LNCS 2555

Ee-Peng Lim Schubert Foo
Chris Khoo Hsinchun Chen
Edward Fox Shalini Urs
Thanos Costantino (Eds.)

Digital Libraries: People, Knowledge, and Technology

5th International Conference on Asian Digital Libraries, ICADL 2002
Singapore, December 2002
Proceedings

Springer

Lecture Notes in Computer Science 2555

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Tokyo

Ee-Peng Lim Schubert Foo Chris Khoo
Hsinchun Chen Edward Fox Shalini Urs
Thanos Costantino (Eds.)

Digital Libraries: People, Knowledge, and Technology

5th International Conference on Asian Digital Libraries,
ICADL 2002
Singapore, December 11-14, 2002
Proceedings

Springer

Volume Editors

Ee- Peng Lim
Nanyang Technological University
E-mail: aseplim@ntu.edu.sg

Schubert Foo
Nanyang Technological University
E-mail: assfoo@ntu.edu.sg

Chris Khoo
Nanyang Technological University
E-mail: assgkhoo@ntu.edu.sg

Hsinchun Chen
University of Arizona
E-mail: hchen@eller.arizona.edu

Edward Fox
Virginia Tech
E-mail:fox@vt.edu

Shalini Urs
University of Mysore
E-mail: shalini@vidyanidhi.org.in

Thanos Costantino
IEI-CNR
E-mail: thanos@iei.pi.cnr.it

Cataloging-in-Publication Data applied for

A catalog record for this book is available from the Library of Congress.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data is available in the Internet at http://dnb.ddb.de

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

ISSN 0302-9743
ISBN 3-540-00261-8 Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Printed in Germany

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Stefan Sossna e. K.
Printed on acid-free paper SPIN 10871754 06/3142 5 4 3 2 1 0

Preface

The *International Conference on Asian Digital Libraries* (ICADL) is an annual international forum for participants to exchange research results, innovative ideas, and state-of-the-art developments in digital libraries. Built upon the successes of the first four ICADL conferences, the 5th ICADL Conference in Singapore was aimed at further strengthening the position of ICADL as a premier digital library conference that draws high quality papers and presentations from all around the world, while meeting the needs and interests of digital library communities in the Asia-Pacific region.

The theme of the conference, “Digital Libraries: People, Knowledge & Technology,” reflects the shared belief of the organizers that success in the development and implementation of digital libraries lies in the achievement of three key areas: the richness and depth of content to meet the needs of the communities they intend to serve; the technologies that are employed to build user-centered environments through organization, interaction, and provision of access to that content; and the human elements of management policies, maintenance, and vision necessary to keep pace with new content, new technologies, and changing user needs.

For the first time, ICADL was jointly held in conjunction with the 3rd International Conference on Web Information Systems Engineering (WISE) that focuses on Web-related database technologies. To promote interaction between ICADL and WISE participants, the two conferences shared a common opening ceremony and two joint keynote addresses. ICADL 2002 also included 4 tutorials, 1 keynote address, 6 invited talks, and 54 research paper presentations in 12 parallel sessions. These sessions, ranging from Digital Libraries for Community Building, Information Retrieval Techniques, and Human Computer Interfaces, to Digital Library Services, directly reflect and support the theme of the conference.

A record number of 110 regular and 60 short papers were submitted to the conference from researchers and practitioners from 29 countries. Another record registered by ICADL 2002 was the formation of a 72-member program committee composed of well-known digital library researchers from Pacific Asia, USA, and Europe. The large program committee was motivated by the need for ICADL to reach out to more people and to ensure good-quality paper submissions. Additional reviewers were also enlisted to assist in the review process. The reviewing process was managed by the ConfMan Conference Management Software, developed and supported by the University of Oslo, Brandenburg Technical University of Cottbus, and Darmstadt University of Technology. Thirty-four regular papers, 20 short papers, and 16 poster papers were accepted by the program committee.

On behalf of the Organizing and Program Committees of ICADL 2002, we thank all authors for their submissions and preparation of camera-ready copies of papers, and all delegates for their participation in the conference. We also acknowledge our sponsors, the Conference Support Committee, and all others for their strong support and generous help in making the conference a success. We hope that the conference will continue to grow from strength to strength as it travels to different host countries in Asia, and will continue to provide a forum for the stimulating exchange of ideas to enable us to build future digital libraries that surpass anything we currently can envision.

November 2002

Schubert Foo
Ee-Peng Lim
Hsinchun Chen
Edward Fox
Shalini Urs
Thanos Costantino

Conference Organization

The *Fifth International Conference on Asian Digital Libraries (ICADL 2002)* was organized by the Division of Information Studies, School of Communication and Information, Nanyang Technological University, Singapore, in collaboration with the National Library Board of Singapore, Nanyang Technological University Library, National University of Singapore Libraries, Library Association of Singapore, and the Library and Information Technology Association of Singapore.

Honorary General Chair

Cham Tao Soon (President, Nanyang Technological Univ., Singapore)

Advisors, Organizing Committee

Christopher Chia (Chief Executive, National Library Board, Singapore)
Choy Fatt Cheong (President, Library Association of Singapore)
Foo Kok Pheow (University Librarian, Nanyang Technological Univ. Library, Singapore)
Eddie Kuo (Dean, School of Communication & Information, Nanyang Technological Univ., Singapore)
Sylvia Yap (Deputy Director, National Univ. of Singapore Libraries, Singapore)

General Chair

Schubert Foo (Head, Division of Information Studies, School of Communication and Information, Nanyang Technological Univ., Singapore)

Program Co-chairs

Ee-Peng Lim (Nanyang Technological Univ., Singapore)
Hsinchun Chen (Univ. of Arizona, USA)
Edward Fox (Virginia Tech, USA)
Shalini Urs (Univ. of Mysore, India)
Thanos Costantino (Inst. di Elaborazione della Informazione, Italy)

Program Committee

Asia-Pacific

Abdus Sattar Chaudhry (Nanyang Technological Univ., Singapore)
Abhijth Lahiri (National Information System for Science and Technology, India)
Chao-chen Chen (National Taiwan Normal Univ., Taiwan)
Chin-Choo Wong (Nanyang Technological Univ., Singapore)
Ching-chun Hsieh (Academia Sinica, Taiwan)
Christopher Yang (Chinese Univ. of Hong Kong, Hong Kong)

Chu-Keong Lee (Nanyang Technological Univ., Singapore)
Diljit Singh (Univ. of Malaya, Malaysia)
Dion Hoe-Lian Goh (Nanyang Technological Univ., Singapore)
Hsueh-hua Chen (National Taiwan Univ., Taiwan)
Hwee-Hwa Pang (Laboratories for Information Technology, Singapore)
Ian Witten (Waikato Univ., New Zealand)
Jerome Yen (Chinese Univ. of Hong Kong, Hong Kong)
Jianzhong Li (Harbin Inst. of Technology, China)
Jieh Hsiang (National Chi-nan Univ., Taiwan)
Ji-Hoon Kang (Chungnam National Univ., Korea)
Jun Adachi (National Institute of Informatics, Japan)
K.S. Raghavan (Univ. of Madras, India)
Key-Sun Choi (KAIST, Korea)
Liddy Nevile (Univ. of Melbourne, Australia)
Li-zhu Zhou (Tsinghua Univ., China)
Man-Ho Lee (Chungnam National Univ., Korea)
Masatoshi Yoshikawa (NAIST, Japan)
Mun-Kew Leong (Laboratories for Information Technology, Singapore)
N. Balakrishnan (Indian Institue of Science, Bangalore, India)
N.V. Sathyanarayana (Informatics, India)
Noriko Kando (NII, Japan)
S. Sadagopan (Indian Institute of Information Technology, India)
San-Yih Hwang (National Sun Yet-Sen Univ., Taiwan)
Shigeo Sugimoto (Univ. of Tsukuba, Japan)
Soon J. Hyun (Information Communication Univ., Korea)
Steve Ching (Feng Chia Univ., Taiwan)
Suliman Hawamdeh (Nanyang Technological Univ., Singapore)
Sung Hyun Myaeng (Chungnam National Univ., Korea)
T.A.V. Murthy (INFLIBNET, India)
T.B. Rajashekar (Indian Institute of Science, India)
Vilas Wuwongse (AIT, Thailand)
Wai Lam (Chinese Univ. of Hong Kong, Hong Kong)
Wee-Keong Ng (Nanyang Technological Univ., Singapore)
Wen Gao (Chinese Academy of Sciences, China)
Yahiko Kambayashi (Kyoto Univ., Japan)
Yin-Leng Theng (Nanyang Technological Univ., Singapore)

USA

Carl Lagoze (Cornell Univ., USA)
Ching-chih Chen (Simmons College, USA)
Christine Borgman (UCLA, USA)
Edie Rasmussen (Univ. of Pittsburg, USA)
Gary Marchionini (Univ. of North Carolina, USA)
Howard Wactlar (CMU, USA)
Jim French (Univ. of Virginia, USA)
Jonathon Furner (UCLA, USA)
Judith Klavans (Columbia Univ., USA)
Marcia Lei Zeng (Kent State Univ., USA)

Richard K. Furuta (Texas A&M Univ., USA)
Robert Allen (Univ. of Maryland, USA)
Stuart Weibel (OCLC, USA)

Europe

Alan Smeaton (Dublin City Univ., Ireland)
Andreas Rauber (Vienna Univ. of Technology, Austria)
Ann Blandford (Univ. College London, UK)
Carol Peters (Italian National Research Council (IEI-CNR), Italy)
Donatella Castelli (Italian National Research Council (IEI-CNR), Italy)
Erich Neuhold (Darmstadt Univ. of Technology/Fraunhofer-Gesellschaft (FhG), Germany)
Gobinda Chowdhury (Univ. of Strathclyde, UK)
Harold Thimbleby (Univ. College London, UK)
Ingeborg Solvberg (Norwegian Univ. of Science and Technology (NTNU), Norway)
Jose Borbinha (National Library of Portugal, Portugal)
Keith van Rijsbergen (Univ. of Glasgow, UK)
Marc Nanard (LIRMM Montpellier, France)
Mike Papazoglou (Tilburg Univ., The Netherlands)
Norbert Fuhr (Univ. of Dortmund, Germany)
Thomas Baker (Fraunhofer-Gesellschaft (FhG), Germany)
Traugott Koch (Lund Univ. Libraries, Sweden)
Yannis Ioannidis (Univ. of Athens, Greece)

Reviewers

Aixin Sun
Andreas Meissner
David Woon Yew Kwong
Diego Klappenbach
George Buchanan
J.K. Vijayakumar
Klaus Maetzel
Kok-Leong Ong
Leonardo Candela
Liang-Tien Chia
Maria Bruna Baldacci
Myo Myo Naing
Na Jin Cheon
Robert Donald
Shiguang Shan
Shiyan Ou
Thomas Risse
Tiejunhuang
Trond Aalberg
Ulrich Thiel
Weiqiang Wang
Wensi Xi
Xiaoming Zhang
Xing, Chunxiao
Zehua Liu

Publication and Tutorial Committee

Christopher Khoo, Chair (Nanyang Technological Univ., Singapore)

Workshop Committee

Elisabeth Logan, Chair (Nanyang Technological Univ., Singapore)

Publicity and Sponsorship Committee

Abdus Chaudhry, Chair (Nanyang Technological Univ., Singapore)
Dion Goh (Nanyang Technological Univ., Singapore)
Wong-Yip Chin Choo (Nanyang Technological Univ. Library, Singapore)
Samantha Ang (Nanyang Technological Univ. Library, Singapore)
Cecilia Lee (National Univ. of Singapore Libraries)

Local Arrangements Committee

Suliman Hawamdeh, Chair (Nanyang Technological Univ., Singapore)
Shaheen Majid (Nanyang Technological Univ., Singapore)
Akbar Akim (Nanyang Technological Univ. Library, Singapore)
Tan Keat Fong (National Library Board, Singapore)
Tan Lay Tin (National Univ. of Singapore Libraries)

Sponsoring Institutions

The *Fifth International Conference on Asian Digital Libraries (ICADL 2002)* wishes to thank the following organizations for their support:

Thompson-ISI
iGroup Asia Pacific
Lee Foundation
National Library Board of Singapore
Infocomm Development Authority of Singapore
IBM
Elsevier Science
SIRSI Corporation
Adroit Innovations Ltd.
Blackwell's Book Services
John Wiley & Sons
Asia Library News-InfoMedia Asia Ltd.
Information and Knowledge Management Society
EBSCO Information Services
Thomson Learning – Gale
Kinokuniya Bookstores of Singapore Pte Ltd.
Internet Securities, Inc.

Table of Contents

Keynote and Invited Papers

Challenges in Building Digital Libraries for the 21st Century 1
Christine L. Borgman

Building Digital Libraries Made Easy: Toward Open Digital Libraries 14
Edward A. Fox, Hussein Suleman, Ming Luo

Dublin Core: Process and Principles 25
Shigeo Sugimoto, Thomas Baker, Stuart L. Weibel

From Digital Library to Digital Government: A Case Study in Crime Data Mapping and Mining 36
Hsinchun Chen

Progress on Educational Digital Libraries: Current Developments in the National Science Foundation's (NSF) National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program 53
Jane C. Prey, Lee L. Zia

Examples of Practical Digital Libraries: Collections Built Internationally Using Greenstone 67
Ian H. Witten

Data Mining Technologies for Digital Libraries and Web Information Systems 75
Ramakrishnan Srikant

Papers

Information Retrieval Techniques

Chinese Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis 76
Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang

A Linear Text Classification Algorithm Based on Category Relevance Factors 88
Zhi-Hong Deng, Shi-Wei Tang, Dong-Qing Yang, Ming Zhang, Xiao-Bin Wu, Meng Yang

A Hierarchical Framework for Multi-document Summarization of Dissertation Abstracts 99
Christopher S.G. Khoo, Shiyan Ou, Dion Hoe-Lian Goh

Generality of Texts 111
R.B. Allen, Yejun Wu

The Effectiveness of a Graph-Based Algorithm for Stemming 117
Michela Bacchin, Nicola Ferro, Massimo Melucci

Multimedia Digital Libraries

Searching Digital Music Libraries 129
David Bainbridge, Michael Dewsnip, Ian Witten

The NUS Digital Media Gallery – A Dynamic Architecture for Audio, Image, Clipart, and Video Repository Accessible via the Campus Learning Management System and the Digital Library 141
Phoon Moi Ying, Junius Soh Hock Heng

A Schema Language for MPEG-7 153
Ano Chotmanee, Vilas Wuwongse, Chutiporn Anutariya

Bitmap-Based Indexing for Multi-dimensional Multimedia XML Documents . . . 165
Jong P. Yoon, Sung H. Kim, Go U. Kim, Venu Chakilam

What People Do When They Look for Music: Implications for Design of a Music Digital Library 177
Sally Jo Cunningham

Distributing Relevance Feedback in Content Based Image Retrieval Systems . . . 179
Raghu Menon, Raj Acharya

Data Mining in Digital Libraries

Multistrategy Learning of Rules for Automated Classification of Cultural Heritage Material 182
G. Semeraro, F. Esposito, S. Ferilli, N. Fanizzi, T.M.A. Basile, N. Di Mauro

VideoCube: A Novel Tool for Video Mining and Classification 194
Jia-Yu Pan, Christos Faloutsos

Special Purpose Digital Libraries

Developing Tsinghua University Architecture Digital Library for Chinese Architecture Study and University Education 206
Chunxiao Xing, Lizhu Zhou, Zhiqiang Zhang, Chun Zeng, Xiaodan Zhou

Retrieving News Stories from a News Integration Archive 218
Arnulfo P. Azcarraga, Tat Seng Chua, Jonathan Tan

Digital Library Services

A Data Mining Approach to New Library Book Recommendations 229
San-Yih Hwang, Ee-Peng Lim

Grouping Web Pages about Persons and Organizations for Information Extraction 241
Shiren Ye, Tat-seng Chua, Jimin Liu, Jeremy R. Kei

Personalized Services for Digital Library 252
Chun Zeng, Xiaohui Zheng, Chunxiao Xing, Lizhu Zhou

Automatic References: Active Support for Scientists in Digital Libraries 254
Harald Krottmaier

Organizing and Maintaining Dynamic Digital Collections 256
M. Nikolaidou, D. Anastasiou, D. Anagnostopoulos, M. Hatzopoulos

Navigation, Organization, and Retrieval in Personal Digital Libraries of Email 258
Benjamin M. Gross

A Work Environment for a Digital Library of Historical Resources 260
Dion Goh, Lin Fu, Schubert Foo

Digital Libraries for Community Building

A Personalized Collaborative Digital Library Environment 262
M. Elena Renda, Umberto Straccia

Virtual Tutor: A System for Deploying Digital Libraries in Classrooms 275
Joemon M. Jose, Hywel Braddick, Innes Martin, Brian Robertson, Craig Walker, Greg MacPherson

Resource Annotation Framework in a Georeferenced and Geospatial Digital Library 287
Zehua Liu, Ee-Peng Lim, Dion Hoe-Lian Goh

Building Policy, Building Community: An Example from the US National Science, Technology, Engineering, and Mathematics Education Library (NSDL) 299
Ellen Hoffman, Edward A. Fox

Building a Digital Library from the Ground Up: An Examination of Emergent Information Resources in the Machine Learning Community 301
Sally Jo Cunningham

Subscription Clubs for E-journals: Indian Initiatives 303
I.R.N. Goudar, Poornima Narayana

Information Retrieval in Asian Languages

A Multilingual Multi-script Database of Indian Theses: Implementation of Unicode at Vidyanidhi 305
Shalini R. Urs, N.S. Harinarayana, Mallinath Kumbar

A Workbench for Acquiring Semantic Information and Constructing Dictionary for Compound Noun Analysis 315
Kyung-Soon Lee, Do-Wan Kim, Kyo Kageura, Key-Sun Choi

Building Parallel Corpora by Automatic Title Alignment 328
Christopher C. Yang, Kar Wing Li

Offline Isolated Handwritten Thai OCR Using Island-Based Projection with N-Gram Models and Hidden Markov Models 340
Thanaruk Theeramunkong, Chainat Wongtapan, Sukree Sinthupinyo

A Cache-Based Distributed Terabyte Text Retrieval System in CADAL 352
Jun Cheng, Wen Gao, Bin Liu, Tie-jun Huang, Ling Zhang

Building and Using Digital Libraries

Rural Digital Library: Connecting Rural Communities in Nepal 354
Binod Vaidya, Jagan Nath Shrestha

