing newspaper formats and research book/journal covers, for example. When conventions differed from their own experiences, participants became confused; for example, the historic newspaper collection posed difficulties to one participant who felt that the older newspaper formats resembled journals. Participants also pointed out that random sampling of collection images may not present a representative view over a short time period.

## 4 Conclusions

These results suggest that streaming collage can be an effective, albeit time-consuming, technique for grasping the gist of a collection, if the user has some previous familiarity with its domain or publishing formats, and under the condition that the user focuses on the collage. Ambient collage displays are proposed in [2] to promote incidental learning while interacting with a collection (or possibly during background viewing while engaging in other tasks). Given that incidental learning extends time requirements, the viewing times reported in our study indicate that understanding through ambient collaging may be too time inefficient for practical use.

The collages in this present study were image-only, were presented in their original sizes, and were randomly selected. Further research on collages that support more efficient learning could include incorporating text snippets; presenting images from a representative sample of collection sub-topics rather than randomly selecting them; and investigating optimal size and presentation speed for enhancing comprehension.

## References

1. Bainbridge, D., Cunningham, S.J., Downie, J.S.: Visual collaging of music in a digital library. In: *Procs. of the International Symposium on Music Information Retrieval (ISMIR 2004)*, Barcelona, October 2004, pp. 397–402 (2004)
2. Chang, M., Leggett, J.J., Furuta, R., Kerne, A.: Collection Understanding. In: *Procs. of JCDL 2004*, Tucson (AZ, USA), pp. 334–342 (2004)
3. Kerne, A., Koh, E., Dworraczyk, B., Mistrot, M., Choi, H., Smith, S., Graeber, R., Caruso, D., Webb, A., Hill, R., Albea, J.: 2006 cominFormation: A mixed-initiative system for representing collections as compositions of image and text surrogates. In: *Procs. of JCDL 2006*, Chapel Hill (NC, USA), pp. 11–20. ACM, New York (2006)
4. Nielsen, J., Landauer, T.K.: A mathematical model of the finding of usability problems. In: *Proceedings of INTERCHI 1993*, Amsterdam (Netherlands), pp. 206–213. ACM, New York (1993)
5. Ong, T.J., Leggett, J.J.: Collection understanding for OAI-PMH compliant repositories. In: *Proceedings of JCDL 2005*, June 7-11, pp. 258–259 (2005)

# Person Specific Document Retrieval Using Face Biometrics

Vikram T.N, Shalini R. Urs, and K. Chidananda Gowda

International School of Information Management, University of Mysore, Manasagangotri,  
Mysore – 570006, Karnataka, India  
shalini@isim.ac.in, tnvikram@gmail.com, kcgowda12@yahoo.com

**Abstract.** In this work we propose a person specific document image retrieval system based on face images. Face images are extracted from all the documents and the document labels are tagged to the face images. We created synthetic documents, borrowing face images from ORL [7] face database. Experimental results based on Principal component analysis have revealed an improvement in retrieval time without any compromise in accuracy. This work address an important requirement in case of business and legal document image management system.

**Keywords:** Document retrieval, Face recognition, Face Indexing.

## 1 Introduction

Document image retrieval has been one of the most important functionality of the digital libraries. The digitization of land records, passports applications, driver's licenses, university documents, scientific literature etc. has propelled more research to be carried out in this area. Managing this huge information and indexing them has become one of the most important tasks. Document image retrieval aims at identifying relevant documents relying on image features only. In so far documents have been archived and retrieved based on the paradigms like Word indexing, Logo identification, Handwriting , Signature and physical layouts [1].

Currently documents are indexed using metadata like unique identification numbers, date of issue, registration number of the issuing authority etc. Though these are efficient, additional features like signature, face photograph and thumb impression are embedded to confirm the identity of the person whom the documents is pertaining to. Biometric features confirm the identity of the person across domains. There have been many instances where there is a wrong mapping our owners to documents due to near similar metadata records. Many instances exists where there is ambiguity or wrong mapping between of documents and its owners or authors. This can be particularly seen in author based indexing of published scientific literature. There are many such instances in the DBLP, Web of Science indexes, where inappropriate articles are associated with an author, since the tagging is based only on attributes like name, field of work etc. To resolve this issue of ambiguity and also to give wider publicity to the authors, journal like *Nerocomputing*, *IEEE transactions on pattern analysis and machine intelligence*, *IEEE transactions on knowledge and data engineering* have made

it mandatory to publish author profile with their photograph at the end of the article. The other advantage of having biometric features embedded is its ability to establish the ownership of the document, in case the document gets degraded or when printed metadata records gets distorted.

## 2 Methodology and Experimentation

We recommend the usage of face images to index documents. Face biometric is far more stable as compared to signatures[2], in presence of noise. Even though it has behavioral changes like variations in moods and expressions, the contemporary face recognition algorithms are matured enough to handle this. This can be seen in the works of Guru and Vikram [3], Nagabhushan et al [7], Yang et al [6]. The proposed methodology of person specific document retrieval based on faces is explained in the following section.

We recommend to extract the face images from the documents, and tag the face images with their respective documents. As this is a person specific retrieval, it is required to retrieve all the documents relevant to a person for a given query. In order to achieve this, we recommend to compute the average of the face images, by normalizing the size and tagging the all the documents to the average face image and storing in the database. An example of the same is given in Illustration 2. Principal component analysis and linear discriminant analysis based subspace models for face retrieval scheme are employed.

Principal component analysis (PCA) and linear discriminant analysis (LDA) are the most popularly used subspaces methodologies for subspace based face recognition. Coupled by simplicity and practical efficiency, PCA and LDA based approaches have carved a niche in the domain of face recognition. A plethora of PCA and LDA based schemes like 2D-PCA [4], 2D-FLD [5], alternate 2D-PCA [6], alternate 2D-FLD [7], have high recognition rates when compared to the original PCA and LDA schemes. Despite the fact that the 2D variants of PCA and LDA schemes have better recognition performance, they require more coefficients for image representation than the conventional PCA or LDA

We created a set to test documents, with various layouts and degradations and styles, and used to the ORL face database to add the face biometric feature to each of them. The motivation to not consider real documents was, that it is extremely difficult to find different documents pertaining to the same person at a single place. Also the issue of legalities in using them for such kind of activities. ORL face database consists of face images pertaining to 40 distinct users, with ten samples each. We generated 200 different documents with five documents pertaining to any given user and manually affixed with the ORL face images. The remaining 200 face images were used as query samples. Artificial degradation were introduced and the documents were scanned at 300 dpi.

The extraction of the face images from the document set was automated by employing the face detection methodology of Wu et al [9]. Experiments were conducted using Matlab version 7, on Pentium III, 256 MB Ram, Windows XP operating system.

The first level index is created by projecting the class average of every individual in the training set on to the optimal projection axes. A linked list is maintained along with the class average, with all the document identification numbers relevant to that. The query is also projected and then the Euclidean distance based nearest neighbour classifier is applied to find the  $k$  nearest classes in the first level index. Recognition of the query is carried out with respect to those images whose class labels are present in the index. In the conventional recognition schemes all the 200 projected training samples have to be scanned to identify the class label of a given query image. However as in the case of the proposed indexing scheme the part of the database to be scanned is reduced to  $k \times 5$  where  $k \leq 40$  and 5 is the number of training samples per class. Table.1 describes the optimal recognition performance of the various subspace based recognition models on the conventional and the proposed indexed training set.

It is inferred from Table.1 that the effective scanning of the database is reduced as compared to that of the conventional one. It shall also be noticed that the recognition time has considerably reduced throughout, without any loss in optimal recognition rates.

**Table. 1.** Subspace based recognition schemes on ORL images

Method		Classes Selected	Dimension of feature vector	Recognition Rate (%)	Effective % of database scanned	Recognition time	
						Secs	Reduction %
2D-PCA[4]	Conventional	40	2x112	95.5	100	3.03	53.8
	Proposed	12	2x112	95.5	50.0	1.40	
Alt.2D-PCA[6]	Conventional	40	9x92	94.5	100	5.02	52.4
	Proposed	15	9x92	94.5	57.5	2.39	
2D2-PCA[6]	Conventional	40	4x4	92.5	100	1.74	47.36
	Proposed	13	4x4	92.5	52.5	1.09	
2D-FLD[5]	Conventional	40	3x112	94.5	100	2.48	67.64
	Proposed	13	3x112	94.5	52.5	0.81	
Alt.2D-FLD[7]	Conventional	40	10x92	96	100	5.64	80.14
	Proposed	7	10x92	96	37.5	1.12	
2D2-FLD[7]	Conventional	40	7x7	94	100	2.00	74.5
	Proposed	10	7x7	94	45.0	0.51	

## 5 Discussion and Conclusion

We have presented here, a person specific document image retrieval system based on face images in the documents. Face biometrics are far more robust when it comes to manual human verification of the system generated recognition matches. We have recommended to use the average face image of a person and tagging all the document images to it and just retrieve the top ‘ $k$ ’ hypothesis for further recognition purposes. This reduces the scan time and improves the retrieval time without compromising the recognition rate. We have fixed ‘ $k$ ’ empirically in this work. This can also be fixed analytically by the methodology presented in Ghosh [8]. Person specific document image retrieval can be improved further by taking into account the fused scores of the

multiple biometrics (face, signature and fingerprint) that are available in the documents. Developing such a multimodal biometrics based person specific document image retrieval would be end objective of this research.

## References

- [1] Marinai, S., Marino, E., Soda, G.: Exploring digital libraries with document image retrieval. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 368–379. Springer, Heidelberg (2007)
- [2] Srihari, S.N., Shetty, S., Chen, S., Srinivasan, H., Huang, C., Agam, G., Frieder, O.: Document image retrieval using signatures as queries (2006)
- [3] Guru, D.S., Vikram, T.N.: 2D Paiwise FLD.: A robust methodology for face recognition. In: Proceedings of IEEE AutoID, pp. 99–102 (2007)
- [4] Yang, J., Zhang, D., Frangi, A.F., Yang, J.: Two dimensional PCA: a new approach to appearance based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(1), 131–137 (2004)
- [5] Li, M., Yuan, V.: 2DLDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters* 26(5), 527–532 (2005)
- [6] Yang, J., Zhang, D., Yang, X., Yang, J.: Two dimensional discriminant transform for face recognition. *Pattern Recognition* 38(7), 1125–1129 (2005)
- [7] Nagabhushan, P., Guru, D.S., Shekar, B.H.: 2D<sup>2</sup>-FLD: An efficient approach for appearance based object recognition. *Journal of Neurocomputing* 69, 934–940 (2006)
- [8] Ghosh, A.K.: On optimum choice of k in nearest neighbor classification. *Computational statistics and data analysis* 50(11), 3113–3123 (2005)
- [9] Wu, H., Chen, Q., Yachida, M.: Face detection from color images using fuzzy pattern matching methods. *IEEE transactions on Pattern Analysis and Machine Intelligence* 21(6), 557–563 (1999)

# The Potential of Collaborative Document Evaluation for Science

Jöran Beel and Béla Gipp

Otto-von-Guericke University, Department of Computer Science,  
Universitätsplatz 2, 39106 Magdeburg, Germany  
joeran@beel.org, bela@gipp.com

**Abstract.** Peer review and citation analysis are the two most common approaches for quality evaluations of scientific publications, although they are subject to criticism for various reasons. This paper outlines the problems of citation analysis and peer review and introduces Collaborative Document Evaluation as a supplement or possibly even a substitute. Collaborative Document Evaluation aims to enable the readers of publications to act as peer reviewers and share their evaluations in the form of ratings, annotations, links and classifications via the internet. In addition, Collaborative Document Evaluation might well enhance the search for publications. In this paper the implications of Collaborative Document Evaluation for the scientific community are discussed and questions are asked as to how to create incentives for scientists to participate.

**Keywords:** open peer review, citation analysis, alternative, research policy.

## 1 Introduction

Searching and evaluating scientific publications is a time-consuming activity. Synonyms, a growing number of publications and ambiguous nomenclature impede the search for relevant documents. Sometimes nomenclature itself changes over time. Therefore, researchers using keyword-based search engines miss out documents, if they do not know all relevant keywords or authors did not use the commonly known terms in their publications.

Once a publication is found, the reader needs to assess its quality and credibility. Usually scientists assume a publication's quality from the reputation of the issuing journal. Journals in turn select publications based on their peer reviewers' recommendations. However, peer review is often criticised for leading to non-objective decisions caused by incompetent reviewers and reviewers following own interests, whether due to competition, alliances or economical reasons [1,2]. The increasing amount of interdisciplinary articles encumbers the peer review process, too. Imagine an empirical study about the influence of music on online shop visitors' behaviour. A thorough evaluation would require experts in the fields of music, psychology, neuroscience, business, computer science and statistics. Hardly any journal has access to reviewers that could competently review such an interdisciplinary paper. This is especially the case for conferences.



Due to the limitations of the peer review process, scientists attempt to evaluate the quality of a publication by its citation counts. The assumption is that the more often publications and authors are cited, the better they are. However, citation analysis is subject to criticism as well. Citation databases are incomplete and sometimes erroneous; citation counts are spoiled by ceremonial citations, self and negative citations, cronyism, citation oblivion and the fact that authors tend to cite secondary sources rather than the original authors; authors are biased and do not cite all influences, while they sometimes cite publications they have never read [3,4]. Most importantly, citation counts can only measure ‘impact’, but impact does not necessarily correlate with quality [5].

Another drawback of citation analysis and peer review is their lack of capability to accomplish post-publishing quality evaluations. Once a paper is published, it is associated with the journal’s reputation even if at a later point in time new insights lead to a different assessment. For instance, John Darsee published dozens of articles in reputable journals. Later, most of his articles were proven to be fraudulent or at least questionable. Nevertheless, his flawed work was cited 298 times during the following ten years. An astonishing 86% of the citations approved of his work [6]. Apparently, the citing authors were unaware of the flaws and relied on the reputation of the publishing journals.

These shortcomings demonstrate the need for improving the existing quality evaluation approaches for scientific publications. The Otto-von-Guericke University is researching ‘Collaborative Document Evaluation’ as part of the *Scienstein.org* project. Collaborative Document Evaluation aims to let the scientific community evaluate publications and share the gathered information for everyone’s benefit.

## 2 Related Work

Some attempts have been made, by letting the scientific community rate and/or annotate papers – including others such as Arxiv.org, Bibsonomy.org, Naboj and Nature [7]. However, none of these attempts has been totally successful. Incentives for participation are not sufficient and the competence and trustworthiness of those participating are unclear. As a consequence, very few annotations or ratings exist and their reliability remains unclear.

Nevertheless, in other domains comparable projects have been successful. For instance, Wikipedia succeeded to let the ‘crowd’ create and evaluate content in a decent manner [8] and the United States Patent and Trademark Office introduced successfully the public reviewing of patent applications [9].

We believe that Collaborative Document Evaluation could be equally successful for evaluating publications.