Collection Development for the Digital Age: The Case of Malaysia 366
Ahmad Bakeri, Abu Bakar, Kaba Abdoulaye

Digital Divide: How Can Digital Libraries Bridge the Gap? 379
Gobinda G. Chowdhury

Digital Libraries in Academia: Challenges and Changes 392
Anne Adams, Ann Blandford

Building Digital Libraries for Children: Reviewing Information Literacy of Students and Teachers 404
Marcia Mardis, Ellen Hoffman

Metadata Issues

The Use and Functionality of the Environmental Data Registry: An Evaluation of User Feedback 406
Panayiota Polydoratou, Michael Pendleton, David Nicholas

Adding Semantics to 3D Digital Libraries 419
Anshuman Razdan, Jeremy Rowe, Matthew Tocheri, Wilson Sweitzer

Algorithms and Protocols

INEXP: Information Exchange Protocol for Interoperability 421
Hongzhi Wang, Jianzhong Li, Zhenying He

Study on Data Placement and Access Path Selection in an FC-SAN Virtual Storage Environment 423
Chao Li, Chun-xiao Xing, Li-zhu Zhou

Building an OAI-Based Union Catalog for the National Digital Archives Program in Taiwan 425
Chao-chen Chen, Hsueh-hua Chen

Human Computer Interfaces

Intergenerational Partnerships in the Design of a Digital Library of Geography Examination Resources 427
Yin-Leng Theng, Dion Hoe-Lian Goh, Ee-Peng Lim, Zehua Liu, Natalie Lee-San Pang, Patricia Bao-Bao Wong, Lian-Heong Chua

Pie Charts for Visualizing Query Term Frequency in Search Results 440
Terry Anderson, Ali Hussam, Bill Plummer, Nathan Jacobs

Information Therapy in Digital Libraries 452
Yin-Leng Theng

Evaluation of Task Based Digital Work Environment 465
Narayanan Meyyappan, Schubert Foo

A Framework for Flexible Information Presentation in Digital Collections 467
Unmil P. Karadkar, Jin-Cheon Na, Richard Furuta

Electronic Journal of the University of Malaya (EJUM): An Attempt to Provide a Truly Electronic Environment 469
A.N. Zainab, N.N. Edzan, T.F. Ang

Digital Library Infrastructure

Patenting the Processes for Content-Based Retrieval in Digital Libraries 471
Hideyasu Sasaki, Yasushi Kiyoki

Secure Content Distribution for Digital Libraries 483
Mariemma I. Yagüe, Antonio Maña, Javier López, Ernesto Pimentel, José M. Troya

A Strategic Level for Scientific Digital Libraries 495
Ling Feng, Manfred Jeusfeld, Jeroen Hoppenbrouwers

Bridging the Gap between Information Resource Design and Enterprise Content Management 507
Sue Fowell

A Digital Content Management Model for Making Profits in Digital Content Sites 516
Hyeonjeong Mun, Sooho Ok, Yongtae Woo

Posters

Digital Library Initiatives and Services

The Idea of a Digital Library: Issues of Today 518
Swapan K. Dasgupta, Asoke P. Chattopadhyay

US-Korea Collaboration on Digital Libraries: An Overview and Generalization for Pacific Rim Collaboration 519
Edward A. Fox, Reagan W. Moore, Ronald L. Larsen, Sung Hyon Myaeng, Kim Sung-Hyuk

ETDs at HKU: A Spearhead for Digital Library Growth 520
David T. Palmer

MiMedicalLibrary: A Digital Health Library for Michigan 521
Harvey R. Brenneise

Reference Services in a Digital Library of Historical Artifacts 522
Roy Chan, Dion Goh, Schubert Foo

An Integrative User-Centered Purchase Request Service in the Age of Digital Library Development 523
Ching-Fan Wu, Hui-Chen Hsieh

Vitalising Library and Information Science Education: A Challenge in the Digital Information Environment 524
K. Padmini

Technology

Developing a Dialogue Library System 525
Ivan Kopecek, Miroslav Bartosek

WebClipper: A Personal/Community Link Library Builder Based on Web Link Management Technique 526
Young S. Kim, Jae H. Lim, Soon J. Hyun, Dongman Lee

Hiding a Logo Watermark in an Image for Its Copyright Protection 527
Jun Zhang, Feng Xiong, Nengchao Wang

Searching Video Segments through Transcript, Metadata, and SVG Objects 528
M. Karthikeyan, D. Uzagare, S. Krishnan, Meng Yang, Gary Marchionini

Similar Sub-trajectory Retrieval Based on k-Warping Distance Algorithm for Moving Objects in Video Databases 529
Choon-Bo Shim, Jae-Woo Chang

A Keyword Spotting System of Korean Document Images 530
Il-Seok Oh, Yoon-Sung Choi, Jin-Ho Yang, Soo-Hyung Kim

An Efficient Strategy for Adding Bulky Data into B+-Tree Indices in Information Retrieval Systems 531
Jin-Ho Kim, Ju-Young Kim, Sang-Wook Kim, Soo-Ho Ok, Hi-Young Roh

Author Index 533

Challenges in Building Digital Libraries for the 21st Century

Christine L. Borgman

Professor & Presidential Chair in Information Studies
Dept of Information Studies
235 GSE&IS Bldg, Box 951520
University of California, Los Angeles
Los Angeles, CA 90095-1520
cborgman@ucla.edu
http://is.gseis.ucla.edu/cborgman/

Abstract. After a decade of research and development, digital libraries are becoming operational systems and services. This paper summarizes some of the challenges required for that transition. Digital libraries as systems are converging with digital libraries as institutions, particularly as we consider the service aspects. They are enabling technologies for applications such as classroom instruction, information retrieval, and electronic commerce. Because usability depends heavily upon context, research on uses and users of digital libraries needs to be conducted in a wide array of environments. Interoperability and scaling continue to be major issues, but the problems are better understood. While technical work on interoperability and scaling continues, institutional collaboration is an emerging focus. Concerns for an information infrastructure to support digital libraries is moving toward the concept of "cyberinfrastructure," now that distributed networks are widely deployed and access is becoming ubiquitous. Appropriate evaluation methods and metrics are requirements for sustainable digital libraries that have received little attention until recently. We need to know what works and in what contexts. Evaluation has many aspects and can address a variety of goals, such as usability, maintainability, interoperability, scalability, and economic viability. Lastly, two areas that have received considerable discussion elsewhere are noted -- digital preservation and the role of information institutions such as libraries and archives.

1 Introduction

We now have about a decade's experience in the research and development of digital libraries, and that experience builds upon several decades of prior research on information storage and retrieval systems. Digital libraries are beginning to move from research to practice, and from prototypes to operational systems. It is time to turn projects into programs [15].

Building operational systems and services is much different than conducting research and development. As Dan Greenstein commented, "In computer science, if it works, it's not research." [15] Moving to operational systems and services will require addressing a host of challenges. These challenges are not necessarily new.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 1-13, 2002.

Rather, some appear simpler and some appear more complex than a decade ago. In the interim, the community has learned more about the nature of digital libraries, their uses, their users, and the technical and institutional requirements for their support. Here I revisit the evolving definitions of digital libraries; the relationship between uses, users, and usability; interoperability and scaling; information infrastructure; and evaluation; and touch upon known problems such as digital preservation and the role of libraries as institutions.

2 What Are Digital Libraries?

First, what are these entities that we propose to turn into operational systems and services? I examined the proliferating definitions of the term "digital library" several years ago in an article [4] and later extended the discussion in a book [5]. Initially, the computer science community was focused on digital libraries as new forms of information retrieval systems that were distributed and that usually contained media in multiple formats (text, numeric, audio, and visual) rather than text alone. The library community considered digital libraries to be a new form of information institution, with staff and long-term responsibility for maintaining digital collections. In the interim, these perspectives have converged toward a hybrid view of digital libraries that has both technical and institutional components. However, "digital library" remains a contested term, meaning different things to different groups. The two-part definition established in [7] continues to be useful:

1. Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense, they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, image, sound; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g., representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library.
2. Digital libraries are constructed—collected and organized—by [and for] a community of users, and their functional capabilities support the information needs and uses of that community. They are a component of communities in which individuals and groups interact with each other, using data, information, and knowledge resources and systems. In this sense they are an extension, enhancement, and integration of a variety of information institutions as physical places where resources are selected, collected, organized, preserved, and accessed in support of a user community. These information institutions include, among others, libraries, museums, archives, and schools, but digital libraries also extend and serve other community settings, including classrooms, offices, laboratories, homes, and public spaces.

Several aspects of this definition are important in considering the shift from digital library research to practice. One aspect is a broad conceptualization of library "collections." The notion of collection is problematic when libraries provide access to resources that they do not own. Another aspect is that digital libraries encompass the full information life cycle: capturing information at the time of creation, making it accessible, maintaining and preserving it in forms useful to the user community, and sometimes disposing of information. With physical collections, users discover and retrieve content of interest; their use of that material is independent of library systems and services. With digital collections, users may retrieve, manipulate, and contribute content. Thus users are dependent upon the functions and services provided by digital libraries; work practices may become more tightly coupled to system capabilities [6].

3 Uses, Users, and Usability

Building a sustainable digital library presumes that the systems and services meet the needs of some user community. This is not a simple presumption, however. We need a much better understanding of who the users of these systems are and what use they make of them. Digital libraries support specific activities in specific contexts – scholarship, classroom instruction, distance learning, digital asset management, virtual museums, and so on. Usability must be considered in the context of specific uses and users.

We are now building digital libraries for a generation of users that has grown up with MTV, mobile phones, computer games, email, and the World Wide Web. Today's entering college students learned to search for information on Google – not in card catalogs. They are more likely to learn biology by "dissecting" virtual frogs than by watching filmstrips. They write their term papers on computers connected (with or without wires) to the Internet, where they can look up references and read current journal articles without setting foot in a traditional library. Many continue to use physical libraries, but they expect their libraries to provide a rich mixture of physical and digital collections. We need to consider a new generation of users, with experiences and expectations different than those on whom the first generations of online catalogs, information retrieval systems and digital libraries were tested.

Although today's scholars learned their craft in a world of print publications, print indexes, and card catalogs, most are taking full advantage of new information technologies that have emerged over the course of their careers. They are retrieving, reading, and writing electronic publications, and are sharing their research results via distributed, networked information systems. In the future, researchers will have yet more facilities to share data, more means to utilize tools from distributed sources to analyze those data, and more opportunities to collaborate across national and disciplinary boundaries [2]. At the same time, print publication continues unabated, and will continue to complement electronic sources of information for the foreseeable future.

To move toward operational systems and services, more research is needed on digital library uses and users, and this work needs to be conducted in many different environments. Two short case studies offer examples of usability research that is contributing to the goal of sustainable digital libraries.

3.1 Undergraduate Students and Faculty in Geography: The ADEPT Project

Exploratory research on uses and users of digital libraries is best accomplished via large, long term, collaborative ventures. We found this opportunity in the Alexandria

Digital Earth ProtoType Project (ADEPT), a five-year effort (1999-2004)[1] involving scholars of geography, computer science, psychology, and information studies. Our concern in the Education and Evaluation component of ADEPT[2] is the efficacy of geographical digital libraries for undergraduate instruction. Despite the policy juggernaut to place computers in classrooms, the effectiveness of technology-based instruction to improve learning is far from proven. Digital libraries have great potential to enrich learning by providing access to new forms of content. They also can support independent, guided, or collaborative learning [9,28,29].

Geography is a fruitful area of study, because the discipline depends heavily on primary source data such as observations from satellites and sensor networks. Many geographers gather and analyze these dynamic data for their research, yet students typically learn about geography from static, processed forms of data such as textbooks, slides, maps, and displays on overhead projectors. If students can use and manipulate the primary source data available to scholars, they may learn to "think like scientists" and to develop a much richer understanding of geographic concepts.

We began the project by identifying the users and uses of ADEPT. The system will serve faculty members in their role as instructors to gather and organize geo-spatial resources for instruction and in their role as researchers to gather, manipulate, and display geo-spatial data. Teaching assistants will use ADEPT to assist faculty members in assembling lectures, to reinforce concepts in laboratory sessions, and to assist students in performing interactive course assignments[3]. Students will attend lectures and laboratory sessions that incorporate ADEPT resources and will use ADEPT modules to manipulate geo-spatial data and to learn scientific concepts. This project is among the first to go beyond studies of searching to look at multiple uses of digital libraries, including creating and using new information resources.

A central concern of our research is whether scientific learning is taking place. We are applying theories of mental model to study students' comprehension of the digital library and of the geography modules for course instruction [5]. Our team members in psychology are leading this part of the effort[4]. Another guiding principle is that of least effort on the part of our users [14]. We assume that students (and probably faculty members) will spend no more than 30 minutes learning to use the ADEPT technology. The technology should be a minimal barrier to learning the scientific concepts that are being conveyed.

The second central concern of our research is building a technology that instructors will want to use. No matter how useful and usable ADEPT may be from a technological perspective, it must offer sufficient advantages over present practices for university faculty members to choose to employ it in their teaching. Thus we are interviewing faculty about their motivations and requirements for implementing information technology in their teaching, and then conducting iterative assessments of

[1] Funded by the U.S. National Science Foundation, Digital Libraries Initiative, Phase II, grant no. IIS-9817432, Terence R. Smith, University of California, Santa Barbara, Principal Investigator.

[2] The current members of the ADEPT Education and Evaluation team are Christine Borgman, Anne Gilliland-Swetland, Gregory Leazer, Laura Smart, Rich Gazan, Kelli Millwood, and Jason Finley at UCLA; and Richard Mayer, Rachel Nilsson, and Tricia Mautone at UCSB.

[3] Results of a study of the role of teaching assistants in geography, by Rich Gazan et al, will be available in early 2003.

[4] Richard Mayer, Rachel Nilsson, and Tricia Mautone at UCSB.

prototypes in their classrooms. We are also studying their teaching styles, to determine what behaviors are common to multiple instructors of the same course and what styles are unique. We are observing classrooms to identify instructors' use of concepts and the relationships among those concepts. These observations provide a basis to determine requirements for metadata and for functionality of the ADEPT digital library.

A complementary thread of our research is studying the information-seeking behavior of geographers in support of their teaching and research. While the scholarly activities of university faculty have been studied extensively [21], surprisingly little research has been done on how faculty locate information resources for their teaching. We are currently analyzing these data and expect to report on them in early 2003.

3.2 High School Students in Biology and Physics: The CENS Project

We are pursuing research questions similar to those of ADEPT, but with different content and different populations, as part of the Center for Embedded Networked Systems (CENS), a new National Science Foundation Science and Technology Center based at UCLA (http://www.cens.ucla.edu). Our role in CENS[5] is to deploy scientific data in school classrooms, grades 7-12 (ages 12-18), studying course design and inquiry-based student learning. CENS consists of research teams at multiple universities, in multiple disciplines, who are embedding sensor networks to gather data for biology, physics, environmental sciences, seismology, and other applications.

Some aspects of the uses, users, and usability of information systems in CENS are less difficult to study than in ADEPT and some aspects are more difficult. Studying the uses is somewhat easier because course design follows educational standards established by the State of California. Thus the requirements for educational content are relatively well understood. However, inquiry-based learning is a new approach that is far more difficult to implement than textbook-based instruction. Students are given the opportunity to construct and carry out experiments, and these experiments may be longitudinal, extending far beyond a single class period. Teachers must design an environment in which students can explore and in which answers may be ambiguous. The challenge here is providing primary source data to support inquiry learning.

The aspect of CENS that will be even more difficult than in ADEPT is that we must support streaming data, rather than data that has been processed and organized into manageable packets such as documents. Digital library technology should help us to manage these data and to provide content and services to this diverse scientific community. An important sub-project of CENS is to assess how researchers will gather and use these data, what standards for data and metadata are required, and how these data can be organized for use by CENS scholars and by high school students[6]. The use and re-use of scientific data for multiple audiences that have a diverse range of domain knowledge (e.g., scholars and high school students) is one of the great

[5] The education and evaluation team of CENS initially consists of Christine Borgman, UCLA Information Studies; William Sandoval, UCLA Education; Kathy Griffis, biology teacher at Buckley School; and Joe Wise, physics teacher at New Roads School.

[6] The data management research project of CENS is led by Kalpana Shankar, UCLA Information Studies.

challenges for digital library research. Data management decisions made early in the project will determine what questions can be asked, how the content can be searched, how it can be displayed, and what types of longitudinal data analysis will be possible.

4 Interoperability and Scaling

By the mid-1990s, interoperability and scaling were identified as key challenges for achieving sustainable digital libraries [19]. Clifford Lynch and Hector Garcia-Molina recently revisited the issues raised in that report, in a talk to the principal investigators of currently funded digital library projects [20]. They concluded that, for the most part, the report identified the right set of issues, although not as much progress has been made in the intervening 7 years as predicted. Design has been less user-centered than expected, and sustainability is farther from being achieved than hoped. While they did foresee the problems of digital preservation, they did not anticipate the rapid commercialization of digital libraries nor the changes in U.S. copyright law that would create new barriers to interoperability. Metadata was thought to be the key to interoperability, but it has turned out not to be a magic bullet. The community's assessments assumed high quality data – clean and honest – as is generally the case with library resources. They did not anticipate the large amount of self-published content, mis-represented data, spam, and deliberately incorrect metadata (such as the trashed music files distributed online to complicate music retrieval) that now exists. Scaling is at least as large a challenge as anticipated, with the growth of large video collections such as those of the Survivors of the Shoah Visual History Foundation (http://www.vhf.org).

The theme of interoperability is again high on the digital library agenda. Arms et al. [1] revisited the range of issues in interoperability and it was the topic of the Digital Library Federation Forum in May, 2002 [12]. Besser [3] and Greenstein [15] address needs for modular architectures and interoperability protocols. Progress is being made in these areas, such as METS (metadata encoding and transmission standard) (http://www.loc.gov/standards/mets/), OAIS (Open Archival Information System), and OAI (Open Archives Initiative) (http://www.openarchives.org) [18]. Besser [3] explores the various types of metadata that will contribute to interoperability.

What is new since the mid-1990s is an emphasis on interoperability as institutional cooperation. Technical connections between systems are much easier to achieve if they are based upon agreements that organizations will work together for common goals. Libraries are now working toward consensus on how-to guides, best practices, and benchmarks for data, metadata and digital library services. These activities are essential to move digital library services fully into the library's mainstream [10,15]. More technical efforts such as the OAI will enable users to search across multiple digital libraries provided by one organization (e.g., their institution's library) or multiple organizations.