## 3 Scienstein and Collaborative Document Evaluation

Collaborative Document Evaluation is about creating and sharing metadata of scientific papers by the scientific community via the internet. The metadata gathered in the Scienstein project includes ratings, annotations, links, classifications and highlighted

passages within documents. *Collaborative ratings* are quantitative ratings given in different categories such as *originality*, *significance*, *readability*, *correctness of methods and analysis* and *overall quality*. *Collaborative annotations* are comments for entire documents or parts of it. They can be classified, for instance as *critique*, *addition* or *misc* and may include *collaborative links*. These links can point generally to other documents or to specific passages, just as hyperlinks do. In contrast to hyperlinks, collaborative links can also be classified, similar to collaborative annotations. *Collaborative classifications* are similar to tags, but more structured [10].

Collaborative Document Evaluation enhances document search by various techniques. Via annotations and classifications, new terms can be assigned to publications. This way, older publications can be updated with modified or currently-used terminology.

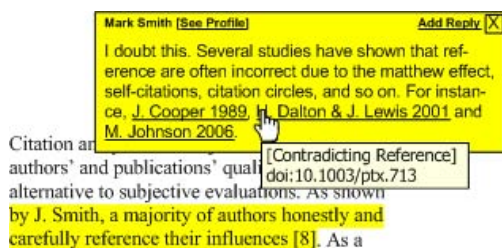


Fig. 1. Annotations, Highlighted passage and Links

Highlighted passages and annotations indicate which parts of a publication are considered particularly relevant by the scientific community. Accordingly, keyword-based search engines could attach greater significance to the words in the highlighted or annotated passages. Ratings given by researchers with similar interests can be used by a research paper recommender system [11].

Collaborative links not only show *that* publications are related to each other, but also *how* they are related, which enables a new type of search for related work.

Collaborative Document Evaluation may enhance quality evaluations in a number of ways. First, in contrast to citation counts, ratings measure the real (subjective) quality perception of the community members. Based on these ratings, the overall rating and the ratings of subgroups with similar interests can be displayed for each publication. Second, due to low entry barriers, new groups of people can act as reviewers. For instance, professionals could communicate their knowledge effortlessly via annotations. Since many readers from various research fields can participate, evaluating interdisciplinary work would be facilitated. Third, highlighted passages and annotations can provide more concise information about a publication in addition to the abstract. Finally, one community member would be sufficient to inform other scientists via annotations about new insights. Herewith, Collaborative Document Evaluation would enable the first continuous post-publishing quality evaluation.

Scienstein aims to motivate researchers to participate in Collaborative Document Evaluation with various incentives. It will be an open platform, available to everyone from every application. This way, metadata can be submitted and retrieved, for instance, from PDF-readers, reference managers or browsers. This is certainly more convenient than the current procedure. Moreover, researchers can directly benefit from participating. New software will help them to manage their electronic documents based on their evaluations. For instance, Scienstein could display all documents a researcher has recently read, classified as *peer review* and rated as *good*. Another positive aspect is that researchers could improve their visibility and the visibility of

their publications within the scientific community by annotating, rating, and classifying publications of colleagues. Last but not least, ratings of publications can be used by research paper recommender systems to generate recommendations. Accordingly, the more publications a researcher has rated, the better the recommendations. Also fundamental for the success of Collaborative Document Evaluation is the ability to determine the participants' competence and trustworthiness. These key success factors are covered in more detail in [10].

## 5 Summary

In this paper we presented Collaborative Document Evaluation as a supplement or even alternative to citation analysis and classic peer review. We outlined many advantages, such as an improved evaluation of (interdisciplinary) work, a continuous post-publishing quality evaluation of publications and improved search possibilities. As part of the Scienstein project, we are currently implementing the presented concept and are looking forward to the results. Particularly the researchers' motivation to participate and methods to determine their trustworthiness and competence will finally decide to what extent Collaborative Document Evaluation actually will be an effective and efficient method for evaluating scientific publications.

## References

1. Godlee, F., Gale, C., Martyn, C.: Effect on the Quality of Peer Review of Blinding Reviewers and Asking Them to Sign Their Reports. In: JAMA, pp. 237–240 (1998)
2. Relman, A.S.: Peer Review in Scientific Journals - What Good Is It? *New England Journal of Medicine* 153, 520–522 (1990)
3. Lee, D., Jaewoo, K., Prasenjit, M., Giles, L., Byung-Won, O.: Are your citations clean? *Communications of the ACM* 50, 33–38 (2007)
4. MacRoberts, M.H., MacRoberts, B.: Problems of Citation Analysis. *Scientometrics* 36, 435–444 (1996)
5. Yates, L.: Is Impact a Measure of Quality? *European Educational Research Journal* 4, 391–403 (2005)
6. Kochan, C.A., Budd, J.M.: The persistence of fraud in the literature: the Darsee case. *JASIS* 43, 488–493 (1992)
7. Nature's peer review trial, *Nature* (2006),  
<http://www.nature.com/nature/peerreview/debate/nature05535.html>
8. Ball, P.: The more, the wikier, *Nature* (2007) ,  
<http://www.nature.com/news/2007/070226/full/news070226-6.html>
9. Nyblod, R., Byrne, J.: USPTO Extends and Expands Peer Review Pilot (July 2008),  
<http://www.uspto.gov/web/offices/com/speeches/08-26.htm>
10. Beel, J., Gipp, B.: Collaborative Document Evaluation: An Alternative Approach to Classic Peer Review. In: proceedings of World Academy of Science, Engineering and Technology, vol. 31, pp. 410–413 (2008) ISSN 1307-6884
11. Gipp, B., Beel, J.: Scienstein: A Research Paper Recommender System (not published yet)

# Automated Processing of Digitized Historical Newspapers: Identification of Segments and Genres

Robert B. Allen, Ilya Waldstein, and Weizhong Zhu

College of Information Science and Technology, Drexel University  
rba@drexel.edu, imw22@drexel.edu, wz32@drexel.edu

**Abstract.** Many historical newspapers are being digitized. We aim to support access to them via text analysis of the OCRd content. However, the OCR includes many errors; so extracting meaningful content from it is difficult. A pipeline of processing steps is proposed. Here, we describe the first two steps: segmentation and genre identification. The segmentation procedure based on headings was quite successful. Genre identification worked well for easily defined genre categories such as weather reports. We also propose additional techniques which may improve the accuracy still farther.

**Keywords:** Categorization, Genres, Historical Newspapers, OCR, Segmentation.

## 1 Introduction

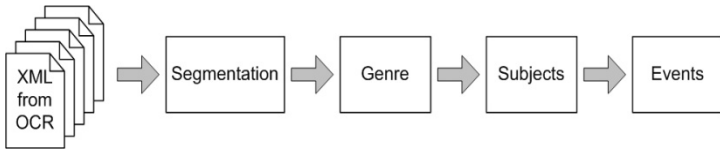
### 1.1 Digitized Collections of Historical Newspapers

Historical newspapers are an important source of past local news. They are being digitized to preserve content and enable search and retrieval. In the U.S., the National Digital Newspaper Project (NDNP) has several hundred thousand pages of historical newspapers digitized from microfilm, processed with OCR, placed online, and made accessible through search-engine interfaces. These materials should be more browseable if individual components are identified. Because of the large volume of material this would be costly to do manually, however, automated processing could be used to augment or replace manual classification. While there has been other work on processing modern news articles, there has been surprisingly little text processing research on historical newspapers or even on newspaper in their entirety. Because of the difficulty with OCR, our goal is not perfect recognition but enough improvement in accuracy that the process will eventually converge. We do not aim to develop new algorithms but to find robust solutions to a practical problem.

### 1.2 Automated Processing with the Pipeline

There are four main steps in the classification process (Figure 1). The modules take into account the properties of the data set, such as the date and page number, and the results of the previous processing step. We have developed an ad hoc XML wrapper for each of the newspaper segments which identifies the date, page, and coordinates of each

segment. The wrapper accumulates information as it passes down the pipeline. For example, once the genre is assigned it is stored in the XML wrapper. Eventually, this ad hoc XML scheme could be formalized as an article-level METS description.



**Fig. 1.** Pipeline for processing the OCR'd text

The pipeline provides a convenient model for conceptualizing these complex processes. However, it is a simplification in regard to the distinction between genres and subjects. The vagueness between genres and subjects makes a clear separation between them difficult. For example, a description of a baseball game can fall into the genre of sports notices or it can be counted as an article about baseball.

### 1.3 Test Corpus

NDNP files for the years 1900-1910 from Washington DC were obtained. From those, we focused on the *Washington Times* for 1904. Because they conformed to the NDNP specification, they included OCR which was wrapped into METS Alto files [7]. Thus, in addition to the OCR text itself, other attributes were included such as the coordinates for each word and the fonts. It is worth noting that the Alto files differ across digitization projects in the level of detail they include; for example, a few digitization projects have developed Alto files that include keystroked article headings and image captions. As with many newspapers, there is substantial variation from year to year. The OCR for 1904 was selected as a moderately difficult data set due to its quality. Figure 2 shows text samples from two OCR records from our test collection. To reduce the introduction of errors, only minimal processing was applied to the original OCR files. Thus, some corrections are relatively easy to make while others are more difficult. Indeed many other researchers have posited that OCR of this quality is too complex to process automatically. We argue that there are sufficient constraints that allow this text to be processed automatically. In particular, we believe that the tasks which involve categorization such as genre and subject identification will succeed even though some of the words in the OCR are not intelligible.

STATEHOOD MEASURE WILL PASS THE HOUSE Republicans Determine to Rush Hamilton Bill Through to Be Ready for Senate in December margin but NSW Mexico which has nearly I nearly double the population of Arizona is largely Republican at present The Republicans in their rule will provide that no amendment shall be considered

I THE T MEs71 I world Fair Contests it OFFEH NO ITf acid the three employes of the District or National + t tional Government collecting respectively the < Uteat number ofLouis Sti 4 Louis Exposition coupons to the Worlds Fair for 4 one week and payalixpenses i pxpenses Note District or National Government ewtptoUli es SUKonly Uli only the coupon

**Fig. 2.** Samples of relatively good (above) and poor (below) OCR from the 1904 *Washington Times*

## 2 Segmentation

The goal for the first step in the classification process is to identify and categorize meaningful segments of the newspapers. Across NDNP projects and collections, the METS Alto files differ in the detail with which they identify regions, thus we explored extracting segments from the OCRd text; however, none capture smaller segments such as classified advertisements. Segmentation is challenging because of content which varies widely in size and shape, sometimes even wrapping around pictures. There has been some previous work on segmentation (also called zoning) that identified many spatial factors and other segmentation issues [5, 6]. The OCR text can contribute provide semantics.

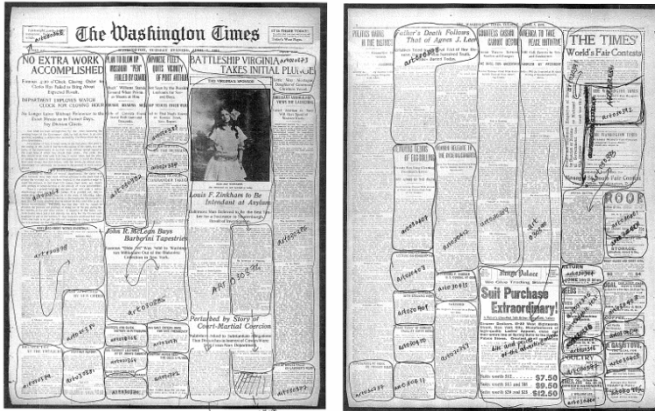
### 2.1 Method

Because we have OCRd text with word coordinates, we decided to use methods which incorporated that information. Three techniques of article segmentation were explored. Technique 1 relied on the identification of news article titles determined by several lines of capital letters. In Technique 2, a latent semantic indexing (LSI)-based linear segmentation technique [4] was used to divide the OCR text into blocks. The edges of the blocks were further identified by the closest lines of capital letters which might indicate titles. With Technique 2, the 5594 pages of the *Washington Times* in 1904 yielded 78297 regions. This approach seems to work well for simple segments but not for complex segments such as those split across several columns. Moreover, due to the poor quality of the OCR text, capital letters in the titles of many news articles were originally incorrectly identified, which caused many errors in determining block boundaries. In Technique 3 a combination of approaches was applied. The ratio of capital letters in each text line and the average font size of the titles (which should be larger than the body text) were used to developed a more accurate title detector. If the news articles are located in either the same block or in different blocks of one column, the text blocks are separated into regions by these titles. If the news articles fall across blocks in several columns, these text blocks are linked with the nearest titles that have smaller vertical coordinates compared to the vertical coordinates of the first text lines of the blocks.

### 2.2 Results

In Technique 3, the pages of the *Washington Times* for 1904 are divided into 116964 regions, an average of about 320 per day (though the Sunday was much higher). The number of segments identified is larger than with the other techniques apparently because it was better at distinguishing short items such as notices and advertisements.

We focused on the performance for one day, picked at random, April 5, 1904. Figure 3 shows images from pages 1 and 3 for that day with segments indicated. As can be seen from inspection of the images, the segmentations are generally accurate. Table 1 divides the results into different types of errors for the first four pages. 67% of the segments were either totally correct or involved only minor errors. Some of the errors affected major articles. For example, the segmentation of the very first article (at the top left of Figure 3) included the newspaper banner but then also incorrectly segmented the lower portion of that article. In the middle of that page, a segment was



**Fig. 3.** Pages 1 and 3 of the *Washington Times* for April 5, 1904 with extracted regions

missed. The caption of the picture on that page was also incorrectly joined to the article below it. For smaller items such as classifieds and notices, the most common error was merging text blocks, but these are often difficult for a human to parse without careful inspection. While the program performed well there were some highly complex headings where it continued to have difficulties. Similarly, it seemed to have particular difficulty with the captions of images and cartoons. Furthermore, spot checking its performance suggests that the performance we report for April 5 is typical.

**Table 1.** Summary of the results for the newspaper segmentation for the first four pages for April 5, 1904

	<i>Count</i>	<i>Percent</i>
Correct	64	67
nor errors (e.g., merging a few words)	6	6
Combined two or more articles	11	12
Too much segmentation	14	15

### 3 Genre Identification

As suggested by the second step in the classification process of the pipeline model, we need to identify the various types of content. This analysis aims to separate the news stories from the wide variety of other material in the newspaper. Some of those contents will be of interest in themselves. Indeed, several NDNP participants have paid to have the obituaries extracted manually because they are of considerable interest for genealogists. Potentially a technique such as ours could automate that process. In other cases, the articles can be analyzed with topic categories and event categorization. The International Press and Telecommunications Council (IPTC) is a newspaper trade organization which has proposed controlled vocabularies for news genres and subjects. We favored these because the terms were designed for news publishing and are used for many modern newspapers (cf., [3]). Even within a specific newspaper, there are often substantial changes in layout and types of coverage across time. While it would

be possible to tune the parameters of the program to this particular newspaper to improve the accuracy of the results, our goal is create a robust program.

### 3.1 Method

We used the IPTC genres as the basis for our categories, but applied the idea of genres somewhat loosely. Our primary goal is to identify different types of content and extract them. Moreover, the IPTC genres include some which are likely to be quite difficult to identify by automatic methods such as separating “background” from the story itself, and omitted other genres such as advertisements. Thus, we added other categories matching what we found in the *Washington Times*, including a set of advertisement sub-genres and game categories such as chess. Table 2 shows an example of genres and sub-genres.