5 Infrastructure for Digital Libraries

Delivering digital library services requires an infrastructure at the local (e.g., university), national, and international level. Visions for future information infrastructures -- distributed computing, grids, collaboratories, e-science – suggest a rich environment of content, computing, services, tools, and institutional structures. Key goals of the U.S.-based Cyberinfrastructure proposal include educating the next generation of scientists with the best technologies and tools, to enable broader participation in national and international collaborations [2].

How do we get from here to there, and where do digital libraries fit in? First, some definitions of these fuzzy terms. The Cyberinfrastructure report defines IT-based infrastructure as "a set of functions, capabilities, and/or services that make it easier, quicker, and less expensive to develop, provision, and operate a relatively broad range of applications. This can include facilities, software, tools, documentation, and associated human support organizations." [2] Digital libraries are an essential component of the content management requirements for IT-based infrastructure [2].

Greenstein [15] offers the example of the University of California's (UC) infrastructure for digital libraries, whose goal is the construction of a persistent research collection available to the entire UC community irrespective of location. This infrastructure includes a union bibliographic catalog, a shared approach to bibliographic cataloging, a "buying club" for commercial e-content, guidelines and tools for content creators, shared print repositories, digital archival repositories, and an array of end-user services such as document delivery and subject portals. The UC expects this infrastructure to lower the cost of high-quality, locally tailored, online library service environments.

Infrastructure can be defined more broadly as a social and technical construct. The eight dimensions identified by Star and Ruhleder [26] are still a useful framework: An infrastructure is *embedded* in other structures, social arrangements, and technologies. It is *transparent*, in that it invisibly supports tasks. Its *reach or scope* may be spatial or temporal, in that it reaches beyond a single event or a single site of practice. Infrastructure is *learned as part of membership* of an organization or group. It is linked with *conventions of practice* of day-to-day work. Infrastructure is the *embodiment of standards*, so that other tools and infrastructures can interconnect in a standardized way. It builds upon an *installed base*, inheriting both strengths and limitations from that base. And infrastructure becomes *visible upon breakdown*, in that we are most aware of it when it fails to work—when the server is down, the electrical power grid fails, or the highway bridge collapses.

The advantage for digital libraries of the broader definition of infrastructure is that the social context is acknowledged. Digital libraries are not an end in themselves; rather they are enabling technologies for other applications such as delivering intellectual content to students and teachers (as in the ADEPT and CENS projects described earlier) and electronic publishing. As universities develop advanced information infrastructures, digital libraries may facilitate profound changes in the way teaching, learning, and scholarship are conducted. Digital libraries can support learning that occurs in other than "same time / same place" instruction because they support asynchronous interaction (available at anytime) over distributed networks (accessible from any place with a network connection).

Rather than students having to visit a library or laboratory in person to use instructional materials, often competing for one or a few copies (or a limited number of laboratory work stations), one digital document can be accessible to multiple students at multiple places at all times. Documents that are independent in physical form can be interdependent in digital form, hyper-linked. Students and instructors can follow paths between documents linked by citations, common terms, formats, or other relationships. Links can be generated automatically or created manually by instructors and students. The scope of linking is not limited to materials gathered for one course. Links can be followed from one digital library to another, following paths to materials in many countries, cultures, and languages. Students can incorporate still and moving images, sounds, animated models, and other digital resources into their work. Far richer and more complex products can be produced, and they can be tailored to the subject matter of the product. Chemistry projects can include animated models of chemical bonding and dance projects can include films of dancers and animated choreography, for example. One of the greatest values of digital libraries may be in providing access to primary source content. The same "real" data used by scholars in their research can be made available to students – a primary goal of both the ADEPT and CENS projects.

6 Evaluation of Digital Library Systems and Services

If digital libraries are to become sustainable systems and services, we must continually evaluate their quality with respect to goals such as usability, maintainability, interoperability, scalability, and economic viability. Evaluation studies also can provide strategic guidance for the design and deployment of future systems, and assist in determining whether digital libraries address the appropriate social, cultural, and economic problems. Consistent evaluation methods will enable comparison between systems and services.

Despite the advances in digital library technology, we have insufficient understanding of their utility for most applications, and we lack appropriate evaluation methods, metrics, and testbeds for determining their effectiveness relative to various benchmarks. A recent European Union - U.S. workshop (which included Asian participation) addressed the need for evaluation methods and metrics for digital libraries (http://www.sztaki.hu/conferences/deval/presentations.html).

Evaluation is a general term that includes various aspects of performance measurement and assessment. Activities can include laboratory experiments; regional, national, and international surveys or quasi-experiments; time-series analyses; online monitoring of user-system interactions; and other forms of data collection. Evaluation has a long history in fields such as education, communication, health, and criminal justice. The effectiveness of interventions such as new teaching methods, management practices, and policy can be assessed [8,24]. In computer science, systems are benchmarked for various aspects of performance. Quantitative measures are typically specific to applications, such as recall and precision measures in information retrieval. In human-computer interaction, measures include time to learn, error rates, efficiency, memorability, and satisfaction [22,25].

Evaluation methods should meet accepted norms for scientific rigor in the domain of study. In the social sciences, methods should be valid (be a "true" measurement of the quality or concept under study) and reliable (the same measure should achieve the same result at multiple times). Kirk and Miller [17] offer succinct definitions of these concepts:

- *Reliability:* the extent to which the same observational procedure in the same context yields the same information.
- *Validity:* The quality of fit between an observation and the basis on which it is made.

At least four types of evaluation are relevant to digital libraries:

1. *Formative evaluation* begins at the initial stages of a development project to establish baselines on current operations, set goals, and determine desired outcomes. Such evaluation is usually driven by context and project-specific goals.
2. *Summative evaluation* takes place at the end of a project to determine if the intended goals were met. Goals and outcomes must be compared to initial states, so formative evaluation generally precedes summative evaluation.
3. *Iterative evaluation* takes place during a project, such as during the design and development of a digital library. Interim stages of design are assessed in comparison to design goals and desired outcomes, and the results inform the next stages of design. Iterative approaches encourage designers to set measurable goals at the beginning of a project and provide opportunities to re-assess goals throughout the development process.
4. *Comparative evaluation* requires standardized measures that can be compared across systems. Communities can identify and validate measures. If such measures are implemented in a consistent manner, they enable comparisons between systems. Testbeds are another way to compare measures and to compare performance of different functions and algorithms.

7 Further Challenges

The challenges for sustainable digital libraries identified thus far in this paper are the need to understand uses, users and usability; interoperability and scaling; infrastructure; and evaluation. These are daunting tasks in themselves. Other challenges deserve mention but are discussed extensively elsewhere. Two of the most salient are (1) digital preservation, and (2) the role of information institutions such as libraries and archives.

7.1 Digital Preservation

Preservation is a growing problem for digital libraries. Given the rate of advances in information technology, maintaining content in a continuously viable form is a major challenge. Most paper documents can be set on a shelf and remain readable for centuries, under proper storage conditions. Magnetic media (computer disks; audio, video, and data tapes; etc.) must be copied every few years to maintain the readability of content, and must be stored properly to ensure long-term readability [16]. Even if the medium remains viable, finding devices to read older formats is problematic. Already it is difficult to locate operational devices to read media that were widely distributed only a few years ago, such as 5.25” floppy disks or 33-1/3 rpm phonograph records. Devices to read 8” floppy disks, 78 rpm records, Betamax videotapes, and reel-to-reel film are even harder to find. Drives for 3.5” disks already have ceased to be a standard feature of new computers, thus reading these disks will soon be difficult.

Even if the media are readable, finding hardware with the necessary operating systems and application software to read older files can be impractical. Unless files are transferred to the subsequent generation of hardware and software quickly, it is unlikely they ever will be read again. All of the proposed data preservation strategies require active efforts to maintain the data in a readable form, rather than the passive strategies of putting a book on a shelf or a microfilm in a storage vault. Thus when universities create digital libraries, they commit themselves to recurring expenses of maintaining electronic content. Digital preservation is a flourishing research area and one of great import for the use of digital libraries in higher education [11,16,27].

7.2 Role of Information Institutions

Much of the investment (labor and capital) for building digital library systems and services will be the responsibility of libraries and archives. Libraries are institutions that select, collect, organize, conserve, preserve, and provide access to information. Archives perform many of the same functions, but tend to focus on “evidence” rather than on “information” and often have legal requirements for selection and retention of documents [13].

Paradoxically, the massive efforts that libraries and archives devote to digital libraries and information infrastructure are often invisible to their users. People who claim that they never go to the library anymore because everything they need is online are missing the fact that the library has come to them. The invisibility is partly due to the successes of the institution. Good library design means that people can find what they need, when they need it, in a form they want it. Good design is less obvious than bad design, and thus libraries risk being victims of their own success.

Another component is the invisible content and costs of libraries. Many users are simply unaware of the expense of acquiring and managing information resources or the amount of value added by libraries and librarians. Considerable professional time and vast amounts of paraprofessional and clerical time are devoted to the processes of selecting, collecting, organizing, preserving, and conserving materials so that they are available for access. The selection process requires a continuing dialog with the user community to determine current needs, continuous scanning of available resources, and judicious application of financial resources. Once selected, the items are collected, whether in physical form or by acquiring access rights. This process, which requires negotiation with publishers and others who hold the rights to desired items, sometimes takes months or years, depending on the materials and the rights. As new items are acquired, metadata are created to describe their form, content, and relationship to other items in the collection. Once in the collection, resources must be preserved and conserved to ensure continuous availability over time. The invisibility of information work was identified long ago [23], but the implications of this invisibility are only now becoming widely apparent.

The role of information institutions in providing sustainable digital library systems and services are addressed in [5][7]. Research questions that address these issues are explored further in [6].

8 Summary and Conclusions

Digital library research has matured to a phase where systems and services are becoming operational. The transition from research to practice brings many new challenges, some of which we are only now beginning to understand. Indeed, many of the challenges in deploying sustainable digital libraries will themselves require extensive research. Digital libraries are much larger in scale and function than the information retrieval systems that preceded them. They support not only searching, but also the creation, use, and archival storage of information in various formats (text, numeric, audio, visual, and combinations thereof). Digital libraries as systems are converging with digital libraries as institutions, particularly as we consider the service aspects.

Digital libraries are not an end in themselves; rather, they are enabling technologies. They may be a means to provide primary source data for scholarship and instruction, to publish documents in electronic form, or to support electronic commerce, for example. Because usability depends heavily upon context, research on uses and users of digital libraries needs to be conducted in a wide array of applications. We need a better understanding of what is common across contexts and what is distinct, if we are to achieve sustainable digital libraries.

Interoperability and scaling, another set of topics that emerged in the early days of digital library research, are now better understood. Less progress has been made since the mid-1990s than expected, partly due to unanticipated barriers to interoperability such as new copyright laws, and partly because the problem appears even more complex than was known at the time. While technical work on interoperability and scaling continues, institutional collaboration is an emerging focus.

Infrastructure is another challenge that was recognized early on. In this area, focus has shifted from computational power to a broader conceptualization of the requirements for computing and communications. Distributed, networked information systems are now widely deployed and access is becoming ubiquitous, at least for research purposes. The modern research university can now consider their "cyberinfrastructure" requirements for teaching and research. Entire nations, such as the U.S., are considering their cyberinfrastructure requirements.

Appropriate evaluation methods and metrics are a requirement for sustainable digital libraries that have received little attention until recently. We need to know what works and in what contexts. Evaluation has many aspects and can address a variety of goals, such as usability, maintainability, interoperability, scalability, and economic viability. International collaboration on evaluation methods and metrics for digital libraries are under way.

Lastly, two areas that have received considerable attention were noted. These are digital preservation and the role of information institutions such as libraries and archives. We will not achieve sustainability without substantial progress on these fronts.

[7] See especially Chapter 7, "Wither, or whither, libraries?"

While much work remains, the good news is that the challenges are being recognized and that research is being pursued in each of the areas identified. Further, most of the research toward sustainable digital libraries is based on the premise that digital libraries cross international and cultural boundaries. Research teams must collaborate without regard to artificial barriers if we are to achieve the goal of sustainable digital library systems and services.

References

1. Arms, W.Y. et al. (2002). A spectrum of interoperability. *D-Lib Magazine*, 8(2). http://www.dlib.org/dlib/january02/01arms.html
2. Atkins, D.E., et al. (2002). Revolutionizing science and engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Draft 1.0, April 19, 2002. worktools.si.umich.edu/workspaces/datkins/001.nsf
3. Besser, H. The next stage; Moving from isolated digital collections to interoperable digital libraries. *First Monday: Peer-reviewed journal on the Internet.* http://www.firstmonday.dk/issues/issue7_6/besser/index.html
4. Borgman, C.L. (1999). What are digital libraries? Competing visions. *Information Processing & Management,* 38(3), 227-243. In G. Marchionini & E. Fox (eds.), Special Issue: Progress Toward Digital Libraries.
5. Borgman, C. L. (2000). From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World. Cambridge, MA: The MIT Press.
6. Borgman, C. L. (in press). The invisible library: paradox of the global information infrastructure. *Library Trends*, Special Issue on Research Questions for the Field.
7. Borgman, C.L.; Bates, M.J.; Cloonan, M.V.; Efthimiadis, E.N.; Gilliland-Swetland, A.; Kafai, Y.; Leazer, G.L.; Maddox, A. (1996). *Social Aspects Of Digital Libraries*. Final Report to the National Science Foundation; Computer, Information Science, and Engineering Directorate; Division of Information, Robotics, and Intelligent Systems; Information Technology and Organizations Program. Award number 95-28808. Available at: http://is.gseis.ucla.edu/DL/
8. Burstein, L.; & Freeman, H.E. (1985). Perspectives on data collection in evaluations. In L. Burstein, H. E. Freeman, & P.H. Rossi (eds.). *Collecting Evaluation Data*. Beverly Hills: Sage. Pp. 15-34.
9. Criddle, S.; Dempsey, L. & Heseltine, R. (1999). *Information landscapes for a learning society. Networking and the future of libraries, 3.* Bath, UK: UKOLN, the UK Office for Library and Information Networking and London: Library Association.
10. Flecker, D. (2000). Harvard's Library Digital Initiative: Building a First Generation Digital Library Infrastructure. D-Lib Magazine, 6(11), http://www.dlib.org/dlib/november00/flecker/11flecker.html
11. Foster, I. and Kesselman, C. 1999. The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann.
12. George, J. (2002). Digital libraries seek interoperability. CLIR Issues, 28, 3-5. Washington, DC: Council on Library and Information Resources. http://www.clir.org
13. Gilliland-Swetland, A. J. (2000). Enduring Paradigms, New Opportunities: The Value of the Archival Perspective in the Digital Environment. Washington, D.C.: Council on Library and Information Resources.
14. Gilliland-Swetland, A.J., Leazer, G.H. (2001). Iscapes: Digital Library Environments to Promote Scientific Thinking by Undergraduates in Geography. In Fox, E.A.; & Borgman, C.L. (eds*.). Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries*. June 24-28, 2001, Roanoke, VA. New York: ACM. Pp. 120-121.

15. Greenstein, D. (2002). Next Generation Digital Libraries? Keynote address. Marchionini, G.; & Hersh, W. (eds.). (2002). Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries. July 14-18, 2002, Portland, OR. New York: ACM.
16. Hedstrom, M. (1998). Digital preservation: A time bomb for digital libraries. *Computers and the Humanities, 31*, 189-202.
17. Kirk, J.; Miller, M.L. (1986). *Reliability and Validity in Qualitative Research*. Newbury Park: Sage.
18. Lagoze, C. & Van de Sompel, H. (2001). The Open Archives Initiative: Building a low-barrier interoperability framework. In Fox, E.A.; & Borgman, C.L. (eds.). *Proceedings of the Joint Conference on Digital Libraries, June 20-24, 2001, Roanoke, VA*. New York: ACM Press, pp 54-62.
19. Lynch, C.; Garcia-Molina, H. (1995). *Interoperability, scaling, and the digital libraries research agenda.* http://www.hpcc.gov/reports/reports-nco/iita-dlw/main.html
20. Lynch, C.; Garcia-Molina, H. (2002). *Retrospective: The 1995 Interoperability, Scaling, and the Digital Libraries Research Agenda Report.* DLI2/IMLS/NSDL Principal Investigators Meeting, Portland, OR, July 17-18, 2002. http://www.dli2.nsf.gov/dli2pi2002/program.html
21. Meadows, A.J. (1998). *Communicating Research*. San Diego: Academic Press.
22. Nielsen, J. (1993). *Usability Engineering.* Boston: Academic Press.
23. Paisley, W. J. (1980). Information and work. In B. Dervin & M. J. Voigt (Eds), *Progress in the Communication Sciences* (Vol. 2, pp. 114-165). Norwood, NJ: Ablex.
24. Rogers, E. M. (1986). Communication technology: The New Media in Society. New York: Free Press.
25. Shneiderman, B. (1998). Designing the User Interface: Strategies for Effective Human-Computer Interaction, 3rd ed. Reading, MA: Addison-Wesley.
26. Star, S.L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research,* Special issue on Organizational Transformation, J. Yates & J. Van Maanen (eds.). 7(1), 111-134.
27. The State of Digital Preservation: An International Perspective. (2002). Conference Proceedings. Documentation Abstracts, Inc. Institutes For Information Science, Washington, D.C. April 24-25, 2002. Washington, DC: Council on Library and Information Resources. Pub. No. 107.
28. Twidale, M.B.; & Nichols, D.M. (1998a). *A survey of applications of CSCW for digital libraries.* Technical report CSEG/4/98, Computing Department, Lancaster University, U.K. Http://www.comp.lancs.ac.uk/computing/research/cseg/projects/ariadne/docs/
29. Twidale, M.B.; & Nichols, D.M. (1998b). Computer Supported Cooperative Work in Information Search and Retrieval. In Williams, M.E. (ed.), *Annual Review of Information Science and Technology*, *33*, pp. 259-319.

Building Digital Libraries Made Easy: Toward Open Digital Libraries

Edward A. Fox, Hussein Suleman, and Ming Luo

Digital Library Research Laboratory, Virginia Tech
Blacksburg, VA 24060 USA
{fox, hussein, lming}@vt.edu
http://www.dlib.vt.edu

Abstract. Digital libraries (DLs) promote a sharing culture among those who contribute and those who use resources. This same approach works when building Open Digital Libraries (ODLs). Leveraging the intellectual and practical investment made in the Open Archives Initiative through an eXtended Protocol for Metadata Harvesting (XPMH), one can build lightweight protocols to tie together key components that together make up the core of a DL. DL developers in various settings have learned how to apply this framework in a few hours. The ODL approach has been effective with the Computer Science Teaching Center (www.cstc.org), the Networked Digital Library of Theses and Dissertations (www.ndltd.org), and AmericanSouth.org. Hence, to support our Computing and Information Technology Interactive Digital Educational Library (www.citidel.org) and to provide a generic capability for other parts of the US National Science, technology, engineering, and mathematics education Digital Library (www.nsdl.org), we are developing a "DL-in-a-box" toolkit. When lightweight protocols, pools of components, and open standard reference models are combined carefully, as suggested in the OCKHAM discussions, both the DL user and developer communities can benefit from the principle of sharing.

1 Introduction

"Digital libraries" has many definitions and can be viewed from many perspectives [4, 13, 14]. Here we consider it as referring, in different contexts, to two related modern constructs. The first has been called a digital library service system (DLSS, see [6]); it typically is a large, monolithic software package. The second, a type of institution, which is the target of the 5S framework [11], integrates at least: community, services, and content, supported by a DL system. In this paper we focus on the former case, though we are fully aware that the second approach is essential if the DL field to become a science [17].

Our focus is on supporting the large number of people – in libraries, documentation centers, computing centers, and research centers – who deal with the first construct on the way to satisfying requirements implicit in the second type of construct. They face serious problems, so a thoughtful approach is essential.

1.1 Problem

However, instead of building upon the work of others, most DL developers continue to "reinvent the wheel". Why? Here are some of the top reasons given:

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 14-24, 2002.

1. The library budget won't allow purchase of a commercial DL system.
2. Unless the development effort is local, there won't be any control.
3. DLs are extensions of DBMSs, so they are simple applications to develop.
4. Since DLs operate on the Web, one must adopt the newest W3C proposal.
5. Since technology moves so quickly, it is essential to follow the latest fad.
6. CS students always develop from scratch.
7. This team knows it can do it better.
8. This system must have more capabilities than any other system.
9. This DL has to be more flexible and extensible.
10. This is *the* right system architecture – at last!

Note: Lest we be accused of falling prey to the last myth above, it bears stating that our goal is not to develop the best architecture, but rather one that really is simple and easy to use.

1.2 Approach

Simplicity is the driving principle in our approach. This is reflected in our involvement in the OCKHAM initiative [18, 20, 21] as is discussed further in Section 5 below. It is exemplified by our building upon the work of the Open Archives Initiative [24, 28, 37], with its emphasis on low barriers to entry and support of interoperability.

Interoperability is a key goal in the field of DLs [29]. It has led to many investigations of DL architecture. Thus, at Stanford, the InfoBus is the mechanism (building on CORBA) that allows modules to function cooperatively [1]. On the other hand, at the University of Michigan, an agent approach was employed [2].

Since many DLs are built as distributed systems, the parts of such DLs must be able to communicate with each other. Further, since many DLs are in reality federations of independent DLs, these separate systems must speak a common protocol. Clearly we see that a simple protocol must be an essential element of our approach.

In Section 2 we explain that approach: building Open Digital Libraries. To make the idea concrete, we consider in Section 3 its application in the National Science Digital Library (NSDL). Section 4 gives additional examples of ODL's adoption and use in other applications. Then, Section 5 explains how ODL fits into the OCKHAM activities, while Section 6 concludes this paper.

2 Open Digital Libraries

In [38] we sketched the key ideas of the Open Digital Libraries (ODL) approach, and invited the DL community to comment as well as work with us. We set up a web site to provide current information about the freely available software we have been developing, along with related documentation and publications [39]. The following subsections summarize the key ideas.

2.1 Definition

The ODL approach calls for lightweight protocols that allow DL development to proceed simply by interconnecting components. Because of the success of the Open Archives Initiative [28], we build upon the OAI Protocol for Metadata Harvesting (PMH, see [41]). In this regard, we adopt a 2 step approach [35].

First, we developed a new protocol that is an extended form of PMH: XOAI-PMH. Since this was undertaken when PMH was at version 1, we were able to argue for these extensions, and some were incorporated in PMH version 2. The other extensions aimed to support general component-component communication inside a DL. In particular, these allow:

- Response-level containers
- Submission (using PutRecord)
- Ignoring the requirement to support DC (inside the DL)

Second, we developed specialized versions of XOAI-PMH for particular types of components. Examples include:

- Annotate (with PutRecord to add annotations for items whose ID is supplied using the set parameter)
- Browse (with the set parameter encoding the categories and sort order)
- Rate (with a metadata record encapsulating numerical rating and item ID)
- Search (with the keyword list, query language, and bounds for range of returned results all encoded in the set parameter)

Building on this 2 step approach of protocol extension, we then were able to develop components that satisfied these protocols, and thus allowed key DL functionality to emerge, as is explained in the next subsection.

2.2 Components

From our perspective [14], DLs can be thought of as powerful, high-end information systems that integrate a variety of multimedia, database, information retrieval, and human-computer interface technologies. They encompass creation, discovery, retrieval, and use of information. They support electronic publishing and content management [22]. Thus, a broad range of basic components are needed, and it is essential that they can be composed so that larger and larger systems can be developed. This is possible since an ODL component can be either an OAI data provider, and OAI service provider, or both.

Figures 1-3 illustrate both some of the components developed, and their composition to build a variety of digital libraries. In Figure 1 we see that a small group of individuals, each with a set of suitable XML files, can easily make these available as an OAI data provider. In addition, they can become an OAI service provider, supporting both searching and browsing. All together, this can be thought of as a basic DL.

Figure 2 illustrates a more complex DL. There is one new type of output supported, by a “what’s new” service provider. And there are 3 more types of input. One supports harvesting from open archives. The second allows submission of content, such as by authors or data entry personnel. The third, developed by our partners at NCSA, turns a relational database management system into an OAI data provider, which can be filtered first to ensure that only selected information is passed on.

Finally, Figure 3 illustrates fairly rich services. Composition also is illustrated, such as where IRDB-2 supports searching of annotations. Further, both the Recommend and Rate components can be accessed by a single service provider / interface.

For real-life DLs, however, even more complex systems may be needed. Sections 3 and 4 illustrate this point by way of exploring a variety of DL applications.

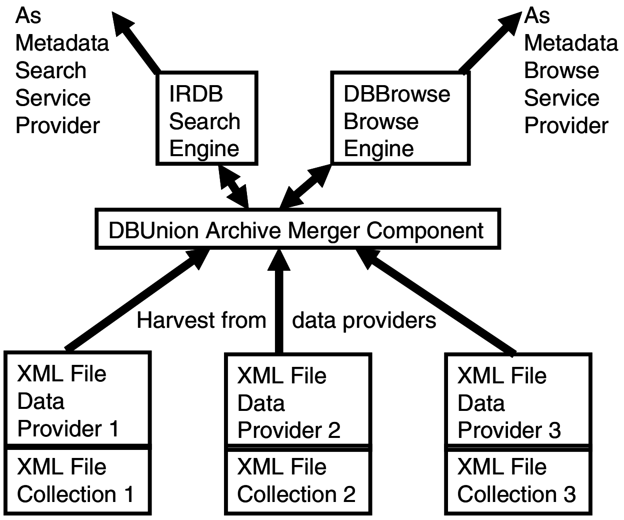


Fig. 1. Simple DL built from 4 basic types of components.

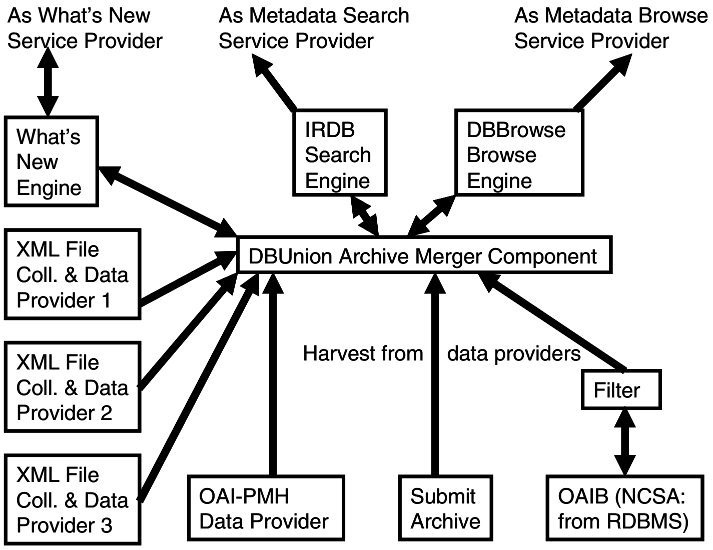

As What's New Service Provider
As Metadata Search Service Provider
As Metadata Browse Service Provider
What's New Engine
IRDB Search Engine
DBBrowse Browse Engine
XML File Coll. & Data Provider 1
DBUnion Archive Merger Component
XML File Coll. & Data Provider 2
Harvest from
data providers
Filter
XML File Coll. & Data Provider 3
OAI-PMH Data Provider
Submit Archive
OAIB (NCSA: from RDBMS)

Fig. 2. Intermediate DL built using 9 types of components.

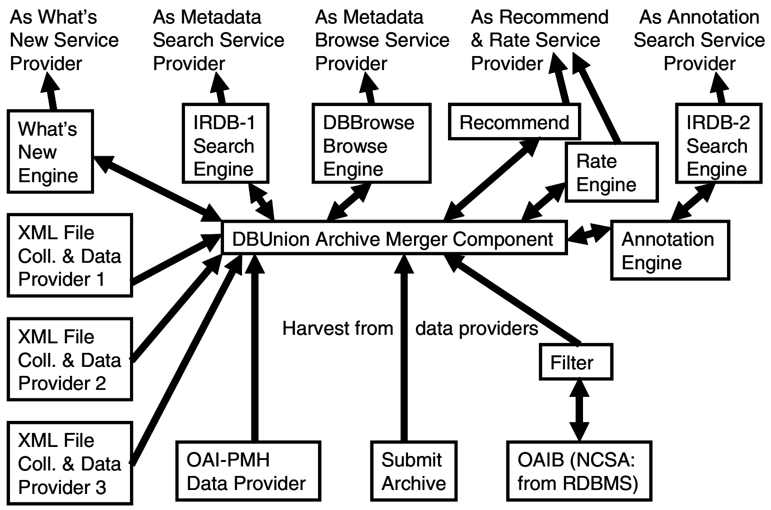

As What's New Service Provider
As Metadata Search Service Provider
As Metadata Browse Service Provider
As Recommend & Rate Service Provider
As Annotation Search Service Provider
What's New Engine
IRDB-1 Search Engine
DBBrowse Browse Engine
Recommend
Rate Engine
IRDB-2 Search Engine
XML File Coll. & Data Provider 1
DBUnion Archive Merger Component
Annotation Engine
XML File Coll. & Data Provider 2
Harvest from data providers
Filter
XML File Coll. & Data Provider 3
OAI-PMH Data Provider
Submit Archive
OAIB (NCSA: from RDBMS)

Fig. 3. More complex DL built using 12 types of components.

3 NSDL Applications

One of the largest DL activities currently underway is the National STEM education Digital Library – NSDL for short [31]. Sometimes, for simplicity, "STEM", which stands for Science, Technology, Engineering and Mathematics (replacing the old form, SMET), is expressed as "Science".

NSDL has 4 tracks. One deals with the Core Integration efforts. A second involves support for key Collections. The third focuses on Services. The fourth, and smallest, involves specialized Research, including evaluation.

By the end of 2002, when an initial version of NSDL will be open for first large-scale testing and deployment, there should be about 90 projects that the US NSF is supporting, in the 4 above mentioned tracks. Clearly, interoperability is essential, so there is widespread use of OAI by the Core Integration and Collection projects. However, while metadata can be harvested easily from a wide variety of sources, integrating a diversity of separately developed services is not planned for 2002.

Fortunately, ODL has been tried in a number of educational settings [8]. We believe that it can be effective with regard to integrating both collections and services, as is explained in the next subsections.

3.1 CITIDEL

Virginia Tech has primary responsibility for the CITIDEL part of NSDL [9]. This Collection project covers the topical areas of computing and information technology. Figure 4 explains both collection and services that are under development.

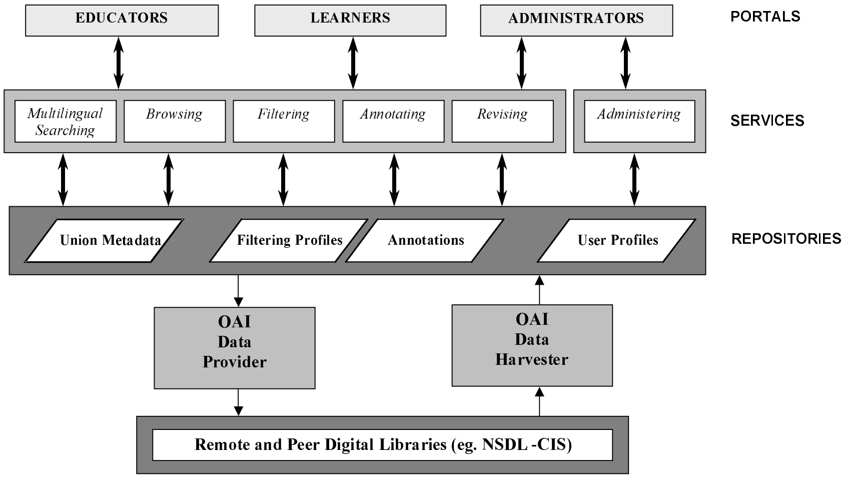


Fig. 4. CITIDEL schematic from original proposal showing collections and services.

Many of the requirements for CITIDEL can be met using existing ODL components, at least for an initial prototype. But when CITIDEL has a union collection with on the order of a million records, and becomes widely used by undergraduate and graduate students, as well as teachers/trainers and other learners (both younger, in public schools) and older (some as lifelong learners), performance may become an issue. Consequently, one of the research activities being explored with regard to ODL is the matter of performance.

Based in part on this experience with CITIDEL, we are working, along with NCSA, on a project awarded to University of Florida, in the NSDL Services track, as is explained in the next subsection.

3.2 DL-in-a-Box

As is explained in Section 1.1, the tendency in NSDL is for each newly funded project to start from scratch in developing software. Consequently, there is a real opportunity to reduce overall costs if new projects can instead begin with a basic but extensible digital library. Our web site for such a digital library in a box [26] aims to support such an approach. We are working to provide additional documentation of components and subsystems (compositions of components), as well as to develop additional components. We are open to requirements statements from others, comments on enhancements and extensions, and will provide full support as funding permits. We hope that gradually others will provide components as well, both those engaged in other NSDL activities, as well as those working on other projects, such as the ones discussed in the next section.

4 Other Applications

Digital libraries can be used in many application domains, and can extend traditional approaches [13]. In the subsections below we explore four other types of applications.

4.1 CSTC

The Computer Science Teaching Center [23] has been one of our test domains for DL development over the last 4 years. It covers the full spectrum of services from author submission to support of end-user searching and browsing. In addition, it supports peer review and editorial control, along with notification through email to editors and reviewers. Further, thanks to support from ACM, CSTC is connected with the ACM Journal of Educational Resources in Computing [5]. This means that submissions to CSTC may be considered for JERIC, and then may appear there if editorial concerns are all addressed. This interconnection is further complicated in that both CSTC and JERIC are collections that are part of CITIDEL, in each case with both their metadata and the full text covered. Fortunately, ODL allows modular development, so parts of CSTC have been replaced by components (e.g., Browse) while the rest of the system has stayed as-is. Clearly such incremental testing and development, and such flexible interconnection, bode well for ODL being deployed in legacy contexts as well as in new situations. A similar situation is considered in the next subsection too.

4.2 NDLTD

The Networked Digital Library of Theses and Dissertations, NDLTD [10], which supports graduate education [8], also has benefited from the ODL approach. This is fortunate, since NDLTD aims to support change [12] and hence must remain agile. The plan in NDLTD is for as many members as possible to become OAI data providers, so metadata can be easily harvested. Already, harvesting into a union catalog [36] occurs, and a set of services is provided (i.e., browse, recent, and search) [33, 34]. In addition, the union catalog feeds into the Virtua DL system developed by VTLS, Inc. Thus, we have services provided both through ODL and through a commercially available monolithic DL – allowing us to undertake scientific comparisons over the next year.

4.3 AmericanSouth.org

While Virginia Tech has lead responsibility in CITIDEL and NDLTD, it only provides technical assistance in the AmericanSouth.org effort, which is led by Emory University [19]. Thus, this activity demonstrates that others can deploy ODL. While we provide assistance, a number of universities around the Southeast, that are willing to employ OAI to make available metadata about local history and culture, can use components to build up their local services as well as their support for interoperability. Further, this effort has in part led to the OCKHAM effort, discussed in the next section.

4.4 Classes

To demonstrate further that ODL can be deployed easily, it was explained to learners, in several class and tutorial settings. These included at library and digital library conferences, as well as in the Virginia Tech course "Information Storage and Retrieval", wherein almost 70 students (in 2 sections, so that each student could work on their own computer):

- Learned about OAI and ODL
- Installed components on their computer
- Configured the components
- Ran the set of components as a small DL

It is clear that all this can take place in less than 3 hours, and that students can both learn a great deal and gain confidence in their understanding of DL practice. But for students focused in this area, it is helpful to set this in a broader context, as explained in the next section.

5 OCKHAM

In the summer of 2002, the Open Community Knowledge Hypermedia Applications & Metadata initiative was launched. A web site was developed [18], and discussion proceeded through a listserv [20]. The concept was disseminated at ECDL'2002 [21] and was well received; feedback suggests that further meetings will gain support.

There are four main ideas:

1. Components
2. Lightweight protocols
3. Open reference models
4. Community perspective and involvement

The first two have already been discussed above, and are the basis for ODL. The other points are explained in the next two subsections.

5.1 Open Reference Models

While it is clear that components and lightweight protocols are helpful when building DLs, more is needed. In the case of the applications discussed above, the context and assumed general architecture / reference model has been well understood. However, this is not always the case! Fortunately, however, the library and information science world has invested considerable time in preparing open reference models [3, 7, 15, 16, 25, 27, 30, 40]. Of particular interest are architectures like DNER, meetings and work encouraged by UKOLN, and efforts related to the archival community. In short, it is important that development of components and lightweight protocols takes place in a suitable framework, where modularity has been carefully thought through.