**Table 2.** The genre categories used in this analysis

ads:autos	ads:travel	congressional notes	lost_found	sports:golf
ads:clothes	ads:whiskey	financial:securities	masthead	sports:horse racing
ads:hats	advertised letters	foreign mails	notices:auctions	sports:tennis
ads:insurance	banner	fraternal organizations	notices:church	sports:track
ads:medicines	bids solicited	games:bridge	notices:death notice	transporation-shipping
ads:paint	classifieds:helpwanted	games:chess	notices:music	vital records
ads:palmistry	commodities:cotton	legal notices: general	notices:navy	weather reports
ads:shoes	commodities:grain	legal notices:probate	obituary	

The primary technique for identification was based on matching words associated with the genre. A sample of terms used for some of the genre categories is shown in Table 3. In some cases, we also used phrases such as “for sale”. In addition, we used newspaper and region-specific terms. For example, for DC we included geographic terms such as “Alexandria”, “Bethesda”, and “Potomac”, and the names of the local baseball team, “The Senators”.

**Table 3.** Terms for some specific genre categories

Ads:	drugs, cure, cures, liver, kidneys, prescriptions, drug, pains, blood, nervous, eye,
Medicines	pain, dying, bone, extract, potency, ache, brain, skin, rectum, chronic, tonic, stomach, remedies, constipation, bottle, bladder, medicine, pills
Chess	chess, check, checkmate, mate, pawn, rook, game, match, win, problems, capture
Weather report	weather, report, temperature, degrees, rain, sun, snow, warmer, colder, temperatures, cloudy, icy, rainy

A score was calculated for each genre for each segment.<sup>1</sup> This score was compared to a genre-specific minimum threshold; typically, that threshold was in the range of 0.002 to 0.010. Other clues were also found to be useful; for example, some of the items had a lot of numbers (e.g., stock market tables, train schedules) in them and others had lists of names (e.g., advertised letters). Thus, we developed counters for

---

<sup>1</sup> 
$$\frac{\left( \frac{\# \text{ occurrences of all terms}}{\# \text{ different terms}} \right)}{\text{length of the segment}}$$



those and used those counters to augment the scores. The segment was assigned to the genre which had the highest score.

### 3.2 Evaluation

For this evaluation we focused on the segments from April 1904. Overall, the program identified and tagged the genre of about half of the total number of segments. These could be separated from the untagged segments. Many of those untagged segments were news stories which should be passed to the following stage of the pipeline for subject categorization. Clearly, not all of these segments were articles. Some of them were very short segments which were essentially incomprehensible. In addition, we continued to find categories which did not match well the categories in Table 3, such as advertisements for beers or advertisements for hens. It is difficult to decide where the subdivisions should stop. For example, we found a gradation of segments which fell between prose articles about baseball and segments which were composed primarily of statistics. Detailed results for some of the narrower genres are shown in Table 4. Performance is measured in terms of precision and recall and is based on the ratings of a human judge who is considered to be an expert. Precision is the number of articles correctly judged by the program to be in the genre divided by the total number judged in that category (including the ones incorrectly identified). Recall is the number of articles correctly identified by the program as matching the genre divided by the total number of articles judged by the expert to be in that genre. Precision and recall scores are common range from 0.0 to 1.0. Out of a total of 1028 articles that contained words that are in the weather genre, 30 were actual weather report files. The program identified 21 of those correctly, missed nine, and misidentified two. Note that weather reports were sometimes confused with genres such as the crop report which also mentioned the effects of weather on the crops but did not contain a specific weather report. Specific techniques could have been adopted to improve performance still further. For example, the word “weather” generally appeared near the beginning of the weather reports and we could have built a detector specifically for that. However, we were not sure that would be generally applicable for other newspapers so we did not implement it here.

**Table 4.** Accuracy for selected categories. Narrow genre categories such are recognized quite well.

	<i>Precision</i>	<i>Recall</i>
Advertised Letters	0.95	0.74
Games: Chess	0.75	0.59
Weather reports	0.91	0.70

## 4 Conclusions

### 4.1 Extensions of the Segmentation Analysis

There are three areas where we believe the segmentation procedure could be productively extended. First, the segmentation algorithm seems to have been confused by pictures and drawings because they leave a hole where there is no text. Automated analysis of the layout of the page-image file should be conducted to identify the locations of pictures and drawings so they can be accounted for in the segmentation process. Second,

some of the segments which were difficult to identify in the genre program were a combination of several unrelated notices. Presumably, this was because they did not include distinct headings but, perhaps, they could be disentangled with semantic analysis. We also need to explore the generality of the results across different years and different newspapers. Finally, as noted earlier Alto files from different digitization projects include different levels of segmentation information. Sometimes that includes keystroked article headings. Additional tests should be conducted to determine how useful that information would be in aiding accurate segmentation.

## 4.2 Extensions of Genre Identification and Adding Feedback to the Pipeline

Compared to the segmentations, quite a lot of improvement should be possible for genre identification. While we have already incorporated a number of factors, there are still other constraints which should be explored. Here, we examine two possibilities. First, in some cases such as weather reports we know there should be one and only one instance each day. Such a constraint could be strictly enforced. Second, we plan to add a named-entity identification utility. Names such as those of government officials would help pin down the nature of the segment which mentions them. In addition, as with the segmentations, we need to explore the generality of the genre identification results for different years and for different newspapers. More ambitiously, we have obtained the best results when using specific genres (e.g., ads:medicine) rather than general ones (e.g., literature or opinion). Although it may be fairly subtle, the text for those general categories should be distinctive and might be able to be processed automatically. We plan to apply and test several machine learning techniques (e.g., [9]) for categorization. The Pipeline Model has proven to be an effective basis for initial analyses, but for further analyses, it seems that feedback should be helpful. For example, the genre and subject categories could improve the original term disambiguation. Similarly, using known headings to look for segment headings such as “Special Notices” could be used to improve the segmentation.

## 4.3 The Big Picture

We will be continuing to the next stages of the pipeline; that is to subject categorization and event identification. Indeed, we are exploring advanced strategies for categorization [9]. In addition to genre codes, the IPTC also provides a controlled vocabulary for subjects (topics) and we will use them. As with genre analysis, there are many factors which contribute to accurate subject categorization. In addition to named entities, there are temporal patterns which could be exploited in an analysis program (Table 5). The year 1901 was selected because it covers the death of President McKinley and the beginning of Theodore Roosevelt’s presidency.

**Table 5.** Month-by-month word counts for selected terms in the *Washington Times* for 1901. For the term “drought” there is a clear seasonal pattern. For other terms such as “President” and “Roosevelt” there was a sharp increase in the count in September when he became president.

<i>Terms</i>	<i>J</i>	<i>F</i>	<i>M</i>	<i>A</i>	<i>M</i>	<i>J</i>	<i>J</i>	<i>A</i>	<i>S</i>	<i>O</i>	<i>N</i>	<i>D</i>
Drought	4	6	4	3	8	7	53	21	10	3	7	3
President	877	746	984	811	787	762	460	358	1798	856	932	1070
Roosevelt	37	50	70	14	13	8	7	38	198	201	200	222

More generally, we believe that providing explicit knowledge about history and about the newspaper will be helpful. However, it does not seem feasible to do that by entering specific facts; there are just too many. Generative models such as cyclic models for the seasons may be better (see [2]). These could be simple such as listing the months of the baseball season, the years in which there are presidential elections, the years during which the Wright Brothers were working, or the locations of major buildings in the city. Beyond categorization, we would like to consider other text-based services such as summarization and the identification of specific events. Doing that would make it much easier to develop timeline interfaces (cf., [1]) and to link to resources such as Wikipedia. Clean text would be very helpful for those services and hopefully, we will be able to use the categories to help correct the text. However, it seems unlikely that automated text correction can yield very high accuracy. To reach that level of accuracy is likely to require some sort of human intervention. Perhaps if the quality of the text is relatively high, the corrections could be made by professionals, but it might also be possible to recruit members of a local historical society to make the changes (cf., [8]).

## Acknowledgements

This work was supported in part by an NEH Digital Humanities Start-up Research Grant. We thank Abhijeet Ganachari for assistance and we thank Ray Murray of the Library of Congress for the *Washington Times* files.

## References

1. Allen, R.B.: A Focus-Context Timeline for Browsing Historical Newspapers. In: ACM/IEEE Joint Conference on Digital Libraries, pp. 260–261 (2005)
2. Allen, R.B., Japzon, A., Achananuparp, P., Lee, K.-J.: A Framework for Text Processing and Supporting Access to Collections of Digitized Historical Newspapers. In: HCI International Conf. (2007)
3. Allen, R.B., Schalow, J.: Metadata and Data Structures for the Historical Newspaper Digital Library Project. In: ACM CIKM, Kansas City, November, pp. 147–153 (1999)
4. Choi, Y.Y.: Advances in domain independent linear text segmentation. In: Proceedings of NAACL, Seattle, USA (2000)
5. Gatos, B., Gouraros, N., Mantzaris, S., Perantonis, S., Tsigris, A., Tzavelis, P., Vassilas, N.: A New Method for Segmenting Newspaper Articles. In: SIGIR, p. 389 (1998)
6. Kanungo, T., Allen, R.B.: Full-Text Access to Historical Newspapers. Technical Report: LAMP-TR-033/CAR-TR-915/CS-TR-4014, University of Maryland, College Park (April 1999)
7. Murray, R.: Towards a Metadata Standard for Digitized Historical Newspapers. JCDL, 330–331 (2005)
8. von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: ReCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321, 1465–1468 (2008)
9. Zhu, W., Allen, R.B.: Topic and Event Categorization of Historical Newspapers (in preparation)

# Arabic Manuscripts in a Digital Library Context

Sulieman Salem Alshuhri

Department of Computer and Information Sciences  
University of Strathclyde  
sulieman.alshuhri@cis.strath.ac.uk

**Abstract.** This paper aims to present related issues of indexing and retrieving Arabic manuscripts in digital libraries. Composed Arabic manuscripts have been taken as samples of presentation in this paper. Comments are commonly found in AMSs, with varying purposes and forms of comments. A digital library should deal with composed AMSs according to their characteristics and users' requirements.

**Keywords:** Digital Library, Arabic Manuscripts, Culture Heritage Resources.

## 1 Introduction

There is no doubt that a good deal of attention has been paid to digitizing cultural heritage resources [1, 2, 3, 4], particularly Arabic manuscripts, in different parts of the world. In this regard, there is a number of international projects have been carried out, for example the World Digital Library [5] led by the Library of Congress in cooperation with several international bodies, such as UNESCO, the Bibliotheca Alexandrina, King Abdullah University of Science and Technology, etc. Additionally, there have been some digitization projects at a regional/local level. Hence, there are certain websites that present AMSs, for instance the Islamic Medical Manuscripts in the National Library of Medicine [6], the Jafet Library in the American University of Beirut [7], the Collection of Arabic Manuscripts in the National Library of the Czech Republic [8], the Memory of the World Programme [9], Digitized Arabic materials at the Royal Library of Denmark [10], etc.

The majority of efforts based on making AMSs available via digital libraries focus on the recognition of handwriting as result of the inefficiency of the state-of-art OCR systems in recognizing the content of ancient manuscripts generally [11, 12, 13], and AMSs in particular [11, 14, 15, 16, 17, 18]. Thus, it has become the trend to build up sophisticated systems to retrieve manuscript images, for example, Content Based Image Retrieval (CBIR) [16, 17, 19]. However, this approach relies on human assistance in its processes, either for indexing or extracting metadata from images.

## 2 The Concept of AMSs

There are several definitions of the term 'manuscript' in the literature, according to culture, and/or discipline in the context of library science. *Harrod's Librarians' Glossary*

defines a manuscript as a "document of any kind which is written by hand, or the text of music or literary composition in hand-written or typescript form, and which in that form, has not been reproduced in multiple copies" [20]. This definition includes also pre-print documents (typescript) before submission for publication. The ALA Glossary of Library and Information Science defines a manuscript as having three thematic purposes as follows: "... The handwritten copy of an author's work before it is printed; or, loosely, the author's typescript.." [21].

The concept of a manuscript in Western countries is quite dissimilar to that in Arab countries in terms of the time of writing the manuscript. In the Arabic context, the concept of a manuscript, and the one adopted in this paper, is any document that was written by hand before the introduction of the printing press. Bearing this in mind, there was a disparity in the introduction of the printing press from one country to another [22]. The printing press in the Arabic peninsula, for example, was introduced in the early twentieth century [23]. Therefore, it can be considered that the period of manuscripts lasted around fourteen hundred years, dating from the first known manuscript book - the *Holy Quran* - in the seventh century AD [24]. This long time of period imparts to AMSs a particularity of structures, content values, and forms [25].

### 3 Commentary of AMSs

It can be observed that there are many patterns of AMSs that can be categorized by their appearance, such as illuminated, illustrated, and commentary manuscript. The term 'commentary AMS' refers to any manuscript containing comments surrounding the main text (*matn*), regardless of the purpose or form of the comments. Unsurprisingly, there were purposes behind and regulations concerning the writing of comments on AMSs. Hence, authors and copyists were required to leave enough space surrounding the *matn* for binding or adding comments.

It is possible to identify a set of common purposes of comments that were well-known and taken into account by authors and commentators. Mashukhi, A.S lists some of the purposes of comments: following the *matn* with critical points, explanation (exegesis), completion, correction, or decorating [26]. Additionally, comments could be for the purpose of linking ideas in different resources, or summarizing [27]. Comments can therefore be divided into two forms, as follows:

1. Writing a separate book: A commentator was often compelled to write a separate book, which included the original text and the comments in his/her authorship style, such as *Fath al-bari sharh Sahih al-Bukhari* by Ibn Hajr Al-'Asqalani. It is worth mentioning that a distinction should be made between the original text and comments by either using different colour, or using a specific word referring to the original text, such as writing before the original text - said قال - referring to the original text's author [27, 28].
2. Merged comments: This method was used by many editors and authors. This approach was based on writing comments surrounding the *matn*, or between lines. It was common to add defined marks or signs to illustrate the position of the comment in the *matn*. These marks or signs were also used to clarify the purposes of the comments[24], **Figure 1**.

Comments could be made by the author of the *matn*, the author's students, the author's followers, the author's opponents, the ordinary reader, or even a combination of these [27]. Therefore, it can be argued that there was more than one creator of the commentary AMSs' contents.



**Fig. 1.** Sample of using words (شرح = Explanation) or initial (ش) to clarify the comments purposes

Comments are one of the most reliable sources in editing and reading AMSs [29]. Thus, comments can provide the AMSs' beneficiaries with valuable information. Such benefits such as verification of the *matn*, finding out how the *matn* circulated among readers, and discovering the impact of the *matn* at different times periods, can be gleaned from the commentary MS [26]. Therefore, Comments should be taken into consideration during digitisation stage, indexing, and previewing to users.

#### 4 Composed AMSs

A composed manuscript is a subcategory of the commentary MS concept. However, this type of MS should involve more than one creator; for example, the main author of the *matn* and a different commentator. Thus, the composed MS excludes comments made by the *matn*'s author only. Comments in composed MS should be made by different body, **Figure 2.**



Fig. 2. Hahiyat 'ala sharh Muhammed Bin Mubarkshah al-Bukhari 'ala Hikmat al-'ayn li-'Ali Bin 'Umar

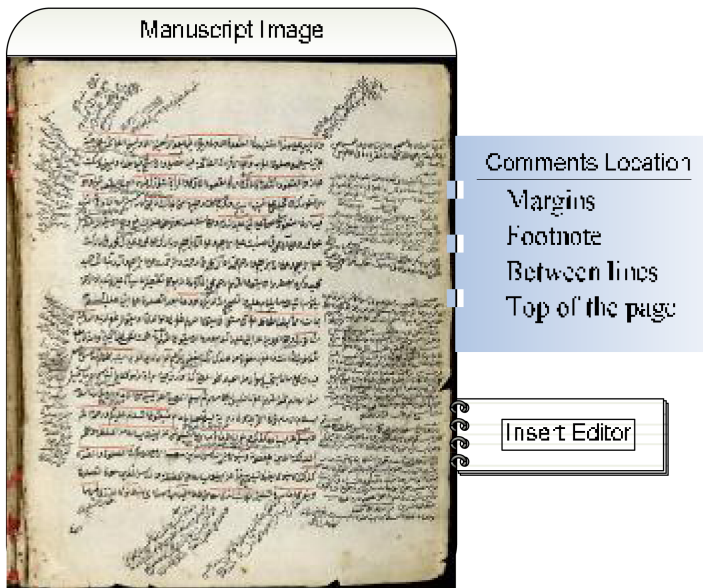


Fig. 3. Comments location

The location of comments in a composed MS can therefore be on a different part of the page. Examples of such locations surrounding the *matn* are in margins, footnotes, or at the top of the page, **Figure 3**. Sometimes comments may be located between lines, or on scraps of paper attached to the AMS [26, 27, 28, 30]. Comments can also be located at the beginning, or the end of the AMS [26].

It is worth mentioning that in some composed AMSs there is a variety of levels of comment; for example, comments that target the *matn's* comments, or, in other words, commented comments. Therefore, composed AMSs can be categorized into three levels: primary level, which refers to comments targeting the *matn*; secondary level, which is that of comments targeting the *matn's* comments; and the tertiary level, which is that of all other comments targeting the comments made at the secondary level.

## 5 The Digital Library in Context

It is certain that the digital library can play a significant role in delivering easily and effectively primary resources such as these AMSs [19, 31]. However, the characteristics and users' requirements of AMSs should be taken into consideration in order to provide an ideal service. In this regards, AMSs can be handled in an appropriate way in the digital library. In this context state-of-the-art digitization, organization, and services can be utilized, although the greatest emphasis in a Digital Library of Arabic Manuscripts (DLAMSs) is on the digitization stage. Content management and services in DLAMSs are still in their infancy in terms of both theory and practice.

The composed AMS in the digital library context might be rare case in terms of handling with a diversity of contents on one page. This could be a challenge, especially in the shadow of the ineffectiveness of the current OCR technologies. Compounding this is the fact that manual indexing of AMSs is generally a very expensive task, and could also give rise to human error [17]. However, many studies and bibliographic works have addressed AMSs. Consequently, a good number of these works have turned their attention to composed AMSs as palaeographic, codicologic, and bibliographic studies, such as *Tarikh al-adab al-'Arabi* by Carl Brockelman [32], *Kashf al-zunun 'an asami al-kutub wa-al-funun* by Hajji Khalifah [28], and *Tarikh al-turath al-'Arabi* by Faud Sezgin [33]. These works are primary tools and references for researchers and editors in their dealings with AMSs. Harun, A. M. [29] stresses that any researcher or editor of AMSs should use such tools as these, as well as catalogues of AMSs, dictionaries, biographies, etc. Therefore, digitising these bibliographic tools and make them in database structure linked with AMSs' metadata and images can facilitate the editing and reading process.

In addition, there are many tools that can be implemented to index and retrieve composed AMSs; for instance, adding appropriate elements to a metadata schema, such as Dublin Core (DC). Up to now, there has been no specific element to describe composed AMSs, apart from the use of the Note element to state if an AMS contains comments or not, while other elements of AMSs, illumination, and illustration, have been added to DC in some AMSs' metadata schema. So, it could be a necessary to use the expanded DC schema to describe AMSs, and adding some sub-elements, such as under Contributor element adding commentator, corrector, etc. In addition, if the purpose of comments known should be mentioned in Note element. Bearing in mind, AMSs cataloguing rules state the need of adding reference field for external resources which have been used to



obtain AMS information to the catalogue. Therefore, reference element that includes external resources data should be added to AMSDC schema.

Functional Requirements for Bibliographic Records (FRBR) and Resource Description and Access (RDA) can be utilized in terms of building a relationship system among collections of AMSs. It is worth mentioning that handling a composed AMS requires deep content analysis.

## 6 Conclusion

This paper reveals characteristics of rare cultural heritage documents that have been digitized in several digital libraries of Arabic manuscript projects. Composed Arabic manuscripts were produced from the accumulation of knowledge and culture in Islamic civilization. This type of document provides its potential users with very rich information. Thus, digital libraries of Arabic manuscripts should realize the characteristics and user requirements of composed AMSs. In addition, DL is supposed to have an advantage over traditional libraries in terms of handling AMSs. There are reference works and tools that can be implemented to facilitate access, content indexing, and retrieval. Moreover, there is a potential possibility of creating an ideal model to be implemented in digital library content management. In applying this model, AMS disciplines, authorship, geography and chronology need to be taken into account.

## References

1. O'Keefe, E.: Medieval manuscripts on the Internet. *Journal of Religious and Theological Information* 3, 9–47 (2000)
2. Boserup, I.: The manuscript and the Internet: digital repatriation of cultural heritage. *IFLA Journal* 31, 169–173 (2005)
3. Liew, C.L.: Online cultural heritage exhibitions: a survey of information retrieval features. *information systems* 39, 4–24 (2005)
4. Nicolas, S., Paquet, T., Heutte, L.: Digitizing cultural heritage manuscripts: the Bovary project. In: *Proceedings of the 2003 ACM symposium on Document engineering*, pp. 55–57 (2003)
5. World Digital Library,  
<http://www.worlddigitallibrary.org/project/english/index.html>
6. Islamic Medical Manuscripts at the National Library of Medicine,  
<http://www.nlm.nih.gov/hmd/arabic/arabichome.html>
7. Jafet Library at American University of Beirut,  
<http://ddc.aub.edu.lb/projects/jafet/manuscripts/>
8. The Collection of Arabic Manuscripts in the National Library of the Czech Republic,  
[http://digit.nkp.cz/knihcin/digit/KatalogCD/EN/COLLEC\\_1/Collecti.htm](http://digit.nkp.cz/knihcin/digit/KatalogCD/EN/COLLEC_1/Collecti.htm)
9. Memory of the World Programme,  
<http://www.unesco.org/webworld/mdm/visite/sommaire.html>
10. Digitized Arabic Materials at the Royal Library of Denmark,  
<http://www.kb.dk/en/nb/samling/os/naeroest/arabdigi.html>
11. Le Bourgeois, F., Trinh, E., Allier, B., Eglin, V., Emptoz, H., Liris, C., Villeurbanne, F.: Document images analysis solutions for digital libraries. In: *Document Image Analysis for Libraries, Proceedings. First International Workshop*, pp. 2–24 (2004)

12. Edwards III, J.A.: Easily adaptable handwriting recognition in historical manuscripts. University of California, Berkeley, United States, California (2007)
13. Leydier, Y., Lebourgeois, F., Emptoz, H.: Text search for medieval manuscript images. *Pattern Recognition* 40, 3552–3567 (2007)
14. Khorsheed, M.S.: Recognising handwritten Arabic manuscripts using a single hidden Markov model. *Pattern Recognition Letters* 24, 2235–2242 (2003)
15. Eglin, V., Lebourgeois, F., Bres, S., Emptoz, H., Leydier, Y., Moalla, I., Drira, F.: Computer assistance for Digital Libraries: Contributions to Middle-ages and Authors' Manuscripts exploitation and enrichment. In: *Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL 2006)*, pp. 265–280 (2006)
16. Al-Khatib, W.G., Shahab, S.A., Mahmoud, S.A.: Digital Library Framework for Arabic Manuscripts. In: Shahab, S.A. (ed.) *Computer Systems and Applications, AICCSA 2007. IEEE/ACS International Conference*, pp. 458–465 (2007)
17. Shahab, S.A.: Document analysis and indexing of Arabic manuscripts. King Fahd University of Petroleum and Minerals (Saudi Arabia), Saudi Arabia (2006)
18. Lorigo, L.M., Govindaraju, V.: Offline Arabic handwriting recognition: a survey. *Pattern Analysis and Machine Intelligence. IEEE Transactions* 28, 712–724 (2006)
19. Shahab, S.A., Al-Khatib, W.G., Mahmoud, S.A.: Computer Aided Indexing of Historical Manuscripts. In: Al-Khatib, W.G. (ed.) *International Conference on Computer Graphics, Imaging and Visualisation*, pp. 287–295 (2006)
20. Prytherch, R.J.: *Harrod's librarians' glossary and reference book: a directory of over 10, 200 terms, organizations, projects and acronyms in the areas of information management, library science, publishing and archive management*. Ashgate, Aldershot (2005)
21. Young, H., Belanger, T.: *The ALA glossary of library and information science*. American Library Association, London, Distributed by Eurospan, Chicago (1983)
22. Al-showaish, A.: Establishment of a manuscript bibliographical information sharing network among the major libraries in Riyadh, Saudi Arabia. The University of Arizona, United States – Arizona (2000)
23. Saàti, Y.M.: *Al-Tibaàh fi Shihb al-Jazirah al- Àrabiyah fi al-qarn al-tasi àshar al-Miladi (1297-1317 H)*. Dar Aja, al-Riyad (1998)
24. Halwaji, A.: *Al-Makhtut al- Àrabi*. Maktabat Musbah, Jiddah (1989)
25. Mahasini, S.: *Al-Wasail al-tawdhiyah fi al-makhtutat al- ìlmiyah al- Àrabiyah*. Maktabat al-Malik Fahd al-Wataniyah, al-Riyad (2001)
26. Mashukhi, A.S.: *Anmat al-tawthiq fi al-makhtut al- Àrabi fi al-qarn al-tasi al-Hijri*. Maktabat al-Malik Fahd al-Wataniyah, al-Riyad (1994)
27. Nabhan, A.K.: *Al- Àlaqat bayna al-nusus fi al-talif al- Àrabi : dirasah àl' tafaru àl-nusus al- Àrabiyah : manhaj jadid li- ìlm al-bibliyujrafiya al-takwiniyah*. al- Àrabi, al-Qahirah (1993)
28. Khalifah, H.: *Kashf al-zunun àn asami al-kutub wa-al-funun*. Dar al-Kutub al- ìlmiyah, Bayrut, Lubnan (1992)
29. Harun, A.M.: *Tahqiq al-nusus wa-nashruha awwal kitab Àrabi fi hadha al-fann yuwaddihu manahijahu wa-yu àliju mushkilatih*. Muassasat al-Halabi lil-Nashr wa-al-Tawzi, al-Qahirah (1965)
30. Commentary Manuscripts Conference,  
<http://www.manuscriptcenter.org/commentary/About.asp>
31. Crane, G.: Cultural Heritage Digital Libraries: Needs and Components. In: *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 626–637 (2002)
32. Brockelmann, C.: *Tarikh al-adab al- Àrabi*. Dar al-Ma àrif, Misr (1967)
33. Sezgin, F.: *Tarikh al-turath al- Àrabi*. Al-Mamlakah al- Àrabiyah al- Sa ùdiyah, Wizarat al-Ta ìim al- Àli, Jami àt al-Imam Muhammad ibn Sa ùd al-Islamiyah, [Riyadh] (1991)

# Discovering Early Europe in Australia: The *Europa Inventa* Resource Discovery Service

Toby Burrows

ARC Network for Early European Research, University of Western Australia,  
35 Stirling Highway, Crawley WA 6009, Australia  
tburrows@library.uwa.edu.au

## 1 Introduction

The ARC Network for Early European Research (NEER) is funded under the Australian Research Council's Research Networks programme. Its goal is to enhance the scale and focus of Australian research in this multidisciplinary field, and to build collaborative and innovative approaches to the way research is planned and managed. It has more than 350 individual participants, as well as eight industry partners, working across five major research theme areas and fourteen Research Clusters [1].

An integral part of NEER's vision is the development of a digital environment which provides a focus for the work of this national research community. This environment includes a collaborative Web work space (Confluence) [2] and a repository of research outputs (PioNEER) which will be launched in late 2008. This paper looks at the third major component of the digital environment: *Europa Inventa*, which will provide a resource discovery service for Early European objects, artworks and manuscripts held in Australian cultural heritage institutions.

## 2 Europa Inventa

Australia's libraries, galleries and museums hold thousands of rare and irreplaceable European items which pre-date the era of European settlement in the late 18<sup>th</sup> century. The range of objects is extensive: manuscripts, books, maps, artworks of all kinds, furniture, fabrics, ceramics, glassware, silverware, scientific instruments – a record of European culture and history stretching across hundreds of years. Many of these are unique items, which will never be digitised or catalogued as part of any collections in Europe itself. NEER researchers regard a resource discovery service for these items as the most important single digital service which NEER can provide.

Until now, systematic research into these objects has been hampered because it has been difficult to identify them effectively as a coherent group. For medieval manuscripts and paintings in Australia, the available printed catalogues are significantly out of date or limited in scope. On-line catalogues and databases vary greatly in their quality and coverage, and in the effectiveness of their search interfaces. Some of them are not available on the Web at all, especially in the gallery and museum sectors.

In the first stage of the *Europa Inventa* project, NEER has established two databases of Early European objects, which currently contain information about nearly 2,000 artworks and 400 medieval manuscripts, drawn from the major Australian libraries,

galleries and museums. The NEER databases provide descriptive information about the objects, with consistent and normalised metadata based on the Getty Trust's *Categories for the Description of Works of Arts (CDWA Lite)* [3] for the artworks, and the Text Encoding Initiative's *Guidelines for Manuscript Description* [4] for the medieval manuscripts. These databases will shortly become available on the Web, providing the first unified access to information about materials of this kind in Australia and increasing the international exposure of Early European items in Australian public collections.

### 3 Future Work

The second stage of *Europa Inventa* will add value to these resources through semantic ontology-based frameworks. Early European research is a difficult area for metadata because of the many European languages used in the original sources and in contemporary scholarship, and the lack of consistent terminology in some fields. Mapping variant forms of names is a particular challenge. An important model for this service will be the MuseumFinland project [5], which has demonstrated the value of an ontology-based, Semantic Web framework for providing unified access to cultural heritage resources from a variety of different museums.

Another key goal of the *Europa Inventa* project is to implement tools for researchers to contribute annotations and commentaries on the individual objects, as well as providing facilities for continuing revision and updating of the descriptive information. They will also be able to build links between individual objects and the bibliographies and publications relating to them. The initial software requirements for enabling annotations for individual objects are currently being identified.