5.2 Community Perspective and Effort

Such frameworks, however, are in turn based on community activity. Thus, fundamentally, OCKHAM depends on efforts to achieve consensus by a group with a common aim. Only with a unified perspective can a community develop a reference model that in turn allows efficient and effective development of components and protocols. Fortunately, in cases such as NSDL and NDLTD, years of discussion and prototyping have led to clear understanding by a broad community.

6 Conclusion

As explained above, when the right conditions exist, it is possible to build DLs easily. We argue for the ODL approach, with components and lightweight protocols. Those work best when there is an open reference model, which has arisen to reflect community perspective, and where community effort helps carry the project forward. Yet, we live in a world where other forces also apply. In some cases we have existing subsystems. Thus, for example, in some cases we may want to simply achieve interoperability at the level of interconnecting DL and information visualization systems [42, 43]. Or, we may need to build upon a particular software infrastructure like Web Services [42]. Such situations may occur, and yet the ODL approach may apply, as long as key concepts, and the essential principle of simplicity, are carefully considered.

Acknowledgements. Portions of this work were funded by the US National Science Foundation through grants DUE-9752190, 9752408, 0121679, 0121741, and 0136690; and IIS-0002935, 0080748 and 0086227. Among these are subcontracts with original funding to UNC Wilmington, U. of Arizona, and U. of Florida.

Other aspects of this work were funded in part by the Mellon Foundation, through a subcontract, with original funding to SOLINET for AmericanSouth.org.

References

1. Baldonado, M., Chang, C.K., Gravano, L., and Paepcke, A. The Stanford Digital Library Metadata Architecture. International J. on Digital Libraries 1(2): 108-121, 1997.
2. Birmingham, W.P. An Agent-Based Architecture for Digital Libraries. D-Lib Magazine 1(7), July 1995
3. Blinco, K. Modeling Hybrid Information Environments: The Librarian and the Super Model. PowerPoint presentation for 9th MODELS workshop, 13-14 Oct 1999, UKOLN, http://www.ukoln.ac.uk/dlis/models/models9/presentations/kb-m9.ppt
4. Borgman, C.L. What are digital libraries? Competing visions. Information Processing and Management 35: 227-243, 1999.

5. Cassel, L., Fox, E.A. Introducing the ACM Journal of Educational Resources in Computing (JERIC), editor-in-chiefs' introduction. 1(1), March 2001, http://doi.acm.org/10.1145/376697.382399
6. Castelli, D., Pagano, P. OpenDLib: A Digital Library Service System. In "Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002, Proceedings", eds. Maristella Agosti and Constantino Thanos, pp. 292-308.
7. CCSDS (Consultative Committee for Space Data Systems). Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-R-1, Red Book, May 1999, e Model http://www.ccsds.org/documents/p2/CCSDS-650.0-R-1.pdf
8. Fox, E. Advancing Education through Digital Libraries: NSDL, CITIDEL, and NDLTD. In the Proceedings of Digital Library: IT Opportunities and Challenges in the New Millennium, ed. Sun Jiazheng, Beijing, China: Beijing Library Press, July 9-11, 2002, pp. 107-117.
9. Fox, E. CITIDEL: Computing and Information Technology Interactive Digital Educational Library, 2002, http://www.citidel.org
10. Fox, E. NDLTD: Networked Digital Library of Theses and Dissertations, 2002, http://www.ndltd.org
11. Fox, E. The 5S Framework for Digital Libraries and Two Case Studies: NDLTD and CSTC. In Proceedings NIT'99. Taipei, Taiwan, 1999. http://www.ndltd.org/pubs/nit99fox.doc
12. Fox, E., Gonçalves, M., McMillan, G., Eaton, J., Atkins, A., Kipp, N. The Networked Digital Library of Theses and Dissertations: Changes in the University Community. Journal of Computing in Higher Education, 13(2): 3-24, Spring 2002.
13. Fox, E., Marchionini, G. Digital Libraries: Extending Traditional Values. Guest Editors' Introduction to special section on Digital Libraries. Commun. of the ACM, 44(5): 30-32, May 2001, http://doi.acm.org/10.1145/374308.374329
14. Fox, E. and Urs, S. Digital Libraries. In Annual Review of Information Science and Technology (ARIST), v. 36, B. Cronin, Ed.: American Society for Information Science, 2001.
15. Gardner, T. The MIA Logical Architecture: MODELS Information Architecture (MIA) Requirements Analysis Architecture, UKOLN, 1999, http://www.ukoln.ac.uk/dlis/models/requirements/arch/
16. Garrett, J. ISO Archiving Standards – Overview. Last Revised: 29 July 2002. http://ssdoo.gsfc.nasa.gov/nost/isoas
17. Gonçalves, M.A., Fox, E.A. 5SL – A Language for Declarative Specification and Generation of Digital Libraries. In Proc. JCDL'2002, Second Joint ACM / IEEE-CS, Joint Conference on Digital Libraries, July 14-18, 2002, Portland, pp. 263-272.
18. Halbert, M., ed. OCKHAM: Open Community Knowledge Hypermedia Applications & Metadata. 2002. http://ockham.library.emory.edu
19. Halbert, M. D., ed. AmericanSouth.org: A joint project of Emory University and ASERL, sponsored by the Andrew W. Mellon Foundation. 2002. http://AmericanSouth.org
20. Halbert, M. D., ed. Archives of OCKHAM-SYS@LISTSERV.CC.EMORY.EDU: OCKHAM System Framework Listserv, 2002, http://www.listserv.emory.edu/archives/ockham-sys.html
21. Halbert, M. D., Morgan, E. L., Fox, E. A. OCKHAM: Coordinating Digital Library Development with Lightweight Reference Models. Panel at "Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002", Rome, Italy, September 16-18, 2002
22. Hunter, P. "The Management of Content: Universities and the Electronic Publishing Revolution". Ariadne Issue 28, 22-June-2001. http://www.ariadne.ac.uk/issue28/cms/

23. Knox, D., Fox, E.A., Suleman, H., eds. CSTC: Computer Science Teaching Center. 2002. http://www.cstc.org
24. Lagoze, C., and Van de Sompel, H. The Open Archives Initiative: Building a low-barrier interoperability framework. In Proceedings of JCDL 2001, Roanoke VA, June 2001, ACM Press, pp. 54-62.
25. Library of Congress. METS: Metadata Encoding & Transmission Standard, Official Web Site, Last Revised: February 19, 2002, http://www.loc.gov/standards/mets/
26. Luo, Ming. DL-in-a-box: digital library in a box, website. 2002. http://dlbox.nudl.org
27. Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., and Gupta, A. Collection-Based Persistent Digital Archives - Part 1. D-Lib Magazine, March 2000, 6(3), http://www.dlib.org/dlib/march00/moore/03moore-pt1.html
28. OAI. Open Archive Initiative. 2002. http://www.openarchive.org/.
29. Paepcke, A., Chag, C.-C. K., Garcia-Molina, H., Winograd, T. Interoperability for Digital Libraries Worldwide. Communications of the ACM, vol. 41, pp. 33-43, 1998.
30. Powell, A. JISC Information Environment Architecture, on "DNER Architecture" for the JISC Distributed National Electronic Resource, JISC, 2002 http://www.ukoln.ac.uk/distributed-systems/dner/arch/
31. NSDL. NSDL: The National Science Digital Library. 2002. http://www.nsdl.org
32. Shen, R., Jun Wang, Edward A. Fox. A Lightweight Protocol between Digital Libraries and Visualization Systems. JCDL Workshop on Visual Interfaces to Digital Libraries (see p. 425 of Proc. JCDL 2002), July 18, 2002, Portland.
33. Suleman, H., Atkins, A., Gonçalves, M.A., France, R.K., Fox, E.A., Chachra, V., Crowder, M., Young, J. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress. D-Lib Magazine, 7(9), Sept. 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html
34. Suleman, H., Atkins, A., Gonçalves, M.A., France, R.K., Fox, E.A., Chachra, V., Crowder, M., Young, J. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 2: Services and Research. D-Lib Magazine, 7(9), Sept. 2001, http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html
35. Suleman, H., Fox, E.A. Designing Protocols in Support of Digital Library Componentization. In "Research and Advanced Technology for Digital Libraries, 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002, Proceedings", eds. Maristella Agosti and Constantino Thanos, pp. 568-582.
36. Suleman, H., Fox, E.A. Towards Universal Accessibility of ETDs: Building the NDLTD Union Archive. In Proc. ETD'2002, BYU, Provo, Utah, May 30 - June 1, 2002, preprint at http://rocky.dlib.vt.edu/~hussein/etd_2002/etd_2002_paper_final.pdf
37. Suleman, H., Fox, E.A. The Open Archives Initiative: Realizing Simple and Effective Digital Library Interoperability. J. Library Automation, 35(1/2):125-145, 2002.
38. Suleman, H., Fox, E.A. A Framework for Building Open Digital Libraries. D-Lib Magazine, 7(12), Dec. 2001, http://www.dlib.org/dlib/december01/suleman/12suleman.html
39. Suleman, H. Open Digital Libraries. 2002. http://oai.dlib.vt.edu/odl/
40. Thibodeau, K. Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration. D-Lib Magazine February 2001, 7(2), http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html
41. Van de Sompel, H., Lagoze, C. The Open Archives Initiative Protocol for Metadata Harvesting. Open Archives Initiative, 2001. http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html
42. Vasudevan, V. A Web Services Primer, 2001. http://www.xml.com/pub/a/2001/04/04/webservices/index.html
43. Wang, J. VIDI: A Lightweight Protocol Between Visualization Tools and Digital Libraries. Master's Thesis, Virginia Tech, May 2002.

Dublin Core: Process and Principles

Shigeo Sugimoto[1], Thomas Baker[2], and Stuart L. Weibel[3]

[1] Institute of Library and Information Science, University of Tsukuba, Japan
sugimoto@slis.tsukuba.ac.jp
[2] Institutszentrum Schloss Birlinghoven, Fraunhofer-Gesellschaft, Germany
Thomas.Baker@bi.fhg.de
[3] Dublin Core Metadata Initiative, OCLC Office of Research
weibel@oclc.org

Abstract. The Dublin Core metadata element set has been widely adopted by cultural and scientific institutions, libraries, governments, and businesses to describe resources for discovery on the Internet. This paper provides an overview of its history and underlying principles and describes the activities of Dublin Core Metadata Initiative (DCMI) as an organization.

1 Introduction

In the Web age, metadata is typically defined as “data about data” – a simple definition that embraces a broad range of resources from library catalogues and indexes to thesauri, ratings, reviews, terms and conditions for use. In the Internet, metadata is designed for tasks ranging from resource description and discovery to archiving, trading, content filtering, resource syndication, and information management. This diversity of purpose reflects the variety of information resources available on the Internet, which range from personal Web pages to huge portals for government information, digital libraries, and shopping catalogues. Users of the Internet range from small children to businesses and professionals.

From its origins in the mid-1990s, the Dublin Core has defined itself as a small set of core descriptive attributes by which users can search for information across a broad range of sources [1]. It was recognized from the outset that semantic interoperability across domains implied conceptual simplicity, much in the manner of natural-language pidgins, which use small vocabularies and simple grammars to enable rough comprehension between speakers of different languages. As the Dublin Core became adapted for specialized purposes, however, the focus shifted to methods for qualifying and extending the core vocabulary, and to architectures for encoding Dublin Core descriptions interoperably.

The first, thirteen-element Dublin Core was the result of a workshop held in Dublin, Ohio in 1995, which had been planned as a result of a casual conversation at the Chicago WWW conference in 1994. Since then a series of workshops and conferences has built on this initial consensus, clarifying issues of architecture and extensibility and broadening international participation.

Since the initial “classic” Dublin Core, which stabilized at fifteen elements in 1998, several implementation styles have emerged. “Simple Dublin Core” uses the

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 25-35, 2002.

fifteen elements in a very broad, generic manner. In addition to such simplicity, the ability to use Dublin Core for more precise and detailed description was identified as an important need. "Qualified Dublin Core", therefore, uses additional terms to specify the meaning of the Core elements within the context of specific domains. Since such complexification seemed contrary to simplicity and interoperability among different domains, the Dublin Core community embraced the notion of a modular architecture and elaborated the so-called Dumb-Down Principle. A process model (centered around a Usage Board that acts as an editorial body) is used to approve new descriptive terms. The Dublin Core has evolved among participants who have differing requirements but can integrate their needs through the approval of modular extensions to the basic core.

2 Dublin Core Metadata Element Set (DCMES)

2.1 Simple Dublin Core and Qualified Dublin Core

"The Dublin Core" has been defined since 1998 as a set of fifteen elements for cross-domain resource discovery. The set of elements is shown in Table 1. By design, any of the fifteen elements is optional and repeatable. This set has been approved as an international standard in Europe (CEN/ISSS CWA 13874) and a national standard in the USA (ANSI/NISO Z39.85).

Table 1. The Fifteen Elements of "Simple Dublin Core"

Identifier	Definition
Title	A name given to the resource.
Creator	An entity primarily responsible for making the content of the resource.
Subject	The topic of the content of the resource.
Description	An account of the content of the resource.
Publisher	An entity responsible for making the resource available.
Contributor	An entity responsible for making contributions to the content of the resource.
Date	A date associated with an event in the life cycle of the resource.
Type	The nature or genre of the content of the resource.
Format	The physical or digital manifestation of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A reference to a resource from which the present resource is derived.
Language	A language of the intellectual content of the resource.
Relation	A reference to a related resource.
Coverage	The extent or scope of the content of the resource.
Rights	Information about rights held in and over the resource.

The use of these fifteen elements for metadata records, with no additional qualifiers, and with only plain-text strings as values, is known as "Simple Dublin Core" [2].

"Qualified Dublin Core", in contrast, uses the elements together with qualifiers that increase the richness and precision of description [3][4]. Table 2 shows a list of qualifiers approved by DCMI as “recommended” qualifiers as of September 2002. The approval process and status are explained in detail in section 3.4. Qualified DC has two types of qualifiers – element refinement and encoding schemes. An element refinement narrows the meaning of an associated element; for example, “Date Created” is a more narrowly defined instance of *Date*, and an *Abstract* is seen as a type of *Description*. An encoding scheme qualifier specifies a name of a vocabulary or a name of data encoding scheme used in encoding of a value of its associated element; for example, LCSH encoding scheme qualifier associated with Subject element specifies that a value of the Subject element is expressed in terms of the Library of Congress Subject Headings (LCSH). Qualified DC does not include qualifiers to express components of a value, such as first and last names.

DCMES is a stable but not a closed set. DCMES evolves in accordance with requirements to express resource properties and value-types which are not expressible using existing ones. Table 2 includes a qualifier associated with *Audience* element, which was approved in 2001. The definition of *Audience* element is “A class of entity for whom the resource is intended or useful”. *Audience* element was originally proposed by the working group on educational applications and approved as a recommended element for the global community.

Table 2. DCMI Recommended Qualifiers

Element	Element Refinement	Encoding Scheme
Title	Alternative	
Subject		LCSH, MeSH, DDC, LCC, UDC
Description	Table of Contents, Abstract	
Date	Created, Valid, Available, Issued, Modified	DCMI Period, W3C-DTF
Type		DCMI Type Vocabulary
Format	Extent	
	Medium	IMT
Identifier		URI
Source		URI
Language		ISO 639-2, RFC 1766, RFC 3066
Relation	Is Version Of, Has Version, Is Replaced By, Replaces, Is Required By, Requires, Is Part Of, Has Part, Is Referenced By, References, Is Format Of, Has Format, Conforms To	URI
Coverage	Spatial	DCMI Point, ISO 3166, DCMI Box, TGN
	Temporal	DCMI Period, W3C-DTF
Audience	Mediator	

2.2 Encoding Dublin Core Metadata

DCMI provides documents describing three predominant encoding styles for Dublin Core metadata. The oldest of these styles embeds Dublin Core descriptions in HTML with special tags [5]. In the example below, META tags are used to hold the descriptive elements and their values.

```
 <meta name=”DC.Title”
    content=”Dublin Core Metadata Initiative Home Page”>
<meta name=”DC.Language” content=”en”>
<meta name=”DC.Contributor”
      content=”Dublin Core Metadata Initiative”>
<meta name=”DC.Date” content=”2001-01-16”>
<meta name=”DC.Format” content=”text/html”>
```

In the HTML style, LANG attributes are used to indicate the language of metadata values, as in the following example in German.

```
<meta name=”DC.Title” lang=”de”
    content=”Dublin-Core Metadata-Diskussionen”>
```

In XML encoding, element names appear in the tags with a prefix “dc”, which is associated to a namespace [6].

```
<?xml version="1.0"?>
<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="..." xsi:schemaLocation="..."
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <dc:title>UKOLN</dc:title>
  <dc:description>UKOLN is a national focus of expertise
   in digital information management...</dc:description>
  <dc:publisher>UKOLN, University of Bath</dc:publisher>
  <dc:identifier>http://www.ukoln.ac.uk/</dc:identifier>
</metadata>
```