The project is also aiming to implement methods of cross-linking with European databases which contain records for similar types of material. The semantic framework being developed by the project will be the crucial element in building such links. This work will be coordinated with European efforts, such as those of the CARMEN Medieval Manuscripts Research Group, to apply similar semantic frameworks to discovery services for materials held in European collections.

### References

1. <http://www.neer.arts.uwa.edu.au>
2. <http://confluence.arts.uwa.edu.au>
3. [http://www.getty.edu/research/conducting\\_research/standards/cdwa/cdwalite.html](http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html)
4. <http://www.tei-c.org/release/doc/tei-p5-doc/html/MS.html>
5. Hyvönen, E.: MuseumFinland – Finnish Museums on the Semantic Web. *Journal of Web Semantics* 3(2), 224–241 (2005)

# Mapping the Question Answering Domain

Mohan John Blooma, Alton Yeow Kuan Chua, and Dion Hoe-Lian Goh

Wee Kim Wee School of Communication & Information, Nanyang Technological University,  
Singapore 637718

{b10002hn, AltonChua, ASHLGoh}@ntu.edu.sg

**Abstract.** We present a trend analysis of the question answering (QA) domain. Bibliometric mapping was used to sketch the boundary of the domain by uncovering the topics central to and peripheral to QA research in the new millennium. This paper visualizes the evolution of concepts in the QA domain by studying the dynamics of the QA research during the periods 2000 – 2003 and 2004 – 2007. It was found that question classification, answer extraction, information retrieval, user interface, performance evaluation, web, & natural language were the main topics in current QA research.

**Keywords:** Bibliometric mapping, co-word analysis, question answering.

Question Answering (QA), a pertinent branch of information retrieval is an emerging research field with roots traced back to 1960s. Studies in QA research aim in building intelligent systems that can provide succinct answers to questions constructed in natural language. As the literature in QA research evolves dynamically and proliferates in diverging research directions, the task of charting the intellectual structure of the domain becomes increasingly challenging. This study aims to mark the perimeters of the research trends in the QA domain using widely accepted bibliometric mapping tools [3, 4].

Papers containing the term “question answering” in the title were used as the selection criteria to download from ACM digital library. The documents were restricted to only journals and conferences between 2000 and 2007. The title, abstracts and keywords were extracted. Connexor was used to extract noun phrases and TEXTSTAT2 was used for identifying most frequent phrases. Co-word analysis was performed on phrases in this study. Co-word analysis is built on the assumption that a paper's keywords constitute an adequate description of its content. It is based on the nature of words, which are the important carrier of scientific concepts, idea and knowledge. Two keywords co-occurring within the same paper indicates a possible link between the topics to which they refer. The presence of many co-occurrences around the same word or pair of words highlights a locus of considerable association within papers that may correspond to a research theme. Pajek, a freeware program for visualization developed by the University of Ljubljana, is used in this study for mapping. The QA domain in this study is visualized using the algorithm of Kamada & Kawai as it is available in Pajek. 30 most frequent words to map the domain.

Sketching the boundary of QA domain: 2000 – 2007: The network map illustrated that the core concepts related to QA are “information retrieval”, “question type”, “answer”, and “performance” forming the inner circle. “Information search”, “experimentation”,

“information storage”, “algorithm”, “design”, “natural language” and “web” are the phrases that form the outer circle. Hence it could be concluded that the phrases in the inner circle are highly correlated terms in the QA domain. Analysis of the period 2000-2003: It was evident that during the period 2000 – 2003, QA research were centred on “question classification”, “web”, “performance” studies, “information retrieval”, “answer” extraction techniques, and “search engines”. However, the concepts like “natural language”, “information systems”, “knowledge annotation”, “user”, & “interface” were plotted as distant disciplines to “question answering”. “Ontology”, “predictive annotation” and “machine learning” are other new areas that were mapped in this network analysis. Analysis of the period 2004-2007: It was found that phrases “question type”, “answer”, “information search” and “information retrieval” were highly co-related terms to QA. “Automatic”, “syntactic” “pattern” were terms that were unique to this period indicating that they are the emerging areas in the QA domain.

From the above results, it could be inferred that information retrieval, information storage and search are the subject areas that are highly co-occurring with the phrase question answering. This finding is in evidence to the universal definition of QA that it is a type of information retrieval. “Question type” is a very closely related concept in QA domain and hence Question classification studies are a major research trend in QA domain [1]. The third concept that has been focused in the QA research since 2000 is “performance”. The fourth concept is inevitably the answer extraction [2]. From a bird’s eye view of the QA domain obtained from this study it could be inferred that QA domain is a multi disciplinary domain.

By conducting a co-word analysis for two time periods 2000 – 2003 and 2004 – 2007, it could be concluded that computational linguistics and artificial intelligence are the emerging trends during the last five years. Enhancing the performance, user studies, and enriching the algorithms have gained higher similarity rather than studies on question classification and answer extraction during the recent years. Results of this study educe the multidisciplinary nature of QA. It also paves way to the future of QA and predicts that there will be a tremendous growth in research related to the user interaction and computer linguistics areas other than its paternal domain of information retrieval.

## References

1. Blooma, M.J., Chua, A., Goh, D.: A predictive framework for retrieving the best answer. In: The proceedings of the ACM Symposium on applied computing, pp. 1107–1111
2. Blooma, M.J., Chua, A., Goh, D.: Applying question classification to yahoo answer. In: The proceedings of the first IEEE International Conference on the Applications of Digital Information and Web Technologies (2008)
3. Callon, M., Courtial, J.P., Turner, W., Brain, S.: From translations to problematic networks. An introduction to coword analysis. *Social Science Information* 22(2), 191–235 (1983)
4. Janssens, F., Leta, J., Glanzel, W., Moor, B.D.: Towards mapping library and information science. *Information Processing and Management* 42, 1614–1642 (2006)

# A Scavenger Grid for Intranet Indexing

Ndapandula Nakashole and Hussein Suleman

Department of Computer Science, University of Cape Town  
Private Bag x3, Rondebosch, 7701, South Africa  
{nnakasho,hussein}@cs.uct.ac.za

**Abstract.** Digital library services, such as searching and browsing, are increasingly needed in more restricted environments than the public Web. This paper proposes a scavenger Grid of idle desktop workstations to support computationally-intensive indexing services. A prototype software system was developed using commodity Grid middleware and information retrieval tools. This system demonstrated that the overhead incurred by Grid middleware is manageable and that performance gains are significant.

## 1 Introduction

Companies and educational institutions produce large numbers of electronic documents on a daily basis. Some documents, however, embody the intellectual property of the organization and should not be accessible to the outside world. These documents need to be indexed for rapid retrieval within the organization, but cannot be made accessible to the outside world. Thus, Web search engines may not be appropriate and third party indexing service providers can be used only if there are strong trust relationships and information is guaranteed to be secure. Typically, such a trust relationship already exists within the organization. The natural solution then is to use an in-house information retrieval appliance or system to index documents and enable retrieval.

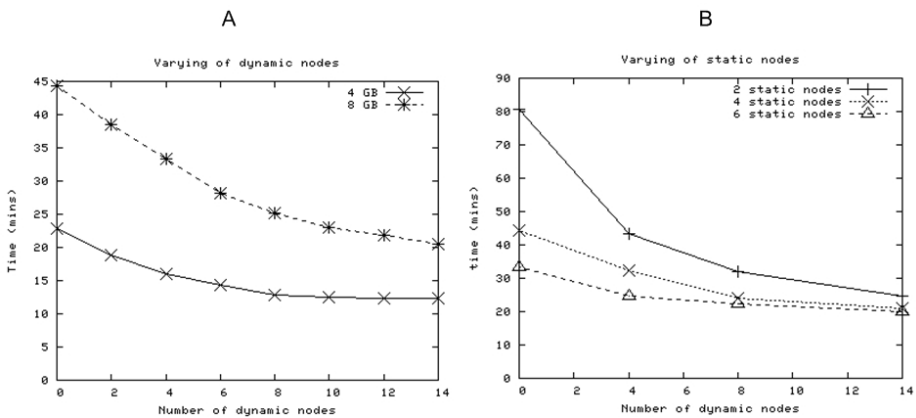
## 2 Scavenger Grid-Based Indexing

The premise of this work is the use of idle workstations in a scavenger Grid environment to handle medium to large-scale search engine indexing, using standard tools and techniques. A scavenger Grid or “cycle-stealing” Grid is a distributed computing environment made up of underutilized computing resources in the form of desktop workstations, and in some cases even servers, that are present in most organizations. The scavenger Grid proposed in this study is a combination of dedicated processing nodes (static nodes) and non-dedicated processing nodes (dynamic nodes) for indexing. A prototype search engine designed for a scavenger Grid environment was developed as a proof of concept. The Grid used in this research makes use of a local scheduler and “cycle stealing” technology, Condor [3,2], and a storage middleware solution, the Storage Resource

Broker (SRB) [1]. SRB is client-server middleware that virtualizes data space by providing a unified view of multiple heterogeneous storage resources over a network.

### 3 Evaluation

Tests were carried out to assess performance of Grid-based indexing in an actual scavenger Grid. In the first scenario, Fig. 1 part A, the number of static nodes was kept constant at 4 while the number of dynamic nodes was varied. In the second scenario, Fig. 1 part B, in addition to the dynamic nodes, the number of static nodes were also varied. The results in Fig. 1 show that the task of indexing can be performed by idle workstations without relying on a large fixed set of dedicated resources.



**Fig. 1.** Indexing performance of different numbers of dynamic nodes (part A) and Indexing performance of different numbers of static nodes (part B)

### 4 Conclusions

This study has looked at the use of resources already at an organization's disposal in the form of a scavenger Grid to provide cost-effective scalability within an intranet while retaining control over who can access the data. The results show that as additional resources become available, there are performance gains which make up for the inevitable grid middleware overhead.

### References

1. Baru, C.K., Moore, R.W., Rajasekar, A., Wan, M.: The SDSC storage resource broker. In: Proceedings of the conference of the Centre for Advanced Studies on Collaborative Research (1998)
2. Litzkow, M., Livny, M.: Experience with the condor distributed batch system. In: Proceedings of the IEEE Workshop on Experimental Distributed Systems (1990)
3. Condor: High Throughput Computing (2007), <http://www.cs.wisc.edu/condor/>



# A Study of Web Preservation for DMP, NDAP, and TELDAP, Taiwan

Shu-Ting Tsai and Kuan-Hua Huang

Institute of Information Science, Academia Sinica, Taipei, Taiwan  
{cxcdx, rane}@iis.sinica.edu.tw

**Abstract.** National Digital Archives Program (NDAP, Taiwan), spanning the period 2002 to 2007, emphasized on both humanities and technology. It has selected and digitized the most representative cultural heritage in Taiwan. Via the Union Catalogs, users can search and browse millions of multimedia resources with proper keywords or the built-in categories. However, since the complete digital resources are distributed in individual websites and databases built by different projects across different institutions, they may become unaccessible due to lack of maintenance after the associated projects have finished. Therefore, a pilot project has been launched to study the Web preservation strategy in order that the related websites of NDAP, its initiative project - Digital Museum Projects (DMP, 1998-2002), and its extended project - Taiwan e-Learning and Digital Archives Program (TELDAP, 2008-2012) can be well preserved for promotion, management, and development. This paper presents the results of the pilot study on the 27 DMP thematic websites.

The pilot study of Web preservation focused on the 27 thematic websites built by the Digital Museum Projects of the initiative phase of NDAP [1]. All websites were built by combining technical skills and content knowledge of experts in different fields; hence they are valuable and should be preserved.

There are two major works in this study. The first work is to investigate the operation of the 27 DMP thematic websites. We tested each website and gathered information from the related literatures and Web. We also contacted these projects' institutions to confirm the conditions of the websites by using questionnaires and phone interviews. The second work includes data collection and online exhibition. We designed several different preservation strategies for different websites.

Our investigation shows that, among the 27 thematic websites, 2 websites have been closed while the URLs of 11 websites have been changed. Some websites have either lost parts of contents or have numerous error or dead links, although they are still basically accessible. Half of the projects refused to provide data to us. The reasons include: 1) they hope to continue maintaining their websites by themselves; 2) they have difficulties in transferring data to us because the project has finished and the related staffs have left; and 3) the right and responsibility are unclear due to the transfer of the principal investigators of the projects. According to the survey results and the experiences of other Web preservation initiatives [2][3], we propose a workflow in Fig. 1. It integrates various preservation methods (cf. grey background items in Fig. 1). In this way, we have preserved 4 websites and imported 10 websites' digital resources into the Union Catalogs (<http://catalog.digitalarchives.tw>).

Web preservation procedure for NDAP and TELDAP will follow the workflow and methods in Fig. 1. Since some projects involved in this pilot study could not grant a sub-license to us for preservation and promotion because they were not authorized from the original copyright owner to do that, we have suggested the TELDAP Program Office and the National Science Council of Taiwan to draw up a contract to make sure that the results of the future projects can be used by Web preservation for promotion, management, and development in the portal <http://digitalarchives.tw>.

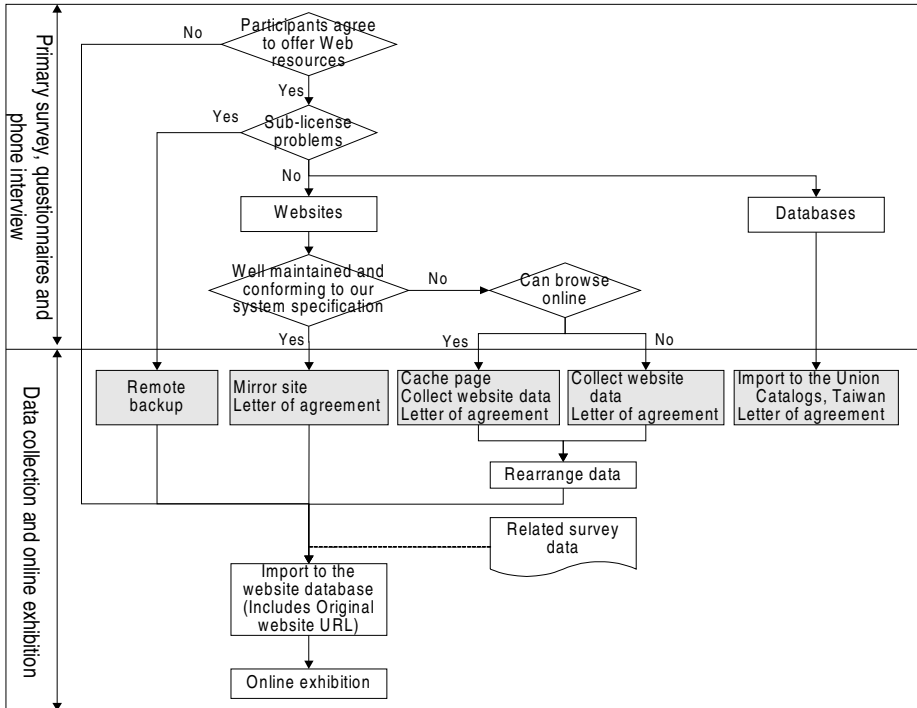


Fig. 1. The proposed Web preservation workflow

**Acknowledgements.** This work was supported by Taiwan e-Learning and Digital Archives Program (<http://teldap.tw>), sponsored by the National Science Council, Taiwan, under Grants: NSC96-2422-H001-009 and NSC96-3113-H001-015.

## References

1. Chen, H.H.: The Development of Digital Library Projects in Taiwan. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 556–558. Springer, Heidelberg (2006)
2. Arms, W.Y., Adkins, R., Ammen, C., Hayes, A.: Collecting and Preserving the Web: The Minerva Prototype. RLG DigiNews 5(2) (2001)
3. Hung, S.F.: Web Preservation: with Introduction to Required Techniques and Applications. National Central Library Bulletin 95(2), 75–121 (2006)

# Measuring Public Accessibility of Australian Government Web Pages

Yang Sok Kim, Byeong Ho Kang, and Raymond Williams

School of Computing, University of Tasmania, Sandy Bay,  
7005 Tasmania, Australia  
{yangsokk, bhkang, r.williams}@utas.edu.au

**Abstract.** We measured the public accessibility of Australian government Web pages, using search engines' crawling coverage and delay. Our results demonstrate that search engines crawled 73% ~ 83 % of newly published Web pages, exhibiting a delay of from 6.7 to 14.0 days.

**Keywords:** e-government, search engines, coverage, delay.

As the Web has become the predominant e-government service, delivering government information through the Web, has substantially replaced traditional approaches and impacted on the traditional roles of libraries [1]. Web search engines are a dominant Web information finding technology and it is important to investigate how well they provide government Web information in relation to public accessibility. This research measured the coverage and delay of search engine services in order to estimate the public accessibility of Australian government Web pages. In this research, the coverage refers to the proportion of the published information collected by a search engine and the delay refers to how long it takes to be made available by a search engine. The newly uploaded web pages were collected from the 249 Australian government Web pages, and were divided into five sub-domains (Federal government homepages (FH), Federal government media release pages (FM), Tasmanian government homepages (TH), Tasmanian government media release pages (TM), and Council homepages (CH)). They were collected from 1st January, 2007 to 15th July, 2007 (28 weeks). 166 Web pages out of 249 registered monitoring Web pages (66.7%) published one or more new Web pages during this experimental period.

**Table 1.** Coverage of Search Engines

Domain	Collected Webpages	Google		Yahoo		MSN		Composite	
			%		%		%		%
FH	947	823	87%	757	80%	786	83%	896	95%
FM	3150	2643	84%	2595	82%	2633	84%	2965	94%
THs	1007	795	79%	493	49%	695	69%	880	87%
TM	703	643	91%	598	85%	230	33%	654	93%
CH	420	242	58%	218	52%	206	49%	266	63%
Total	6227	5146	83%	4661	75%	4550	73%	5661	91%

A total of 6,227 pages were collected from these Web pages during the experimental period, as summarized in Table 1. Table 1 summarizes coverage results for the three major Web search engines, Google, Yahoo, and MSN. It is evident that Google provided the highest coverage (84%), followed by Yahoo (75%) and MSN (73%). This means that search engines still missed from 16% to 27% of all Web pages posted. Google also contributed a higher coverage performance in all domains when compared with other search engines, with the exception of MSN coverage of the Federal media release pages (84%). Table 1 also illustrates the composite coverage of three search engines in the last column, demonstrating that a meta-search engine can potentially collect up to 92% of newly published web information by collectively considering each individual search engine's coverage.

**Table 2.** Delay of Search Engines

Page Type	Google	Yahoo	MSN	AVG(A)	Composite(B)	Gap(A-B)
FH	4.2	13.6	12.6	10.1	3.5	6.6
FM	6.1	14.7	13.6	11.5	4.9	6.6
THs	14.9	21.0	11.9	16.0	10.5	5.5
TM	2.7	5.6	14.6	7.7	2.2	5.5
CH	5.6	15.2	13.1	11.3	4.3	7.0
Overall	6.7	14.0	13.2	11.3	5.1	6.2

Each search engine's overall delay was significantly different, as summarized in Table 2. Google had the shortest delay time (6.7 days), followed by MSN (13.2 days) and Yahoo (14.0 days). Google produced a superior performance in every domain except in the Tasmania Government homepages (14.9 days), where MSN offered the lowest delay time (11.9 days). The composite delay was calculated by taking the minimum average delay of each search engine for each Web page. The overall composite delay of the three search engines significantly decreases, by 1.6 days for Google, by 8.1 days for MSN, and by 8.9 days for Yahoo due to the temporal differences between the search engines. The composite results show that even though Google is the dominant performer in all domains, the delay time in relation to the Tasmania Government homepages may be significantly improved when Google is combined with other search engines.

Our research results demonstrate that commercial search engines cover 73% ~ 83 % of newly published Australian government information, exhibiting a delay from 6.7 to 14.0 days. These results support the possibility that Web monitoring systems may be used as a complementary web information finding method for Web search engines. In addition, coverage and delay characteristics of search engines are significantly improved by combining their capabilities. These results are significant evidence to support the effectiveness of meta-search engines.

## Reference

- [1] Missingham, R.: Access to Australian Government information: A decade of change 1997–2007. *Government Information Quarterly* 25(1), 25–37 (2008)

# Named Entity Recognition for Improving Retrieval and Translation of Chinese Documents

Rohini K. Srihari<sup>1</sup> and Erik Peterson<sup>2</sup>

<sup>1</sup> State University of New York at Buffalo, Buffalo, NY, USA

rohini@cedar.buffalo.edu

<sup>2</sup> Janya Inc., Buffalo, NY, USA

epeterson@janyainc.com

**Abstract.** This paper focuses on named entity recognition corresponding to people, organizations, locations, etc. in Chinese scientific documents. Two key benefits are shown by performing NER: (i) improved quality of semantic retrieval, and (ii) improvement in subsequent machine translation. Experiments using the Semantex platform for information extraction illustrate and quantify the two benefits outlined.

## 1 Introduction

This paper focuses on using native language Chinese entity tagging in order to improve two common tasks: (i) cross-lingual retrieval, and (ii) machine translation of Chinese names. Named entity recognition (NER) focuses on the proper detection and classification of proper noun sequences into semantic categories such as *person name*, *organization name*, *location name* etc. The Automatic Content Extraction (ACE) competition [3] has stimulated much of the recent NER research since it provides annotated data which can be used for training and evaluation of NER systems. NER is the first major step in the more comprehensive task of information extraction (IE).

In this paper, we focus on Chinese named entity recognition. The scenario of interest is cross-lingual retrieval of Chinese scientific documents, with English as the query language. There are two main issues in this task which necessitate the use of NER. The first issue is that simple keyword retrieval systems are not sufficient to handle queries of interest. Researchers are interested in queries such as *find people associated with alternative fuel for aerospace applications*. Keyword querying, while very efficient for document retrieval, is not sufficient to respond to such queries. It is necessary to index the documents, identify key topics, and identify named entities. Thus, a response to such a query could first filter documents based on topic match, and subsequently return people names.

The second issue involves the poor quality of name translation. In cases where Chinese documents are translated by a machine translation (MT) system into English to facilitate searching and browsing, the user often obtains poor search results due to name translation errors.

In this paper, we show that by performing native Chinese tagging and categorizing of named entities prior to MT, the quality of subsequent name translation,

and even overall translation results can be improved. Current MT systems are evaluated by methodology such as Bleu [1] where systems can score highly even with poor name translation, which would cause poor search results for information retrieval. Related work includes Chinese name detection and translation. Once a Chinese-national name has been identified, it is fairly easy to convert it to a English version, using a character to romanization table. However, it is for Chinese transliterations of non-Chinese names that most work has been done [2]. [4] describes an effort which combines monolingual entity detection and coreference in English and Chinese, two different name transliteration schemes, and a commercial machine translation system in order to perform Chinese to English name translation.

## 2 Semantex: Multilingual Information Extraction Platform

Semantex [5] is a domain-independent, modular, scalable information extraction engine. The Semantex engine reflects a hybrid approach to natural language processing, where lexical, grammatical and statistical approaches are exploited at appropriate stages of the pipeline. Semantex now uses Unicode internally in order to process other languages besides English.

The first step necessary for Semantex processing is tokenization, breaking the incoming text is smaller units such as words and punctuation. Since Chinese does not mark word boundaries in text, spaces cannot be used for tokenization. Segmentation is treated as a sequence tagging problem, with the tags marking the start of words. Once the text is tokenized, the entity tagger we have built for Chinese uses three different and complementary paradigms in order to accurately tag entities: (i) lexical look-up, (ii) grammars, and (iii) machine-learning approach. After several months of development, the system f-score for internal tests on Chinese named entities is 72.7%. As an example, 斯洛文尼亚总理扬沙 is translated as “Slovenia premier the sand blowing” by Babelfish. Semantex is able to identify 扬沙 as name, allowing a query looking for all names in the document to find it.

## References

1. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation (2001)
2. Wan, S., Verspoor, C.: Automatic english-chinese name transliteration for development of multilingual resources. In: COLING-ACL, pp. 1352–1356 (1998)
3. Doddington, G., Mitchell, A., Pryzbocki, M., Ramshaw, L., Weischedel, R.: The Automatic Content Extraction (ACE) Program, Tasks, Data, and Evaluation (2004)
4. Ji, H., Blume, M., Freitag, D., Grishman, R., Khadivi, S., Zens, R.: Nyu-fair isaac-rwth chinese to english entity translation 2007 system. In: Proceedings of NIST ET 2007 PI/Evaluation Workshop, Washington, USA (2007)
5. Srihari, R.K., Li, W., Cornell, T., Niu, C.: Infotextract: A customizable intermediate level information extraction engine. *Natural Language Engineering* 12 (2006)

# Current Approaches in Arabic IR: A Survey

Mohammed Mustafa, Hisham AbdAlla, and Hussein Suleman

Department of Computer Science, University of Cape Town  
Private Bag X3, Rondebosch, 7701, South Africa  
{mmustafa,hisham,hussein}@cs.uct.ac.za

**Abstract.** Arabic information retrieval is a popular area of research. This paper presents the current state-of-the-art in Arabic Information Retrieval (IR) approaches. Moreover, it provides general guidance for open research areas and future directions.

## 1 Introduction

Modern Standard Arabic (MSA) is one of the most widely used languages in the world. Previous works in the field of Arabic IR summarize five features of the Arabic language that cause it to be a significant challenge for both information retrieval and search engines: **Orthographic Variations**, its complex **Morphology**, **Diacritization**, prevalence of **Irregular / Broken Plural**, and **Synonyms**. The following are some examples for these challenges: Orthographic Variations - either بُورْتَسُوْدَان or بُورْسُوْدَان for Port-Sudan city; Morphology - لَهْدِيْدِيْهْم (meaning: we will surely guide them) is a word that consists of different parts; Diacritization - الشَّعْر, means hair, while الشِّعْر means poem; Broken Plural - the word قَائِد (meaning: leader) changes to قُوَاد (meaning: leaders); and Synonyms - the word أَسَد (meaning: lion) may have different synonyms to lion according to its age.

## 2 Current Solutions

The above challenges have been solved to different levels. Preprocessing includes removal of non-characters, normalization and removal of stopwords. The non-character removal step [1] includes the removal of punctuation marks, diacritics and kasheeda (the word السُّوْدَان (Sudan) can be written with kasheeda as السُّوْدَان). Normalization is used to represent different forms of a letter with a single Unicode representation as in HAMZA (أَ, إ) and MADDA (آ). For stopwords and their phrases most existing methods use dictionaries or software tools. Tokenization is used intensively in Arabic IR by using different techniques.

A number of studies have been devoted to different approaches of incorporating morphology: stem, root, light stem and no-stemming as well as using

non-rule based statistical or n-gram models. Stemming affects all problems mentioned. Kadri and Nie [3] proposed a new stemming technique based on linguistic removal of affixes. Larkey et al [4] presented the best light stemmers (*light10*). Mansour et al [5] proposed a new technique based on Arabic grammatical rules.

For the problem of broken plural, the most used algorithm was proposed by Goweder et al [2]. Other complementary techniques are in regional variation disambiguation [1]. Also, Abdelali et al [1] presented a query expansion mechanism that has the ability to automatically select a corpus related semantically to the query.

### 3 Current Challenges

Arabic IR has a long road ahead. Major problems of some current Arabic search engines are that queries can only retrieve exact matching documents. Researchers need to devise more effective solutions for Arabic IR. Lack of resources for test-beds, translation and query expansion is one of these challenges. Regional variations, spelling variations and machine translations are contributors to problems in this field. Combinations of resources such as dictionaries, query logs and Web mining techniques is needed for both translation and query expansion. Preprocessing techniques need to be united. Tokenization still has its challenges due to clitics and ambiguity. Linguistic stemming is a very rich area of research. A straightforward algorithm is needed for broken plurals. Automatic diacritization also needs further work. It is expected that the next generations of Arabic search engines will be based on Arabic morphology.

### References

1. Abdelali, A.: Improving Arabic Information Retrieval using Local variations in Modern Standard Arabic, PhD. dissertation, New Mexico Institute of Mining and Technology (2006)
2. Goweder, A., Poesio, M., De Roeck, A.: Broken plural detection for Arabic information retrieval. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, pp. 566–567 (2004)
3. Kadri, Y., Nie, J.Y.: Effective stemming for Arabic information retrieval. In: The challenge of Arabic for NLP/MT Conference, The British Computer Society, London, UK (2006)
4. Larkey, S.L., Ballesteros, L., Connell, E.M.: Light stemming for Arabic information retrieval. Arabic Computational Morphology: Knowledge-based and Empirical Methods (2005)
5. Mansour, N., Haraty, A.R., Daher, W., Hourri, M.: An auto-indexing method for Arabic text. Information Processing and Management 44(4), 1538–1545 (2008)



# A Bilingual Information Retrieval Thesaurus: Design and Value Addition with Online Lexical Tools

K.S. Raghavan and A. Neelameghan

Indian Statistical Institute, Bangalore 560 059, India  
raghavan@isibang.ac.in, anm2002@vsnl.net

**Abstract.** Illustrates value addition to a thesaurus using available online lexical tools to enhance retrieval and support knowledge discovery based on a project for building a bilingual thesaurus as a component of a search interface to a digital library of Tamil classics being developed.

**Keywords:** Bilingual Thesauri, Tamil-English Thesaurus, Information Retrieval, Knowledge Discovery.

## 1 Introduction

A digital library of early Tamil works including Sangam<sup>1</sup> classics and other related works is being developed under a Government of India project. The search interface for the library will have, as a component, a bilingual thesaurus (Tamil – English) to assist intelligent query formulation and query expansion. The issues related to multilingual thesauri in culture-specific domains and handling of semantic relationships have been discussed in earlier papers [1, 2, 3]. The thesaurus being maintained as a WINISIS database has over 85000 descriptors with an index of over one million terms. The complexity of the language and multiplicity of relations necessitated adoption of appropriate strategies. The strategies adopted included:

- Extensive linking of records within the thesaurus database; and
- Linking to other online lexical resources

## 2 Issues and Strategies

- a) Large number of NTs & RTs: Two strategies were employed; postings were made in one record with links from other descriptors or lists of RTs and NTs were maintained as separate external files and database records were linked to this file
- b) Three factors contributed to inadequacies in providing comprehensive semantic maps of descriptors: Large number of Homographs in Tamil, which has to do with the evolution in the meaning and connotation of terms in Tamil (Neelameghan, 2008); Web of relations; and specialized needs of user

---

<sup>1</sup> Historians refer to the Tamil literature from ca. 300 B.C. to 300 A.D. as Sangam literature.

community (E.g. the need to include titles of classics, commentaries, etc as descriptors). It was realized that substantial value addition including providing a more complete semantic map of descriptors and enhancement of retrieval capability could be achieved by linking the thesaurus to online lexical tools. Every record in the thesaurus is linked to the online Tamil Lexicon (See Fig. 1).