At its simplest, encodings in RDF (below) resemble the XML style above [7][8]. As detailed in the draft DCMI recommendations, however, metadata in RDF provides conventional ways to embed diverse types of information into metadata without compromising the Dumb-Down Principle.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description>
    <dc:creator>Karl Mustermann</dc:creator>
    <dc:title>Algebra</dc:title>
    <dc:subject>mathematics</dc:subject>
    <dc:date>2000-01-23</dc:date>
    <dc:language>EN</dc:language>
    <dc:description>An introduction to algebra</dc:description>
  </rdf:Description>
</rdf:RDF>
```

3 Dublin Core Principles

DCMI is a forum for development of metadata standard for resource discovery across domains and is opened to international and multilingual communities. Simplicity, extensibility and semantic interoperability are the fundamental technical characteristics of the development of Dublin Core. This section discusses these key concepts.

3.1 Warwick Framework – Basic Framework for Extensibility

Since the Internet is a very diversified environment, it is useless to assume that a single metadata set will meet the needs of all domains and purposes. It is also impractical to develop metadata sets application by application: the result would be expensive and chaotic, and interoperability would be non-existent. On the other hand, it is desirable for application developers to use established metadata schemas and adopt them in accordance with local requirements. The Warwick Framework, a conceptual model that resulted from the 2nd Workshop in 1996, gave an early expression to the notion of metadata as modular components that may come from more than one metadata schema [9]. In this model, a metadata instance is expressed as a container which contains one or more packages, each of which is expressed in a given metadata schema. The Resource Description Framework (RDF), the development of which began in 1998, provided a practical realization of many of the ideas of the Warwick Framework [10].

The Warwick Framework is important as a model for modular metadata on the Internet. No single metadata schema is sufficient to all applications. Rather, it is necessary to adopt appropriate elements from various schemas in accordance with the functional requirements of an application. The role of Dublin Core in Warwick Framework is to provide a core set of metadata elements for resource discovery in any application domains. In other words, Dublin Core can work as a common schema to find resources across many domains.

3.2 The Dumb-Down Principle as a Basis for Interoperability

The Dumb-Down principle gives a guideline for qualification. The Dumb-Down principle suggests that a value of a qualified element has to be consistent as a value of the element without any qualification. For example, assume the following qualified values;

1. (Element Refinement) Date Accepted: “2002-12-11”,
2. (Encoding Scheme) Language: “en” encoded in RFC 1766, and
3. (Value Structure) Creator: {name: “Sugimoto, Shigeo”, affiliation: “University of Tsukuba”, contact: “sugimoto@slis.tsukuba.ac.jp”}

Then, assume that qualification in the above examples, *Accepted*, *RFC 1766* and the component names of the value structure (i.e., *name*, *affiliation* and *contact*) are removed. The values of example 1 and 2, “2002-12-11” and “en” are consistent with their elements after the removal. However, the value of example 3 {“Sugimoto, Shigeo”, “University of Tsukuba”, ”sugimoto@slis.tsukuba.ac.jp”} causes problems since the second and third values are not valid values of *Creator*.

Dumbing-down is a crucial function for metadata interoperability in the global community; local communities can extend their schemas in accordance with their requirements whereas they are also encouraged to keep their metadata interoperable with other metadata communities.

3.3 Application Profiles

Dublin Core Metadata defines the vocabulary of metadata, i.e., terms and their meanings, but in general does not specify the encoding or syntax characteristics. An exception is the feature included in Simple DC that is “Any of the 15 elements is optional and repeatable.” Local applications, however, may have domain specific requirements appropriate to a given domain or application:

- Title, Creator and Description might be mandated but others are optional,
- Use only Title, Creator, Description, Date and Language elements,
- Use the 15 elements of Simple DC and some elements from other metadata sets such as the IEEE Learning Object Metadata, and so forth.

These requirements can be defined independently of the definition of vocabulary. Description of this application specific syntactic feature is called an *application profile*. Any application can have its own application profile, which specifies a set of metadata vocabulary terms used in the application as well as syntactic or structural features of the particular application. The vocabulary terms could be borrowed from one or more source schemas. More importantly, the application profile could be used to define a mapping between the application’s scheme to global scheme(s), which is crucial for interoperability.

3.4 A Process Model for Maintaining the Standard

To remain relevant in a rapidly evolving Web environment, Dublin Core must be able to grow and evolve in response to user needs. DCMI has therefore instituted a Usage Board and a process model for reviewing proposals for expanding or clarifying the standard. Primary among these functions is the review of proposals for new elements and qualifiers (generically, “terms”). Requirements for new terms may originate in a particular application community. DCMI working groups crystallize these requirements both on mailing lists and in face-to-face meetings and formulate proposals for presentation to the Usage Board.

The Usage Board evaluates such proposals for their conformance to architectural and grammatical principles. This Dublin Core "grammar" includes a typology of Elements, Element Refinements, and Encoding Schemes along with some general principles, such as the axiom that the values of element refinements should be usable as values of the element refined. Proposed elements and element refinements that conform to Dublin Core principles are taken into the standard with the status of *conforming*. To some proposed terms of proven usefulness for resource discovery across domains the Board may assign the status of *recommended*. Proposals for encoding schemes are reviewed for accuracy and given the status of *registered*.

Once approved, each new term is assigned a Uniform Resource Identifier using one of the official namespace URIs maintained by DCMI. A “namespace policy” defines limits within which the metadata terms maintained by DCMI can evolve or change over time. According to this policy, editorial changes or updates are allowed, but changes of semantics (meaning) are not; new semantics require the creation of new element.

The Usage Board has met twice in 2001 and once in 2002, defining formal review processes, developing procedures for registering externally maintained encoding schemes, and approving several proposals for new terms. Proposals which are not approved are sent back to working groups with suggestions on how they might be revised and resubmitted. The process has the feel of the review board for a scientific journal or conference, where reviewers may actively engage with authors for the common purpose of improving the end results.

The underlying motivation for the Usage Board is to provide a framework in which metadata requirements that "bubble up" in particular implementation contexts can be shared in wider circles and eventually be incorporated into a standard where they will be declared in a persistent way and maintained in accordance with known principles. This reflects the conviction that metadata usage, analogously to language usage in general, can only partially be steered from the top down, on the model of traditional standardization activity. In the DCMI model, the art of standards development lies in striking balances between innovation from below and qualified review from above or between domain-specific specificity and cross-domain applicability. The Usage Board process aims to guide the formulation and formalization of community standards for particular domains that integrate well into broader frameworks for interoperability.

3.5 Internationalization and Localization

As described above, Dublin Core is intended for resource discovery on the global Internet. There are several issues which have been identified as crucial for the adoption of Dublin Core by the international community where resources and their metadata are created in different languages under different cultures.

DCMI has been soliciting local communities to translate descriptions of DCMI terms and other documents into local languages in order to promote understanding by non-English speaking people. For example, the DCMI registry described below provides translations of DCMES into 23 languages, which have been translated mostly by volunteers.

A local community which shares a local language and/or culture plays a crucial role for the global use of DC metadata. Translation is a part of the efforts to adopt DC metadata to a local language. Only a local community can identify local requirements based on a local language and culture and let the global community know the requirements. DCMI is promoting the formation of regional organizations to support the local activities and development of local communities.

3.6 Metadata Schema Registry for Information Sharing

A Metadata schema registry is an infrastructure function sharing metadata schemas on the network to enhance interoperability of metadata. DCMI is building a metadata schema registry, which provides reference descriptions of DC terms. The reference descriptions are declared using RDF schema to promote readability and exchange by machines and applications. The reference description encoded in RDF schema provides an identifier given to a term which is uniquely identifiable in the Internet, and also a label and a description which could be given not only in English but also in other languages. Thus, the registry associates human understandable labels and descriptions in multiple languages with a unique identifier for machine understandability.

The registry will play an important role for the long-term maintenance of the reference descriptions, which is a crucial but challenging issue. Every DCMI term has its approval status and human readable label and descriptions, which could change over time. Every term could have translations which may be appended and modified over time [11]. Maintenance of local or domain specific schemas is also an important and challenging issue because the community maintaining the registry has to maintain consistency with other registries such as the central DCMI registry.

4 Dublin Core Metadata Initiative (DCMI)

4.1 Dublin Core Metadata Initiative and History in Brief

The DCMI is built on a community of individuals from many different backgrounds and disciplines located in organizations and institutions all over the world. The mission of the DCMI is to make it easier to find resources using the Internet through the following activities:

- Developing metadata standards for discovery across domains ;
- Defining frameworks for the interoperation of metadata sets;
- Facilitating the development of community or discipline-specific metadata sets that work within the frameworks of cross-domain discovery and metadata interoperability.

The major structural components of DCMI as of 2002 are the Directorate, Board of Trustees, Advisory Board, Usage Board, Working Groups and Interest Groups.

- The Dublin Core Directorate consists of an Executive Director and a Managing Director to supervise the management and coordination of Working Group activities and assist in the development and refinement of techniques promoting metadata interoperability. The directorate also oversees the development of the Web site and related infrastructure.
- The Board of Trustees advises the Directorate on strategic issues and allocation of financial resources, contributes to the promotion of the Initiative through liaisons with the public and private sectors and assists in securing support for the Initiative. The trustees were chosen to provide strategic leadership and support to the organization, and were selected for their leadership and professional abilities in the public, private, and educational sectors. Board members come from six countries on four continents.
- The Dublin Core Advisory Board is comprised of all chairs of DCMI Working Groups and Interest Groups and invited experts. The Advisory Board gives advice to the DCMI Directorate on all technical and strategic issues that occur during the operation of the DCMI. It has a dual role in the DCMI: an internal role to assist in and advise on the developments that take place within DCMI,

and an external role to liaise with the stakeholder community and other global metadata initiatives.

- The mission of the DCMI Usage Board is to ensure an orderly evolution of metadata vocabularies. The Usage Board evaluates proposed vocabulary terms (or changes to existing terms) in light of grammatical principle, semantic clarity, usefulness, and overlap with existing terms. The Usage Committee strives for consensus, justifying its decisions and interpretations in terms both of principle and of empirical practice.
- Working Groups and Interest Groups are formed and dissolved as necessitated by the work at hand and the availability of expertise to accomplish such work. Working and Interest groups will be comprised of volunteers with the interest, expertise, and time to contribute to the solution of problems.

The workshop series and the mailing lists are the major forums for discussion of the development of Dublin Core metadata. Table 3 shows the locations and primary hosts of the Dublin Core workshop series since 1995. There was an active discussion on qualifiers at the 4th Workshop in Canberra, Australia in 1997. Simple DC was fixed at the 5th Workshop in Helsinki in 1997. The Dumb-down principle proposed at the 6th Workshop in Washington DC clarified the qualifier types. Development of the fundamental concept of Dublin Core was completed by the 6th Workshop, and maintenance and evolution of Dublin Core were recognized as an important topic since the 7th Workshop in Frankfurt in 1999. Organizational issue for sustainability of DCMI became one of the key issues since this workshop. The 8th Workshop in Ottawa was the first meeting which included sessions for posters and demos to report implementation experiences and new technologies. At the 9th Workshop the presentation session was extended and the whole event was organized as an international conference, which was named DC-2001 [12].

4.2 DCMI and Other Metadata Initiatives

DCMI and IEEE-LOM. At DC-8 in Ottawa in October 2000, DCMI and representatives of the IEEE-Learning Object Metadata (LOM) working group concluded a memorandum of understanding indicating areas of possible convergence on principles and encoding approaches that have the potential to increase interoperability between the two communities. A subsequent meeting in Ottawa in August 2001 identified specific work items. A prominent deliverable from this activity is the recently published "Metadata Principles and Practicalities," an expression of agreement among leaders in the Dublin Core community and the e-learning community concerning basic principles of metadata [13]. This consensus should value to metadata practitioners in these respective communities as well as among metadata practitioners in general.

Dublin Core and Open Archives Initiative (OAI). The Dublin Core Metadata Initiative and the Open Archives Initiative are actively cooperating on metadata issues. Unqualified DC metadata is the default metadata set used in the OAI Protocol for Metadata Harvesting for the purposes of promoting cross-domain interoperability. Other domain-specific sets are encouraged as well, as envisaged in the modular metadata framework that both communities have been striving for. The OAI-DC schema has been developed for use with the OAI Protocol, and has been discussed at length in the DC-Architecture working group. It is expected that the schema will be of use for other applications as well, and will be hosted on the DCMI Website and maintained by representatives of both groups. This development is an important landmark in the development of Web-based metadata services, reflecting as it does the convergence of community consensus and the development of enabling infrastructure to support that consensus.

Table 3. Dublin Core Workshop Series

	Year	Location	Primary Host
1	1995	Dublin, Ohio, USA	OCLC
2	1996	Warwick, UK	UKOLN
3	1996	Dublin, Ohio, USA	OCLC
4	1997	Canberra, Australia	National Library of Australia
5	1997	Helsinki, Finland	National Library of Finland
6	1998	Washington DC, USA	Library of Congress
7	1999	Frankfurt, Germany	Die Deutsche Bibliothek
8	2000	Ottawa, Canada	National Library of Canada
9	2001	Tokyo, Japan	National Institute of Informatics
10	2002	Florence, Italy	Biblioteca Nazionale Centrale Firenze

W3C Semantic Web Activity. The launch of the Semantic Web activity by the W3C recognizes the increasing importance of supporting the infrastructure for defining, registering, and referencing structured vocabularies and ontologies on the Web. The Dublin Core is an important part of this infrastructure, and the DCMI community has played a major role in laying the foundations for this work. A joint project between DCMI staff and W3C staff now under development will help illustrate the value of combining technologies such as the Resource Description Framework of the W3C with the Dublin Core to advance semantic interoperability on the Web.

The joint project between DCMI and W3C staff will unify access to a substantial amount of data from different sectors in different countries using RDF schema declarations as described above. Participants will be recruited from the government sector, museums, business, trans-governmental organizations, and education. The resulting database will comprise a testbed accessible to researchers and designers to demonstrate and experiment with an operational cross-disciplinary store. It will provide a tutorial by example on a schema-based approach to enhancing cross-domain interoperability.

5 Conclusion

Dublin Core has gained wide acceptance and many metadata applications have been developed based on Dublin Core. Since its beginning in 1995, Dublin Core has evolved; its underlying concepts have been clarified and the community model for maintaining Dublin Core has been accepted. DCMI has been promoting cooperation with other metadata communities, which will greatly enhance semantic interoperability of metadata. On the other hand, there is a lot of work left to do; for example, long-term maintenance of DC metadata in the multi-language community, development of regional communities, and further broadening of the uptake of DC metadata by other communities.

References

1. Dublin Core Metadata Initiative (DCMI) Home Page, http://dublincore.org/ (2001)
2. Dublin Core Metadata Element Set, Version 1.1: Reference Description, http://dublincore.org/documents/1999/07/02/dces/ (1999)
3. Dublin Core Qualifiers, http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/ (2000)
4. Baker, T.: A Grammar of Dublin Core, D-Lib Magazine 6 (2000) http://www.dlib.org/dlib/october00/baker/10baker.html
5. Kunze, J.: Encoding Dublin Core Metadata in HTML, http://www.ietf.org/rfc/rfc2731.txt (1999)
6. Powell A., Johnston, P.: Guidelines for Implementing Dublin Core in XML, http://dublincore.org/documents/2002/09/09/dc-xml-guidelines/ (2002)
7. Beckett, D., et al.: Expressing Simple Dublin Core in RDF/XML, http://dublincore.org/documents/2001/11/28/dcmes-xml/ (2001)
8. Kokkelink, S. Schwaenzl, R.: Expressing Qualified Dublin Core in RDF/XML, http://dublincore.org/documents/2002/04/14/dcq-rdf-xml/ (2002)
9. Lagoze, C.: The Warwick Framework, D-Lib Magazine 2 (1996) http://www.dlib.org/dlib/july96/lagoze/07lagoze.html
10. Resource Description Framework (RDF), http://www.w3.org/RDF/
11. Sugimoto, S. et al.: Versioning the Dublin Core across Multiple Languages and over Time, Proceedings of International Symposium and Applications and Internet 2001 Workshop (2001) 35-45
12. Sugimoto, S., et al.: Report from International Conference on Dublin Core and Metadata Applications 2001, Proceedings of 68th IFLA General Conference and Council, http://www.ifla.org/IV/ifla68/papers/073-151e.pdf (2002)
13. Weibel S. L., et al.: Metadata Principles and Practicalities, D-Lib Magazine 8 (2002) http://www.dlib.org/dlib/april02/weibel/04weibel.html

From Digital Library to Digital Government: A Case Study in Crime Data Mapping and Mining

Hsinchun Chen

Artificial Intelligence Lab and Hoffman E-Commerce Lab
University of Arizona, McClelland Hall 430Z, Tucson, AZ 85721, USA
hchen@bpa.arizona.edu

Abstract. In light of the significant research activities in digital library, digital government, and e-commerce over the past decade, there seems to be common threads among them and unique challenges and opportunities ahead. For digital library, we are beginning to tally its research impacts and contemplate future directions. For digital government, information technologies could offer tremendous opportunities, but will they happen fast enough? We hope by discussing the many unique problems and challenges facing these fast evolving and somewhat related research disciplines, we could share lessons learned and develop insights for advancing our knowledge and achieving successful organizational transformation. A detailed case study of a research project (COPLINK) jointly funded by the NSF Digital Library and Digital Government Programs in the area of crime data mining will be presented in detail. We discuss how advanced visual crime mining techniques such as association rule mining, social network analysis, deception detection, temporal-spatial visualization, and agents could become the catalyst for assisting intelligence and law enforcement agencies in capturing knowledge and creating transformation.