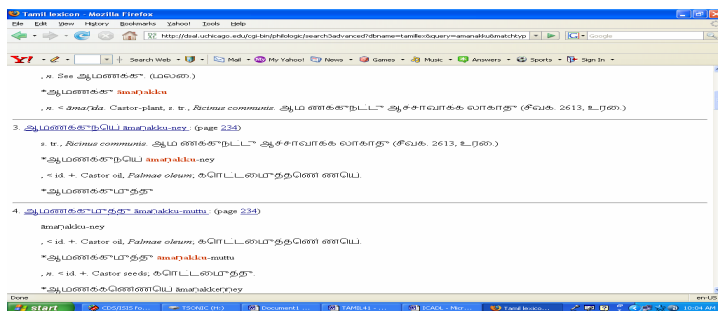


Fig. 1. Display from Online Tamil Lexicon (partial view)

## References

1. Raghavan, K.S., Neelameghan, A.: Design and Development of a Bilingual Thesaurus for Classical Tamil Studies: Experiences and Issues. In: Arsenault, Clement, Tennis, Joseph, T. (eds.) Culture and identity in knowledge organization: Proceedings of the Tenth International ISKO Conference, pp. 70–76. Ergon Verlag (2008)
2. Neelameghan, A., Raghavan, K.S.: An online multi-lingual, multi-faith thesaurus: a progress report on F- THES. Webology (2005), <http://www.webology.ir/2005/v2n4/a19.html>
3. Neelameghan, A., Raghavan, K.S.: Online bilingual thesaurus for subjects in the humanities: a case study. In: Prasad, A.R.D., Madalli, D.P. (eds.) ICSD 2007: International Conference on Semantic Web & Digital Libraries, pp. 489–505. DRTC, Bangalore (2007)
4. Neelameghan, A.: Homographs, homonyms and cultural practices in relation to Tamil-English thesaurus. Information Studies 14, 99–110 (2008)

# Entity-Based Classification of Web Page in Search Engine

Yicen Liu, Mingrong Liu, Liang Xiang, and Qing Yang

Institute of Automation, Chinese Academy of Sciences,  
Beijing 100190, China  
{ycliu, mrliu, lxiang, qyang}@nlpr.ia.ac.cn

**Abstract.** There are several difficulties in integrating traditional classification approaches in a search engine. This paper presents an Entity-Based Web Page Classification Algorithm, which can be embedded in search engine easily. In the algorithm, we build up an Entity System to classify web pages immediately before indexing jobs. It is an assistant system used in text feature selection and can be updated incrementally. Experimental results show its efficiency, compared to the traditional ones and has a good performance.

**Keywords:** information retrieval, web-page classification, entity system.

## 1 Introduction

In this paper, we present an entity-based classification algorithm, which can be embedded in search engine easily. The algorithm is based on the cognition that if most of the classifiable entities in a text belong to certain categories, it is reasonable to classify the text to these categories. In the algorithm, we maintain an entity system, in which all the entities are well classified. With its help, the web page is classified using the categories of its entities. Experimental results show that the best performance of our algorithm is as good as KNN classifier on micro F1 measure [1], and achieves an acceptable result in a search engine.

## 2 Entity-Based Classification of Web Page

Unlike any other traditional classification algorithms [3] [4] [5] [6] [8], we need to construct an entity system in order to classify the web pages. In the system, a dictionary and a principle to recognize an entity should be defined initially, then a part of entities that can be categorized easily and correctly by human editor are organized as a pre-defined category dictionary. At the same time, an entity relationship database should be extracted from certain documents. The approach is to collect all of the entities for each document, each pair of entities in which is saved in the database. When the relationship database is obtained, the method to get the categories of an entity is to check which pre-defined category dictionary the entity and its related entities belong to, and to return the category.

Entity-based classification algorithm makes full use of the entity system to get the text feature for categorization. Before classifying, a web page needs to be transferred to pure text at first, then entities are extracted with an entity system. The categories for each entity are got together, and the category with the highest weight is the result.

### 3 Experiments

In our experiments, all of the web pages are crawled on the Internet. We apply both entity system and  $\chi^2$  statistic [2] [7] with a weighted factor  $\omega$  as follow:

$$W(t, c) = \omega * \chi^2(t, c) + (1 - \omega) * ES(t, c). \quad (1)$$

which means that the weight term  $t$  belongs to category  $c$  is a linear combination value of both  $\chi^2$  statistic and category weight in Entity System.

Experimental results show that when  $\omega = 0$ , the algorithm only use an entity system for feature selection, and the micro-average F1 measure is about 0.752. F1 measure achieves its maximal value 0.837 when  $\omega = 0.7$ , compared to 0.827 when using KNN classifier. When the algorithm is embedded in a search engine and only use entity system for feature selection, F1 measure is about 0.712.

### 4 Conclusions

In this paper, we propose a novel classification algorithm for web pages. Experiments have shown that it is possible to use the local information of a text to classify web pages, and the evaluation of micro-average F1 measure is acceptable. It is proved that the algorithm is efficient and can be used in search engine.

### References

1. Yang, Y.: An Evaluation of Statistical Approaches to Text Categorization. In: Proc. of Information Retrieval, pp. 69–90 (1999)
2. Yang, Y., Pedersen, J.: A Comparative Study on Feature Selection in Text Categorization. In: Proc. of ICML, pp. 412–420 (1997)
3. Kwon, O., Lee, J.: Web Page Classification Based on K-Nearest Neighbor Approach. In: Proc. of IRAL (2000)
4. Shen, D., Chen, Z., et al.: Web-Page Classification Through Summarization. In: Proc. of SIGIR, pp. 242–249 (2004)
5. Christopher, J., Burges, C.: A Tutorial on Support Vector Machines for Pattern Recognition. In: Proc. of Data Mining and Knowledge Discovery, pp. 121–167 (1998)
6. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Proc. of ECML, pp. 137–142 (1998)
7. Liu, T., Liu, S., et al.: An Evaluation on Feature Selection for Text Clustering. In: Proc. of ICML, pp. 488–495 (2003)
8. Han, J., Kamber, M.: Data Mining Concepts and Techniques, 2nd edn., pp. 285–375

# MobiTOP: Accessing Hierarchically Organized Georeferenced Multimedia Annotations\*

Thi Nhu Quynh Kim<sup>1</sup>, Khasfariyati Razikin<sup>1</sup>, Dion Hoe-Lian Goh<sup>1</sup>,  
Quang Minh Nguyen<sup>1</sup>, Yin Leng Theng<sup>1</sup>, Ee-Peng Lim<sup>2</sup>, Aixin Sun<sup>3</sup>,  
Chew Hung Chang<sup>4</sup>, and Kalyani Chatterjea<sup>4</sup>

<sup>1</sup> Wee Kim Wee School of Communication and Information, Nanyang Technological University

{ktnq, khasfariyati, ashlgoh, qmnguyen, tyltheng}@ntu.edu.sg

<sup>2</sup> School of Information Systems, Singapore Management University  
eplim@smu.edu.sg

<sup>3</sup> School of Computer Engineering, Nanyang Technological University  
axsun@ntu.edu.sg

<sup>4</sup> National Institute of Education, Nanyang Technological University  
{chewhung.chang, kalyani.c}@nie.edu.sg

**Abstract.** We introduce MobiTOP, a map-based interface for accessing hierarchically organized georeferenced annotations. Each annotation contains multimedia content associated with a location, and users are able to annotate existing annotations, in effect creating a hierarchy.

**Keywords:** Social tagging, social computing, hierarchical georeferenced annotations, user interface, participatory design.

## 1 The MobiTOP User Interface

With the increasing popularity of mobile devices with GPS capabilities, the tagging or annotating of locations with multimedia content is becoming common. In the spirit of social computing, a system supporting the sharing of georeferenced multimedia content should allow other users to include their annotations to existing content as well. This promotes a community of users to exchange and explore content and ideas.

Existing tagging systems employ tags clouds [2] or map-based visualization [3] which have their drawbacks. Here, we propose a new design (Figure 1) which utilizes both concepts to explore hierarchically organized annotations. This design was the outcome of a participatory design workshop. The resulting system called MobiTOP (Mobile Tagging of Objects and People) consists of: (1) a tree view of hierarchically organized multimedia annotations; (2) a thumbnail placeholder for displaying a series of annotations associated with the current location; and (3) a content panel which displays an annotation's content and its metadata together with a tag cloud showing the tags associated with it and its children.

---

\* This work is partly funded by A\*STAR grant 062 130 0057.

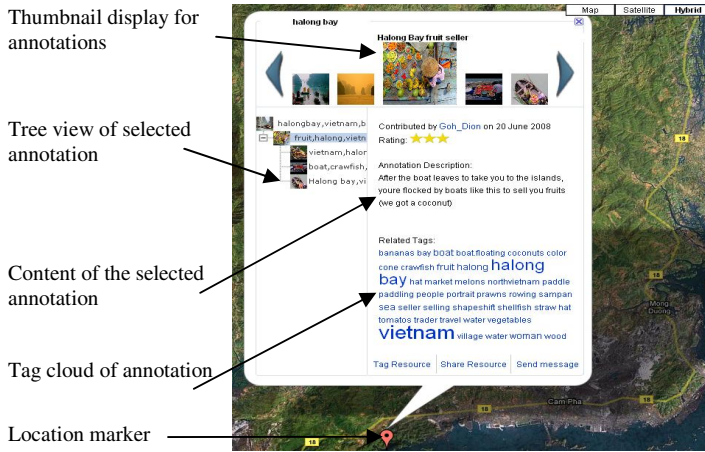


Fig. 1. The main user interface of the MobiTOP system

## 2 Pilot Evaluation

A pilot evaluation of the MobiTOP user interface was conducted to determine its usability by comparing it against a Space-Filling Thumbnail (SFT) [1] interface. Eight participants were given information seeking tasks and were asked qualitative questions. All, except one, of the participants preferred the MobiTOP user interface that presented annotations hierarchically even though it was slightly more difficult to grasp initially than the simpler SFT interface. Our preliminary results suggest that with sufficient time given to learning, participants would be able to effectively access content using the MobiTOP interface.

In the immediate future, and in line with our evaluation findings, more documentation will be built into the system to help users better understand and use the features of the MobiTOP system. Further, we plan to incorporate tag cloud generation algorithms that reduce the number of tags to be displayed in the content panel.

## References

- [1] Cockburn, A., Gutwin, C., Alexander, J.: Faster Document Navigation with Space-Filling Thumbnails. In: CHI 2006, Montréal, Québec, Canada, April 22-27, pp. 1–10. ACM, New York (2006)
- [2] Hassan-Montero, Y., Herrero-Solana, V.: Improving Tag-Clouds as Visual Information Retrieval Interfaces. In: InSciT 2006, Mérida, Spain, October 25-28 (2006)
- [3] Nguyen, Q.M., Kim, T.N.Q., Goh, D.H.L., Theng, Y.L., Lim, E.P., Sun, A., Chang, C.H., Chatterjea, K.: TagNSearch: Searching Photographs in Geo-referenced Collections. In: Christensen-Dalsgaard, B., Castelli, D., Ammitzbøll Jurik, B., Lippincott, J. (eds.) ECDL 2008. LNCS, vol. 5173, pp. 62–73. Springer, Heidelberg (2008)

# Social Tagging in Digital Archives

Shihh-Yuarn Chen<sup>1</sup>, Yu-Ying Teng<sup>2</sup>, and Hao-Ren Ke<sup>2,3</sup>

<sup>1</sup> Department of Computer Science, National Chiao Tung University,  
No. 1001, Ta-Hsueh Rd., Hsinchu 300, Taiwan

<sup>2</sup> Institute of Information Management, National Chiao Tung University,  
No. 1001, Ta-Hsueh Rd., Hsinchu 300, Taiwan

<sup>3</sup> University Library, National Chiao Tung University,  
No. 1001, Ta-Hsueh Rd., Hsinchu 300, Taiwan

sychen@cs.nctu.edu.tw, lal.teng@msa.hinet.net,  
claven@lib.nctu.edu.tw

Visitors of online digital museums or galleries usually encounter a problem – how to find out an artwork if neither the name nor author of the artwork is known. Social tagging can collect tags representing public users' feelings and opinions about an artwork, and these tags can help other users to find out artworks in which they are interested. This research applies social tagging, to Yuyu Yang Digital Art Museum (<http://yuyuyang.e-lib.nctu.edu.tw>)<sup>1</sup>. The advantages include the association of the artworks and visitors' feelings and opinions, the provision of indexes for users to browse and search, and the enrichment of the descriptions of artworks via public users' tags.

To achieve the goal, this research applies CKIP<sup>2</sup> to process the metadata of Yang's artworks, and extracts keywords for representing each artwork. Then, public users' tags of each artwork are collected. Finally, a tag hierarchy generation algorithm based on the co-usage of tags [1]. is applied to group these tags and extracted keywords, and a tag map for each category of Yang's artworks is established. Each tag map is based on professional information (extracted keywords from metadata) and public users' feelings and opinions (tags).

After the tag hierarchy generation process is finished, a tag map for a category can be shown to guide visitors to browse artworks of this category. The left of Fig. 1 is a portion of the tag map for the category *Painting*. Furthermore, as observed in the right of Fig. 1, for each artwork, in addition to the tags and system-generated keywords, the system recommends a few terms as well. These recommended terms are chosen from the parent nodes of tags and keywords of the artwork and the child nodes with significant similarities.

A preliminary evaluation is conducted via online questionnaires. A total of 113 effective responses from students of National Chiao Tung University are collected.

---

<sup>1</sup> Yuyu Yang (楊英風) is one Taiwanese international well-known sculptor and painter. The Yuyu Yang Digital Art Museum is sponsored by the National Digital Archives Program (NDAP).

<sup>2</sup> CKIP (Chinese Knowledge and Information Processing) is a tokenization and part-of-speech (POS) tagging system for Chinese documents, developed by the Institute of Information Science and the Institute of Linguistics of Academia Sinica (<http://rocling.iis.sinica.edu.tw/CKIP/engversion/index.htm>).

Totally 186 artworks are tagged and 1088 tags are given by users. About 50% and 60% people agree that the system-generated keywords and the user-given tags can represent the artworks, respectively. More than 60% people agree that social tagging is helpful to realize the idea of artworks, around 70% agree that social tagging is beneficial for sharing the ideas of public, and more than 60% agree that social tagging can help them find out other Yang's artworks. Furthermore, about 60% agree that the tag map can work as a guide of artworks, supply a quick way for browsing artworks, improve the convenience of browsing, and is really practical and useful.



Fig. 1. (Left) Tag map example. (Right) Snapshot for showing the keywords, tags, recommended terms, and tag map of a painting.

**Acknowledgements.** This work has been funded by the National Science Council through grants NSC97-2631-H-009-002.

**Reference**

[1] Hsieh, W.T., Lai, W.S., Chou, S.-C.T.: A Collaborative Tagging System for Learning Resources Sharing. In: Current Developments in Technology-Assisted Education, FOR-MATEX, Seville, Spain, pp. 1364–1368 (2006)



# Editor Networks and Making of a Science: A Social Network Analysis of Digital Libraries Journals

Monica Sharma and Shalini R. Urs

University of Mysore, Mysore, India  
{monica, shalini}@isim.ac.in

**Abstract.** We present here our research findings of the structural features of editorial boards (EB) of leading journals of digital libraries using social network analysis techniques and tools.

## 1 Introduction

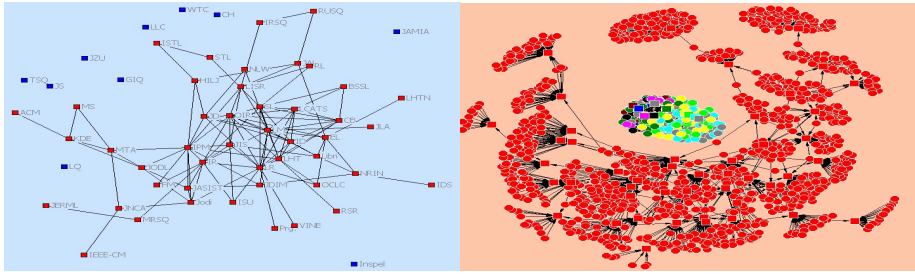
In the making and shaping of a science, academic publications especially peer reviewed journals play a very important role. The landscape and the direction of research are molded by two key players i.e. the editors and the authors of these journals. Our research is in the direction of understanding the centrifugal forces in shaping the field of digital libraries (DL) using social network analysis (SNA) tools. Structure and composition of DL community have been studied by the researchers on co-authorship network of journals and conferences [1,2]. Premising that the community structure of the Editorial Board (EB) reflects the knowledge structure of the field, we draw inferences on the dynamics of the DL. We attempt to find the nature of “Diversified Convergence” of DL as a discipline and also its most central journals and editors.

Using the Thomson Reuters Web of Knowledge (WoK) and publication count, the journals were selected. A topic search on “Digital Libraries” retrieved 7770 papers. A record count threshold of 5 pruned the list to top 250 journals and conferences. A record count threshold of 7 and complete editorial board information yielded 56 journals and 1453 distinct editors which formed the dataset for the study. The SNA metrics- degree, betweenness and closeness centrality form the metrics for analysis.

## 2 DL Journal Network and Editorial Board Network

An analysis of the WoK subject categorization of journals revealed that Computer Science, Information Science and Library Science dominate followed closely by Education and Educational Research. The DL discipline has converged from diverse subjects including social issues, law, health care and geography. Of the 56 journals, 10 journals were isolates (Fig. 1(a)). Library Management (LM) and Library Review (LR) have highest degree centrality (15) followed by Online Information Review (OIR) and Information Processing Management (IPM)(13 each). However IPM gets the most central position in the network acting as a bridge with highest betweenness score. LM and LR follow IPM.

The EB network formed 11 Components. Of these, the giant component constituted 84% of editors indicating that it is very well connected (Fig 1.(b)).



**Fig. 1.** Network of (a) DL Journals and (b) Editorial Board

**Table 1.** Ranking of Top 10 Editors based on centrality scores

Betweenness		Closeness		Degree	
Rank	Authors	Rank	Authors	Rank	Authors
1	Gary E Gorman	1	Gary Marchionini	1	Gary E Gorman
2	Edward A. Fox	1	Edward A. Fox	2	Thomas E. Nisonger
3	Mohammed Atiquzzaman	1	Jim Jansen	2	Gary Marchionini
4	Vijay Atluri	1	Amanda Spink	3	Alan Gilchrist
5	Beng Chin Ooi	1	Michael K Buckland	3	Amanda Spink
6	John Leggett	1	Gary E Gorman	3	Andrew Dillon
7	William Hersh	1	Ricardo Baeza-Yates	3	Charles Oppenheim
8	Andrew Booth	1	Nicholas Belkin	3	Derek Law
9	Ali Asghar Shiri	1	Christine L. Borgman	3	Edward A. Fox
10	Gary Marchionini	1	Alan Smeaton	3	Jennifer Rowley

Gary Gorman, Edward Fox and Gary Marchionini are the three editors who are the core members of the network playing a very important role in shaping policies of the journals and thus the DL communities (Table 1). Majority of editors are from USA, constituting 46.7%. This is followed by UK constituting 11%.

### 3 Conclusion

DL is a well-connected community converging from diverse disciplines with Computer Sciences as the common thread. Though LR and LM top the ranking of journals in terms of degree centrality, IPM is central in terms of betweenness. G E Gorman, E. Fox and G. Marchionini play an important role with a high centrality scores. E. Fox having figured in all the three studies of DL [1, 2, 3] appears to be the star of the DL community.

### References

1. Sharma, M., Urs, S.R.: Small world phenomenon and author collaboration: How small and connected is the digital library world? In: Proceedings of the 10th International Conference on Asian Digital Libraries, pp. 510–511. Springer, Vietnam (2007)
2. Sharma, M., Urs, S.R.: Mapping network structure of digital library research of CiteSeer database. In: Proceedings of the Workshop on Mining Social Data, 18th European Conference on Artificial Intelligence, Greece, pp. 6–10 (2008)
3. Liu, X., et al.: Co-Authorship Networks in the Digital Library Research Community. Information Processing & Management 41(6), 1462–1480 (2005)

# Empowering Doctors through Information and Knowledge

Anjana Chattopadhyay

National Medical Library, Ansari Nagar, Ring Road, New Delhi-110029, India  
anjanachattopadhyay@yahoo.com

**Abstract.** Health information dissemination is essential to empower doctors to prevent death of millions of people from major killer diseases due to unawareness and lack of resources in poor countries. There lies a big '10/90' gap between the magnitude of global health problem and distribution of health resources between developed and developing nations. The existing system is facing serious challenges because of mismatch between the burden of infectious diseases and scarcity of adequate budget to combat these diseases. The World Health Organization started HINARI to provide online medical journals to over 100 developing nations. National Medical Library (India) also started ERMED (Electronic Resources in MEDicine), online journal consortium to provide global medical literature to medical professionals.

**Keywords:** Medical Library, Medical Informatics.

## 1 Introduction

The most important task before us is to provide medical health knowledge to those, who need them most. Approximately 50 million people die every year from poor countries due to infectious killer diseases. We may save at least 2/3<sup>rd</sup> of them by providing appropriate information and facilities to doctors of these regions. 90% of global fund is spent by OECD countries, where only 10% of global killer disease exists. The remaining 10% fund is utilized by developing countries, which are struggling with the 90% of global killed diseases including pneumonia, HIV/AIDS, Diarrhea, tuberculosis, Malaria and Measles. Global Forum for Health Research has been established to look into the existing anomalies and relocation of fund [2].

Heavy clinical duties of doctors and lack of sensitization stands as a hindrance to the progress of vibrant medical research. They remain totally ignorant about new innovations in health research. Access to quality education and opportunity to get exposure to experts in various research fields of specialization to gain experience are very rare. The ever increasing price of international journals cause literature crisis in many academic institutions. They find it very difficult to procure even basic medical books and journals required for their scholars. Students remain deprived of quality information for their research. National Medical Library, New Delhi, procures over 1600 medical journals, and shares them by dissemination of over 8000 photocopy of articles per month to doctors across the country.

## 2 Initiative Taken by Various Organizations

WHO took initiative to bridge the knowledge gap HINARI (Health Inter Network Access to Research Initiative) in January, 2002 for over 100 developing countries [1]. The advantage of HINARI has been restricted to countries having GNP (Gross National Product) less than US\$1000 per capita per year (as per the World Bank Report in 2001). Unfortunately India is not covered under this category.

The National Medical Library, India, started ERMED (Electronic Resources in MEDicine) electronic journal consortium through [www.nlmrmed.in](http://www.nlmrmed.in) since January, 2008 for 40 medical colleges/institutions in India. ERMED provides electronic journal resources from over 1515 medical journal package in a very competitive price. The new cost effective purchasing trend has made valuable impact on medical colleges, which remained deprived of global literature for years. The majority of health research is conducted in rich countries and their output reflects the health problem related to their region. Innovation of new drugs for treatment of tropical diseases is rare. Drugs targeted for Western population are often tested on subjects of poor regions.

## 3 Development of Indigenous Information and Knowledge

Creation of health care literature as per the requirement of local problem is the foremost priority in health research. Very few research papers from developing countries are published in journals indexed by Pubmed, therefore research from these countries are almost invisible on international platform. India took initiative to develop various indigenous information and knowledge resources in the field of biomedical sciences such as IndMed database and MedInd full-text service from Indian medical journals. It also started the Traditional Knowledge Digital Library (TKDL), Revitalization of Local Health Traditions (FRLHT) and National Manuscript Mission (NMM) to conserve old heritage of medical systems.

## 4 Conclusion

Health research may be taken up as the most effective tool to improve public health system of a country. It is hoped that the initiatives taken by various organizations would help to empower doctors through advance information and knowledge to deliver "Better Health for all".

## References

1. Health Inter-Network Access to Research Initiative, <http://www.who.int/hinari/en/index.html>
2. Yong Voice, Research for Health 2007, Global Forum for Health, Lancet (2007)

# Author Index

- Aalberg, Trond 327  
AbdAlla, Hisham 406  
Achananuparp, Palakorn 203  
Adriani, Mirna 276  
Allen, Robert B. 379  
Alshuhri, Sulieman Salem 387  
Antunes, Gonçalo 225
- Bainbridge, David 236, 294  
Ball, Gregory R. 134  
Barateiro, José 225  
Beel, Jöran 375  
Bennett, Erin 367  
Bloom, Mohan John 396  
Borbinha, José 174, 225, 256  
Burrows, Toby 394
- Cabral, Manuel 225  
Chang, Chew Hung 412  
Chang, Hung-Chi 331  
Chang, Kuiyu 266  
Chatterjea, Kalyani 412  
Chattopadhyay, Anjana 418  
Chen, Shih-Yuarn 414  
Chidananda Gowda, K. 371  
Chua, Alton Yeow Kuan 22, 51, 396  
Cunningham, Sally Jo 367
- Damrongrat, Chaianun 339  
Dang, Dinh-Trung 313  
Datta, Anwitaman 266  
Dong, Li 351
- Freire, Nuno 174, 256
- Gerber, Anna 246  
Gipp, Béla 375  
Goh, Dion Hoe-Lian 22, 51, 396, 412  
Gong, Zhiguo 12  
Gunjal, Bhojaraju 194
- Ha, Inay 215  
Haruechaiyasak, Choochart 335, 339  
Hasan, Md Maruf 104  
Hsiao, Jen-Hao 331
- Hu, Xiaohua 203  
Huang, Kuan-Hua 400  
Huang, Michael Bailou 304  
Hunter, Jane 246
- Ishikawa, Yoshiharu 82  
Ito, Tetsuro 309
- Jiang, Airong 347  
Jin, Fusheng 61  
Jo, Geun-Sik 215  
Jones, Matt 294  
Jones, Steve 294
- Kan, Min-Yen 313  
Kaneko, Yasufumi 71  
Kang, Byeong Ho 402  
Ke, Hao-Ren 414  
Khaltarkhuu, Garmaabazar 41  
Khoo, Christopher S.G. 184  
Khy, Sophoin 82  
Kil, Hyunyoung 1  
Kim, Heung-Nam 215  
Kim, Thi Nhu Quynh 412  
Kim, Yang Sok 402  
Kiran, Kaur 285  
Kitagawa, Hiroyuki 82  
Kongthon, Alisa 335  
Körber, Nils 31  
Krapivin, Mikalai 144
- Lang, Haiyang 61  
Lee, Chei Sian 22, 51  
Lee, Dongwon 1  
Lee, Seung-Hoon 215  
Li, Chao 347  
Li, Guohui 304  
Li, Xiaoming 359  
Lim, Ee-Peng 266, 412  
Lin, Jin 351  
Liu, Jia 309  
Liu, Jyi-Shane 343  
Liu, Mingrong 410  
Liu, Yicen 410

- M.A., Angrosh 355  
 Ma, Ningning 347  
 Maeda, Akira 41  
 Marchese, Maurizio 144  
 Martins, Bruno 174, 256  
 Matsushima, Shintaro 317  
 Maureen 266  
 McIntosh, Sam 294  
 Morii, Masumi 317  
 Morishita, Satoshi 114  
 Moyle, Martin 154  
 Mustafa, Mohammed 406  
  
 Na, Jin-Cheon 184  
 Nakamura, Satoshi 71, 124  
 Nakashima, Makoto 309  
 Nakashole, Ndapandula 398  
 Nam, Wonhong 1  
 Nanba, Hidetsugu 114  
 Nantajeewarawat, Ekawit 104  
 Neelameghan, A. 408  
 Nguyen, Quang Minh 412  
 Niu, Zhendong 61  
  
 Ohshima, Hiroaki 71  
 Osborn, Wendy 236  
 Oyama, Satoshi 124  
  
 Peterson, Erik 404  
 Polydoratou, Panayiota 154  
  
 Qin, Kai 61  
  
 Raghavan, K.S. 408  
 Razikin, Khasfariyati 51, 412  
 Rodrigues, Rodrigo 225  
  
 Sari, Syandra 276  
 Sato, Keizo 309  
 Sharma, Monica 363, 416  
 Shi, Hao 194  
 Singh, Diljit 285  
 Srihari, Rohini K. 404  
  
 Srihari, Sargur N. 134  
 Suleman, Hussein 31, 398, 406  
 Sun, Aixin 266, 412  
  
 T.N., Vikram 371  
 Tan, Yee Fan 313  
 Tanaka, Katsumi 71, 124  
 Tani, Seiichi 317  
 Teng, Yu-Ying 414  
 Thaiprayoon, Santipong 335  
 Theng, Yin Leng 412  
 Thet, Tun Thura 184  
 Tsai, Shu-Ting 400  
  
 Urs, Shalini R. 194, 355, 363, 371, 416  
  
 Waldstein, Ilya 379  
 Wang, Jenq-Haur 331  
 Wang, Kehong 351  
 Wang, Weichun 359  
 Wang, Yi 12  
 Williams, Raymond 402  
 Witten, Ian H. 236, 294  
  
 Xiang, Liang 410  
 Xing, Chun-Xiao 164, 347, 351  
  
 Yang, Naomi 93  
 Yang, Qing 410  
 Yeh, Jian-hua 93  
 Yi, Kwan 321  
 Yoshida, Taiga 124  
 Yoshioka, Suguru 317  
  
 Zhang, Ming 359  
 Zhang, Quanxin 61  
 Zhang, Xiaodan 203  
 Zhang, Yong 164  
 Zhou, Li 164  
 Zhou, Xiaohua 203  
 Zhu, Weizhong 379  
 Žumer, Maja 327