1 Introduction

The Internet is changing the way we live and do business. Since the first ARPANET node installed at UCLA on September 1, 1969 and the first paper on Internet, written by Vint Cerf and Bob Kahn on September 10, 1973 (Cerf, 2002), the Internet has evolved from ftp file transfer, gopher information service, and email exchange to supporting seamless multimedia content creation, access, and transactions over the World Wide Web.

Some researchers and practitioners believe that business, technology, and society in general are in a true "Digital Renaissance" (Fiorina, 2000). As HP CEO Fiorina put it:

"Like the first Renaissance, which was the liberation of the inventive imagination, the Digital Renaissance is about the empowerment of the individual and the consumer. And if we can bridge the gap between business, science, and government so that we can all understand and foster the Digital Renaissance then we have a chance to make this second Renaissance truly global and grassroots."

Using HP as an example, she suggested three emerging forces in the technology and business landscape: information appliances, always-on IT infrastructure, and e-

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 36-52, 2002.

services. Information appliances are anything with a chip inside and able to connect to the Internet. The always-on IT infrastructure needs to be as available and reliable as tap water and electricity. And thirdly, e-services will take any process or any asset that can be digitized and deliver it over the Web.

This viewpoint is frequently echoed by other Internet pioneers such as Vint Cerf. He has been preaching of the next-generation Internet as a medium for Internet-enabled appliances (e.g., Internet-enabled automobiles for maintenance and tax collection, Internet-enabled wine corks for ideal storage and drinking condition, etc.), real-time Internet multimedia supports (e.g., Internet multicast video, Internet-enabled VOIP call centers, net-based speech recognition, etc.), and even "Interplanetary Internet" for supporting future NASA Mars and other planetary explorations (and continuous, future data collection and simulation).

In spite of such a positive outlook, many researchers and policy makers caution against the potential pitfalls of technology innovation without careful policy considerations in areas such as privacy/security, cryptography and export, trademarks, domain names and copyright issues, regulatory framework, taxation, liability and dispute resolution, censorship, and digital signatures and certificates, to name a few (Cerf, 2002). While business and technology are in a true Digital Renaissance, we cannot afford to have our whole approach to policymaking remain rooted in the industrial, medieval world.

The Internet offers a tremendous opportunity for many different traditional institutions such as libraries, governments, and businesses to better deliver their contents and services and interact with their constituents – citizens, patrons, businesses, and other government partners. In addition to providing information, communication, and transaction services, exciting and innovative transformation could occur with new technologies and practices. Data and information can begin to become knowledge assets. Digital Library (e-library), digital government (e-government), and e-commerce research have many common threads, yet each faces some unique challenges and opportunities. This paper provides a brief review of digital library and digital government research and presents a case study in using data mapping and visual data mining techniques for a digital government application in crime analysis.

2 Digital Library: The Field

The location and provision of information services has dramatically changed over the last ten years. There is no need to leave the home or office to locate and access information now readily available on-line via digital gateways furnished by a wide variety of information providers, e.g., libraries, electronic publishers, businesses, organizations, individuals. Information access is no longer restricted to what is physically available in the nearest library. It is electronically accessible from a wide variety of globally distributed information repositories (Schatz & Chen, 1996) (Schatz & Chen, 1999).

Information is no longer simply text and pictures. It is electronically available in a wide variety of formats, many of which are large, complex (i.e., video and audio) and often integrated (i.e., multimedia). This increased variety of information allows one to take virtual tours of museums, historical sites and natural wonders, attend virtual concerts and theater performances, watch a variety of movies, and read, view or listen to books, articles, lectures and music, all through digital libraries.

The World-Wide Web has made access to the Internet part of everyday life. At the same time, over the past few years, the primary interface to the Web has evolved from browsing to searching. Millions of people all over the world perform web searching every day. But the commercial technology of searching large collections has remained largely unchanged from its roots in US government-sponsored research projects of the 1960s. This public awareness of the Net as a critical infrastructure in the 1990s has caused a new revolution in the technologies for information and knowledge management in digital libraries.

Digital libraries represent a form of information technology in which social impact matters as much as technological advancement. It is hard to evaluate a new technology in the absence of real users and large collections. The best way to develop effective new technology is in multi-year large-scale research projects that use real-world electronic testbeds for actual users and aim at developing new, comprehensive, and user-friendly technologies for digital libraries. Typically, these testbed projects also examine the broad social, economic, legal, ethical, and cross-cultural contexts and impacts of digital library research.

2.1 Digital Library Challenges

Unlike digital government or e-commerce, digital library researchers face some unique challenges:

- Cultural and historical heritage: Many digital library and museum collections contain artifacts that are fragile, precious, and of historical significance. Many different countries are quickly moving towards digitizing their unique cultural and historical collections. However, the selection and digitization process has not been easy, both for technical, organizational, and economic reasons.
- Heterogeneity of content and media types: Digital library collections have the widest range of content and media types, ranging from 3D chemical structures to tornado simulation models, from the statue of David to paintings by Van Gogh. A mix of text, audio, and video is common among digital library applications. Collection, organization, indexing, searching, and analysis of such diverse information content continues to create unique technical challenges.
- Intellectual property issues: Unlike digital government or e-commerce applications that often generate their own content, digital libraries provide content management and retrieval services to many other content owners. The intellectual property issues (rights and fee collection) surrounding such diverse collections need to be addressed.
- Cost and sustainability issues: Many patrons often would like library services to be “free” or at least extremely affordable. Compounding the issue further is the notion of “free” Internet content. However, for high-quality, credible content to be accessible through digital libraries, cost and sustainability problems needed to be resolved. Different digital library pricing models would need to be developed for different contents and services.

- Universal access and international collaboration: Digital library content is often of interest to not just people in one region, but possibly all over the world. Many content creation and development processes also require collaboration among researchers and librarians in different parts of the world. Digital library researchers are facing the unique challenge of creating a global service that bridges cultural and language barriers.

3 Digital Government: The Field

Unlike digital library and e-commerce, which have attracted significant attention and research since 1994, various levels of governments, both in US and international, have been slow to adopt new Internet-enabled information technologies or to develop consistent research programs.

3.1 US Government Goes Electronic: The History

The US Government's painstaking process of "going electronic" is a good illustration of some of the unique challenges and issues facing government. Many legislations and regulations concerning information technologies and Internet in particular were developed only recently (Taschek, 2002), such as:

- 1986 Brooks Act amended: This is the first act to reduce government costs through volume buying, including IT purchases.
- 1996 Information Technology Management Reform Act: Not until 1996, did the US Government establish the CIO position to manage IT resources.
- 1998 WebGov portal: After seeing many successful Internet applications in the business sector, the US Government announced in August 1998 the WebGov portal project that aimed to provide one-stop information dissemination for the government. The project failed and was later replaced by FirstGov portal after a technology donation from Inktomi.
- 2000 Federal Rehabilitation Act: The government requires all IT products to be accessible to the disabled.
- 2000 FirstGov portal (http://firstgov.gov/) unveiled in June 2000.
- 2001 National Security Telecommunications and Information Systems Security Policy No. 11: It mandates all off-the-shelf software used in defense be evaluated by an approved third part (i.e., National Security Agency).
- 2001 Health Insurance Portability and Accountability Act (HIPPA): This important legislation requires all health care information to be in compliance with privacy regulations.
- 2002 E-Government Act: The Act funds additional e-government initiatives and creates the Office of Electronic Government.

As evident in these laws and regulations, even for the US Government Internet-enabled digital government activities did not begin until 1998. In addition, digital government faces many unique, but nonetheless important, policy issues such as providing equal access to the disabled, security, and privacy.

3.2 Information Technology Research, Innovation, and E-government: An NRC Report

In response to a request from the NSF for advice on planning for e-government innovation programs, the Computer Science and Telecommunications Board (CSTB) of the National Research Council (NRC) convened the Committee on Computing and Communications Research to Enable Better Use of Information Technology in Government. The committee was charged with examining how IT research can improve existing government services, operations, and interactions with citizens, as well as create new ones. The committee presented the final results of its study in (NRC, 2002) and offers recommendations intended to foster more effective collaboration between IT researchers and government agencies. Some of the key recommendations include:

- Government should continue to improve its support for transactions with individuals, businesses, and organizations. In doing so, it should emulate, where possible, the commercial trend toward integration of services to improve usability for customers.
- Government should adopt commercial e-commerce technologies and associated practices wherever possible.
- Government should continue to participate actively in developing a full range of information technologies. At the same time, government should leverage its role as a long-term supporter of IT research to embrace the e-government challenge within broad research programs and to stimulate more targeted technology development to meet particular needs.
- Consideration should be given to providing specific mechanisms, such as a centrally managed cross-agency IT innovation fund, as incentives to enable government organizations to undertake innovative and risky IT projects.
- Government should develop more effective means for undertaking multi-agency collaborative efforts that support aggressive prototyping, technology evaluation, and technology transition in support of e-government.

3.3 NSF Digital Government Research Program

As digital library research continues to draw the attention of many researchers and practitioners, the National Science Foundation in the US initiated its first program in Digital Government in 1998. According to the program announcement (NSF, 1998):

"The Federal Government is a major user of information technologies, a collector and maintainer of very large data sets, and a provider of critical and often unique information services to individuals, states, businesses, and other customers. The goal of the Digital Government Program is to fund research at the intersection of the computer and information sciences research communities and the mid- to long-term research, development, and experimental deployment needs of the Federal information service communities. The Internet, which was created from a successful partnership between Government agencies and the information technologies research community, is a major motivating factor and context for this program.

The coming decade will see the potential for nearly ubiquitous access to government information services by citizens/customers using highly capable digital information/entertainment appliances. Given the inexorable progress toward faster computer microprocessors, greater network bandwidth, and expanded storage and computing power at the desktop, citizens will expect a government that responds quickly and accurately while ensuring privacy. Enhancements derived from new information technology-based services can be expected to contribute to reinvented and economical government services, and more productive government employees. As society relies more and more on network technologies, it is essential that the Federal Government make the most effective use of these improvements."

The NSF Digital Government program has attracted participation from numerous federal, state, and local agencies such as: Bureau of Labor Statistics, Department of Agriculture, Department of the Interior, National Imagery and Mapping Agency, National Institute of Standards and Technology, US Patents and Trademark Office, National Institutes of Health, National Archives and Records Administration, Environmental Protection Agency, National Institute of Justice, Tucson Police Department, and Phoenix Police Department. By working closely with agency partners, the digital government projects emulate the successful partnership model adopted in digital library research and have begun to generate significant research findings in many areas (http://www.digitalgovernment.org/archive/projects.jsp) (Ambrite et al., 2001) (Elmagarmid and McIver, 2001).

Many of the NSF-funded digital government projects are related to geo-spatial content and services, ranging from event and process tagging for the international Gulf of Maine watershed to geo-spatial data mining, and from a geo-spatial decision support system for drought to spatial analysis for the Oregon coast. Another similar category of research is related to ecosystem and biodiversity, including projects such as: biodiversity information organization using taxonomy, infrastructure for heterogeneity of ecosystem data, collaborators, and organizations, digital aerial video mosaics for ecosystem carbon cycle modeling, and radar remote sensing of habitat structure for biodiversity informatics. Many federal agencies, including NASA, NIMA, NIH, NASA, EPA, are clearly interested in geo-spatial and ecosystem content and services.

Federal statistical data related research also is a major digital government research category. Sample projects include: adaptive interfaces for collecting survey data from users, energy data collection and access, quality graphics for statistical summaries, etc. Other related applications include: state and federal family and social services, FedStats secure collaborative environment, and improved privacy for census data through statistical means.

A few policy-oriented digital government projects are quite innovative. These include studies on Internet voting (e-voting), multinational investigation of new models of collaboration for government services, and building leadership for a digital government. The COPLINK project developed by the University of Arizona is also one of the successful projects under the NSF Digital Government program. In addition to solving the information sharing, data/text mining, and knowledge management problems facing local law enforcement agencies, the project has high potential impact for addressing homeland security issues among the law enforcement and intelligence community. A detailed case study of COPLINK will be provided in a later section.

In order to facilitate exchanges between information technology researchers and various federal, state, and local government partners, the NSF has supported the creation of the Digital Government Research Center (DGRC). In addition, an annual NSF-sponsored Digital Government Conference helps bring researchers and government partners together for an active exchange of research results and collaborative projects (http://ww.dgrc.org/dgo2002). Several research centers and organizations are dedicated to e-government related research such as: Center for Technology in Government at Albany (http://www.ctg.albany.edu), Digital Government Research Center at the University of Southern California and Columbia University (http://www.isi.edu/dgrc), and COPLINK Center for Law Enforcement (http://ai.bpa.arizona.edu/COPLINK).

3.4 European Union and Other International Digital Government Programs

In the EU, many digital government initiatives are under way. These programs vary widely and include several major research areas (ERCIM, 2002):

- Online public service and one-stop shopping for information content,
- E-politics, e-democracy, e-voting,
- Transactions, security, and digital signatures for e-government, and
- Business and political issues of relevance to e-government.

The European project "eGOV" is an example of the type of projects that focus on implementing one-stop government. Developed at the University of Linz, Austria, its major components comprise an integrated platform, a standardized data and communication exchange format, and process models for online public services. HELP (http://www.help.gv.at) is the national e-government portal of the Austrian government and a platform for all Austrian authorities to support official proceedings. The One-Stop-Shop model developed in Italy is another example of an integrated architecture and interface between the citizens and the European public administrations. Similarly, the FASME project (http://www.fasme.org/), developed at the University of Zurich, aims to offer a holistic architecture for international e-government services.

Several EU projects are related to e-politics, e-democracy, and e-voting. The DEMOS project provides Internet services facilitating democratic discussions and participative public opinion formation (http://www.demos-project.org). In the city of Esslingen in Germany, the Internet was used to involve citizens in an informal discussion about plans for a neighborhood development project. Researchers at the University of Amsterdam developed the "Coalition Logic", which could lead to automatic generation of the "fairest" voting procedure (http://www.cwi.nl/~pauly/).

Several municipal projects are related to digital signatures including Esslingen's e-government project in Germany and the Italian judicial system that relies on the safe and secure transmission of legal documents. In addition, many projects are related to unique business, political, and national issues and priorities. For example, the Italian approach to e-government is based on the development and deployment of a nationwide Public Administration Network, a "secure Intranet" (http://www.aipa.it/). In Hungary, its e-government program aims to integrate the strategies and IT development projects of various sectors and institutions in order to provide citizens with better services (http://www.ikb.hu/ekormanyzat/pdf/angol_ekp.pdf).

Many successful ongoing e-government initiatives have also emerged in Asia and Pan-Pacific countries such as: Singapore (http://www.ecitizen.gov.sg), China, Japan, Korea, Taiwan, India, New Zealand, Australia, etc. E-government projects in Latin American countries have also been reported. Most, however, are of the information dissemination or "one-stop shop" types.

3.5 Digital Government Challenges

Despite similar reliance on Internet technologies, digital government faces some issues and challenges uniquely different from those of e-commerce. Gordon made a clear distinction between e-commerce and e-government in (Gordon, 2002):

> …But e-commerce is not at the heart of e-government. The core task of government is governance, the job of regulating society, not marketing and sales. In modern democracies, responsibility and power for regulation is divided up and shared among the legislative, executive and judicial branches of government. Simplifying somewhat, the legislative is responsible for making policy in the form of laws, the executive for implementing the policy and law enforcement, and the judiciary for resolving legal conflicts. E-government is about moving the work of all of these branches, not just public administration in the narrow sense.

In addition to having different roles in government, digital government applications also face some unique challenges that are different from those of e-commerce and digital libraries.

- Organizational and cultural inertia: Most government entities are not known for their efficiency or willingness to adopt changes. Organizational bureaucracy and lack of clear communication channels or collaboration culture are some of the difficult problems to resolve before any e-government initiatives can become successful. Some (federal, state, and local) government agencies or departments are known for being non-responsive, closed, secretive, arrogant, bureaucratic, and resistant-to-change. Organizational and cultural changes often are more difficult than technological changes.
- Government and legal regulations: Governments at all levels are often faced with numerous laws and regulations intended to make their rights and obligations clear and provide some supervisory and/or balancing functions. Although well-intended, such laws and regulations often inhibit innovation or thinking "out-of-the-box."
- Security and privacy issues: E-government applications on the Internet face the daunting task of protecting the privacy of citizens (and their transactions) in an open (and often not-so-secure) Internet environment. Although e-commerce applications may also stress security and privacy issues (e.g., for credit transactions and customer information), government-provided services have the extra burden of guaranteeing security and privacy for citizens. Many digital government projects are currently under way to explore public key encryption and digital signature issues unique to e-government.
- Disparate and out-dated information infrastructure and systems: Many government departments at all levels often face budget shortfalls for years. As a result, their information infrastructure and systems may be out-of-date. Mainframe computers and applications from the 1970s may constitute a significant part of their computing infrastructure. Some applications may be LAN or Windows-based (technology choice of the 1980s), but most are not Web-enabled or Internet based. Different departments often purchase their own computers and software at different times based on their immediate needs. As a result, disparate legacy,

"stovepipe" systems are created, which prevent departments from sharing information and/or streamlining their businesses.

- Lack of IT funding and personnel: Some government agencies are affluent, but most are not. IT spending often is not a priority (e.g., in light of the more visible or pressing need to put cops on the street, or to purchase additional fire trucks for an under-served community). Furthermore, IT personnel in government often lack resources for training and re-education to update their technical skills. The Internet e-commerce boom (and resulting brain drain) over the past decade also has accelerated the recruitment and retention problem for government IT divisions.

4 Trailblazing a Path towards Knowledge and Transformation

From the above reviews and discussions, it appears true that digital library and digital government have many common threads and yet there are also many unique challenges facing each discipline. It is our belief that regardless of their surface differences, each discipline requires a switch of focus from simple data organization and information access, to the more fundamental process of knowledge creation and sharing. It is well recognized that "knowledge is power," but not data or information (which creates "data/information overload"). We also believe that fundamental "transformation" is required of these institutions to adopt new technologies and the associated processes, instead of relying only on technologies to provide information, communication, and transactions over the Internet (Chen, 2002).

5 A Case Study in Digital Government: Information Sharing and Analysis

5.1 Introduction

In response to the September 11 terrorist attacks, major government efforts to modernize federal law enforcement authorities' intelligence collection and processing capabilities have been initiated. At the state and local levels, crime and police report data are rapidly migrating from paper records to automated records management systems in recent years, making them increasingly accessible.

However, despite the increasing availability of data, many challenges continue to hinder effective use of law enforcement data and knowledge, in turn limiting crime-fighting capabilities of related government agencies. For instance, most local police have database systems used by their own personnel, but lack an efficient manner in which to share information with other agencies (Lingerfelt, 1997) (Pilant, 1996). More importantly, the tools necessary to retrieve, filter, integrate, and intelligently present relevant information have not yet been sufficiently refined. According to senior Justice Department officials quoted on MSNBC, September 26, 2001, there is "justifiable skepticism about the FBI's ability to handle massive amounts of information," and recent anti-terrorism initiatives will create more "data overload" problems.

As part of nationwide, ongoing digital government initiatives, COPLINK is an integrated information and knowledge management environment aimed at meeting some of these challenges. Funded by the National Institute of Justice and NSF, COPLINK has been developed at the University of Arizona's Artificial Intelligence Lab in collaboration with several local, state, and federal law enforcement agencies. The main goal of COPLINK is to develop information and knowledge management systems technologies and methodology appropriate for capturing, accessing, analyzing, visualizing, and sharing law enforcement related information in social and organizational contexts (Chen, Schroeder, et al. 2002).

5.2 Main Components of COPLINK

The COPLINK system consists of two main components: COPLINK Connect and COPLINK Detect. COPLINK Connect is a system designed to allow diverse police departments to share data seamlessly through an easy-to-use interface that integrates different data sources. COPLINK Detect uncovers various types of criminal associations that exist in law enforcement databases.

5.3 COPLINK Connect

The targeted users of COPLINK Connect are law enforcement personnel who are typically not experienced IT users and have pressing, oftentimes mission-critical, information needs. The design of COPLINK Connect was closely guided by user requirements acquired through multi-phase brainstorming sessions, storyboards, mock system demonstrations, focus groups, and more formally structured questionnaires and interviews. We illustrate the functionality of COPLINK Connect in Figure 1. Key design decisions and lessons learned are summarized below.

One-Stop Data Access. Most police data currently is scattered over distributed information sources. To find relevant information, a police officer not only needs to know which data sources offer what sets of data, and how to access them, but also needs to understand each individual source's query language and user interface. He or she then must manually integrate retrieved data. One of the key functions of COPLINK Connect is to provide a one-stop access point for data to alleviate police officers' information and cognitive overload. In its current version, COPLINK Connect supports consolidated access to all major databases in the Tucson Police Department (TPD). The mug shots illustrated in Figures 1a and 1e are incorporated from a separate state-wide mug shot database as is the gang database illustrated in Figure 1b. Incorporating other data sources, including remote ones managed by other organizations, can be easily accomplished. We are currently expanding the COPLINK Connect data sources to include the Arizona State Motor Vehicle database, the Computer-Aided Police Dispatching database, Tucson city court databases, and other sources considered important by TPD officers.

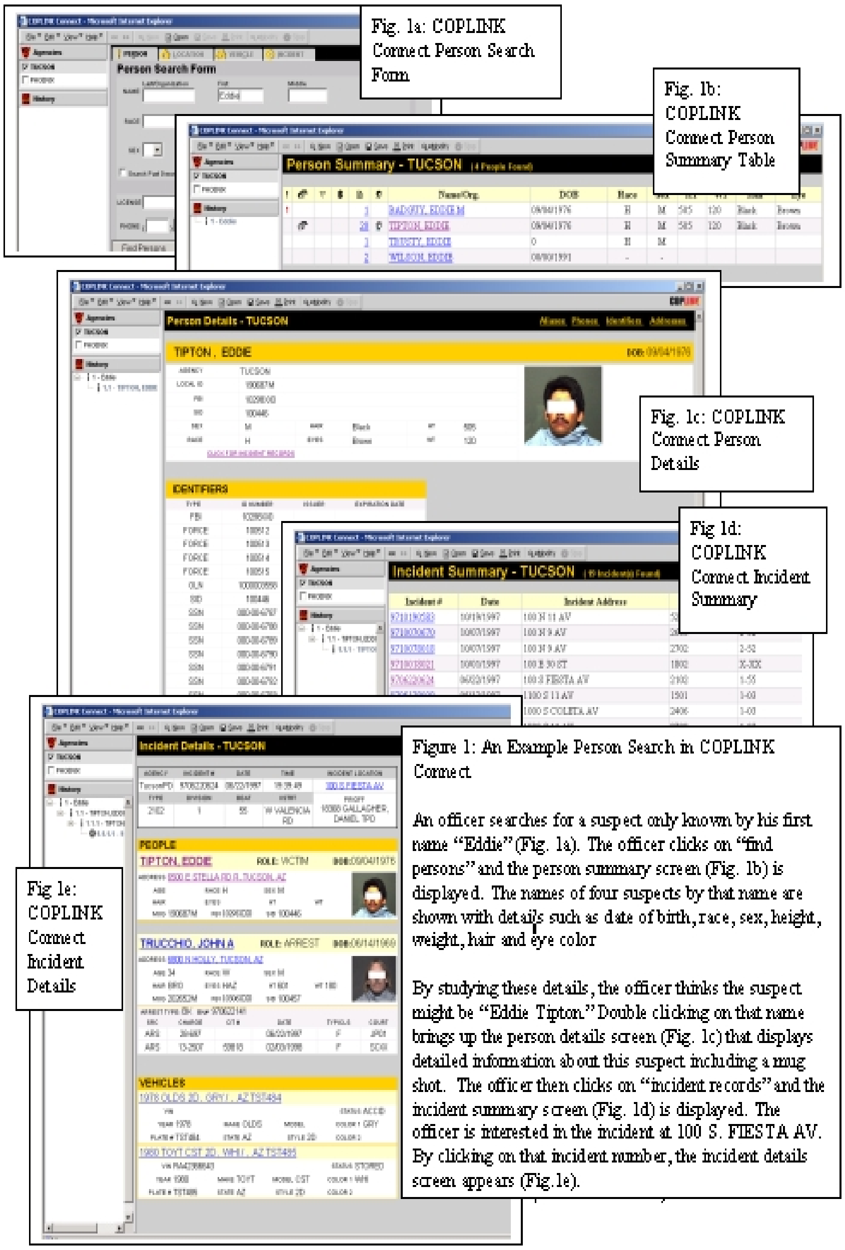


Fig. 1a: COPLINK Connect Person Search Form
Fig. 1b: COPLINK Connect Person Summary Table
Fig. 1c: COPLINK Connect Person Details
Fig 1d: COPLINK Connect Incident Summary
Fig 1e: COPLINK Connect Incident Details
Person Search Form
Person Summary - TUCSON
Person Details - TUCSON
TIPTON , EDDIE
IDENTIFIERS
Incident Summary - TUCSON
Incident Details - TUCSON
PEOPLE
TIPTON, EDDIE
TRUCCHIO, JOHN A
VEHICLES

Figure 1: An Example Person Search in COPLINK Connect

An officer searches for a suspect only known by his first name "Eddie" (Fig. 1a). The officer clicks on "find persons" and the person summary screen (Fig. 1b) is displayed. The names of four suspects by that name are shown with details such as date of birth, race, sex, height, weight, hair and eye color

By studying these details, the officer thinks the suspect might be "Eddie Tipton." Double clicking on that name brings up the person details screen (Fig. 1c) that displays detailed information about this suspect including a mug shot. The officer then clicks on "incident records" and the incident summary screen (Fig. 1d) is displayed. The officer is interested in the incident at 100 S. FIESTA AV. By clicking on that incident number, the incident details screen appears (Fig. 1e).

Search Functionality. As illustrated in Figure 1, four types of searches are made available to the user: *person, vehicle, incident, and location*. These forms are chosen because they cover the primary search tasks that police officers normally perform. Based on the user requirement for simplicity (especially from field officers), we decided that these four types of search can only be performed independently of each other (i.e., the user is not allowed to perform a combined search such as one involving both person and vehicle search terms). A follow-up user study justified such a simplified design.

User Interface. Many user interface design tradeoffs have been made in developing COPLINK Connect.

- *Partial and phonetic-based matching*. Police officers often need to conduct searches based on sketchy and incomplete information. Extensive support for partial and phonetic-based matches are built into COPLINK Connect to facilitate such searches. For instance, in Arizona, police and other law enforcement officers (e.g., border patrol agents) often have to deal with misspelled Spanish names (e.g., "Gweesty"). In COPLINK Connect, such names would match plausible Spanish names (e.g., "Guisti").
- *Search history*. As illustrated in Figure 1d, the end user has access to his or her own search history, making data entry a bit easier and keeping the user appraised of all outstanding tasks. More importantly, the search history mechanism provides important documentation to justify and corroborate related information inquires and subsequent actions in legal proceedings. Furthermore, search logs become an important part of organizational memory by representing training cases for new police officers.

System Architecture. The current version of COPLINK Connect follows a 3-tier system architecture. The user accesses the system through a Web browser. The GUI part of the system is enabled through standard HTML. The middle tier connects the user GUI and the backend databases using Java Servlet and JDBC, and implements the business logic using Java. This system architecture decision was based on careful consideration of the law enforcement domain.

- *Ease of installation and maintenance.* Similar to other government agencies, the IT departments of law enforcement agencies are typically understaffed. The current system architecture eliminates the needs to install or maintain software on end users' local machines.
- *Cost effectiveness*. Earlier versions of COPLINK Connect adopted a proprietary software architecture based on Oracle products, which had the advantage of excellent system performance and the availability of a rich set of integrated development tools. It could, however, incur significant cost for software licensing. On the other hand, the current open architecture based on JDBC-compliant databases (e.g., open-source mySQL and MS SQL Server) can lead to substantial savings.
- *System extensibility*. The current architecture can support access to both remote and local databases, making very limited assumptions regarding information providers. This capability to easily incorporate and make use of additional information sources is important for law enforcement applications due to the frequent need for cross-jurisdictional collaborations in dealing with crimes that are typically not confined to one geographical location.

5.4 COPLINK Detect

COPLINK Detect, targeted at detectives and crime analysts, shares the same incident record information as COPLINK Connect and utilizes the database indexes generated for COPLINK Connect. It, however, has a completely redesigned user interface, and employs a new set of intelligence analysis tools to meet its user needs. Figure 2 shows a sample search session.

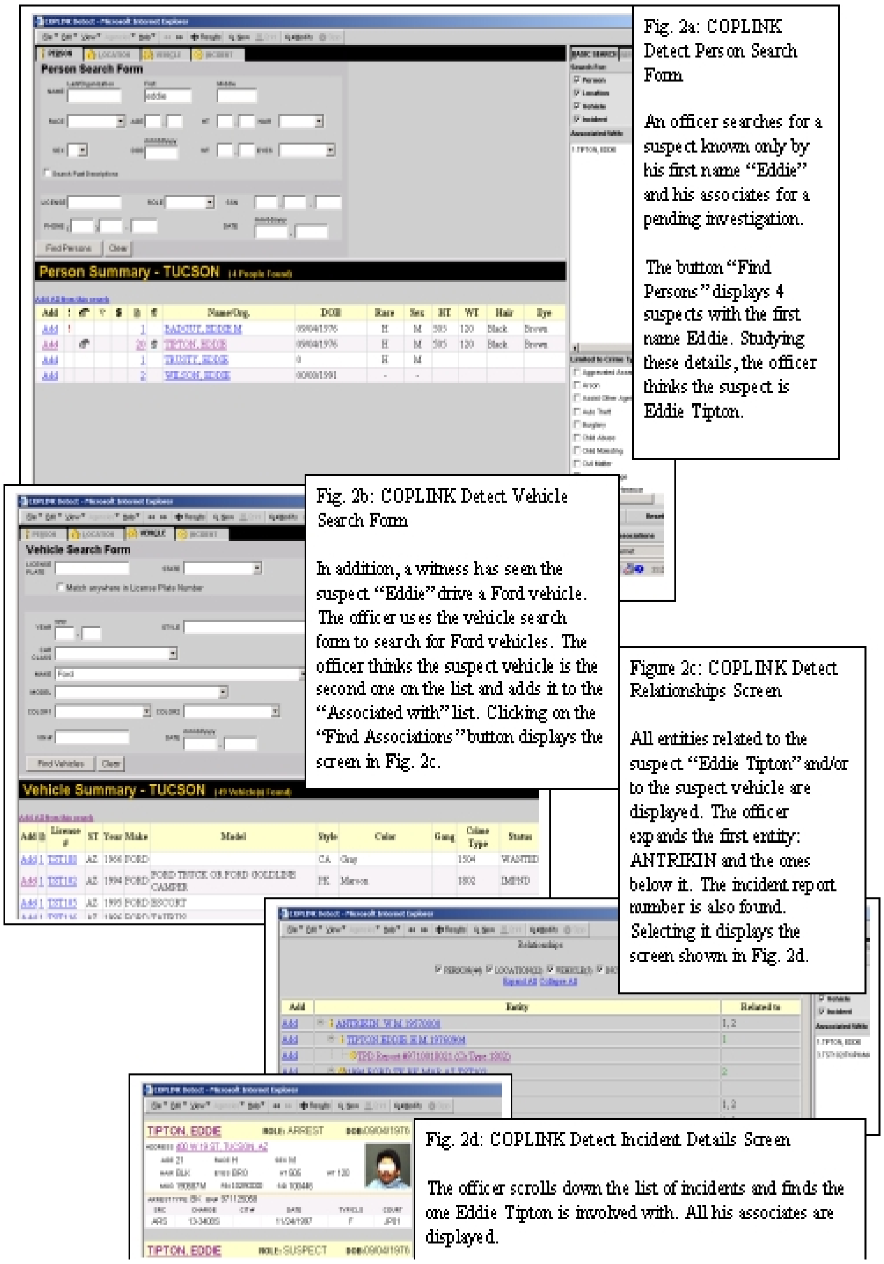
Link and Association Analysis. Much of crime analysis is concerned with creating associations or linkages between various aspects of a crime (similar to association rule mining in traditional data mining research). COPLINK Detect uses a technique called Concept Space (Chen, Schatz, et al., 1996) to identify such associations from existing crime data automatically. COPLINK Detect uses statistical techniques such as co-occurrence analysis and clustering functions to weight relationships between all possible pairs of criminal concepts.

In COPLINK Detect, detailed criminal case reports are the underlying information space and concepts are meaningful criminal elements occurring in each case (Chen, Schroeder, et al., 2002). These case reports contain both structured (e.g., database fields for incidents containing the case number, names of people involved, address, date, etc.) and unstructured data (narratives written by officers commenting on an incident, e.g., "witness1 said he saw suspect1 run away in a white truck"). Using COPLINK Detect, investigators can link known objects (e.g., a given suspect) to other related objects (e.g., people and vehicles related to this suspect) that might contain useful information for further investigation. At present, COPLINK Detect has access to a collection of 1.5 million TPD case reports, spanning a time frame from 1986 to 1999. The system is capable of automatically identifying relationships among Person, Organization, Location, Vehicle, and Incident/Crime type.

5.5 Summary of COPLINK User Studies

Several field user studies have been conducted to evaluate the COPLINK system. Detailed reports are available in (Hauck & Chen, 1999). We summarize two studies below.

A group of 52 law enforcement personnel from TPD representing a number of different job classifications and backgrounds were recruited to participate in a study to evaluate COPLINK Connect. Both interview data and survey data analyses support a conclusion that use of COPLINK Connect provided performance superior to that of the legacy police Records Management System (RMS). In addition to the statistical data, these findings were supported by qualitative data collected from participant interviews. Comments collected from interviews indicate that COPLINK Connect was rated higher than RMS in terms of interface design, performance, and functionality. Participants indicated that the quality and quantity of information from COPLINK Connect surpassed that of RMS. During the time of user evaluation, use of COPLINK Connect had led to the investigation of cases that otherwise might not have been picked up, as well as aided in making multiple arrests.



Person Search Form

Person Summary - TUCSON (4 People Found)

Add	!			$			Name/Org.	DOB	Race	Sex	HT	WT	Hair	Eye
Add	!				1		BAIDUT, EDDIE M	09/04/1976	H	M	505	120	Black	Brown
Add					20		TIPTON, EDDIE	09/04/1976	H	M	505	120	Black	Brown
Add					1		TRUSTY, EDDIE	0	H	M				
Add					2		WILSON, EDDIE	00/00/1991	-	-				

Vehicle Search Form

Vehicle Summary - TUCSON (49 Vehicle(s) Found)

Add	ID	License #	ST	Year	Make	Model	Style	Color	Gang	Crime Type	Status
Add	1	TST180	AZ	1966	FORD		CA	Gray		1504	WANTED
Add	1	TST182	AZ	1994	FORD	FORD TRUCK OR FORD GOLDLINE CAMPER	PK	Maroon		1902	IMPND
Add	1	TST183	AZ	1995	FORD	ESCORT					

Fig. 2a: COPLINK Detect Person Search Form

An officer searches for a suspect known only by his first name "Eddie" and his associates for a pending investigation.

The button "Find Persons" displays 4 suspects with the first name Eddie. Studying these details, the officer thinks the suspect is Eddie Tipton.

Fig. 2b: COPLINK Detect Vehicle Search Form

In addition, a witness has seen the suspect "Eddie" drive a Ford vehicle. The officer uses the vehicle search form to search for Ford vehicles. The officer thinks the suspect vehicle is the second one on the list and adds it to the "Associated with" list. Clicking on the "Find Associations" button displays the screen in Fig. 2c.

Figure 2c: COPLINK Detect Relationships Screen

All entities related to the suspect "Eddie Tipton" and/or to the suspect vehicle are displayed. The officer expands the first entity: ANTRIKIN and the ones below it. The incident report number is also found. Selecting it displays the screen shown in Fig. 2d.

Fig. 2d: COPLINK Detect Incident Details Screen

The officer scrolls down the list of incidents and finds the one Eddie Tipton is involved with. All his associates are displayed.

A pilot user study to evaluate COPLINK Detect was conducted with 33 participants, including 7 crime analysts and 17 detectives from TPD. Data was collected by three methods: verbal reports, "search notes" for searches they performed, and electronic transaction logs. Participants indicated that (a) COPLINK Detect required minimum training (8 participants started to use the system effectively even without any training), (b) the system improved case closure and crime solvability through uncovering critical associations, and (c) the system was very responsive (all the associations and related incident reports were identified in less than one minute).

5.6 Knowledge and Transformation

Starting in the Spring of 2001, COPLINK Connect was formally deployed at TPD. By December, 2001, all TPD law enforcement personnel, including about 600 police officers and 200 detectives, were using COPLINK Connect and Detect. The deployment of COPLINK in several law enforcement agencies in the Phoenix area is under development. An Arizona state-wide COPLINK system is planned for 2003. Agencies from other states also have shown strong interest in using COPLINK.

Developed to facilitate federal, state, and local law enforcement information sharing and knowledge management, COPLINK could serve as a model for the next-generation information systems aimed at improving the government's crime fighting capabilities and facilitating its homeland security effort. Several research directions are under way to extract intelligence and knowledge from criminal data.

We are currently working on a new module called COPLINK Collaboration which will enable sharing of crime data and information search experience among law enforcement team members (Atabakhsh, et al., 2002). One of the most intriguing aspects of developing such a collaborative system in law enforcement concerns information privacy, security, and the legal ramifications of having to keep track of information search logs for an extended period of time. COPLINK Collaboration will also include a wireless access and alerting component using cell phones, wireless laptops, and PDAs, to meet the needs of mobile law enforcement officers.

We are also experimenting with several crime visualization techniques such as using hyperbolic trees to better present identified associations in COPLINK Detect. A user can search all entities having a relationship with a given search term and view the relationships in the form of a hyperbolic tree as well as in a hierarchical tree structure. Lastly, we have also begun to develop promising techniques in the areas of automatic deception detection (Wang & Chen, 2002), and criminal network analysis (Xu & Chen, 2002) based on a visual data mining approach.

We believe the COPLINK approach to information sharing, crime data mining, and knowledge management could have a significant impact on the crime investigation and intelligence analysis practices of law enforcement personnel at the local, state, and federal levels. By developing inter-operable systems, better intelligence information could be shared among different local and federal agencies. By supporting crime relationship identification and visualization, detectives and analysts could reduce their case workloads and increase case closure. Visual data mining with structured network analysis, clustering, and classification techniques for different crime types could help track criminal (e.g., narcotic, gang, terrorist) networks and suggest investigative strategies. Such an IT and Internet-enhanced information sharing and knowledge discovery approach would be crucial to the success in fighting the terrorists after the September 11 tragedy.

In addition to supporting various law enforcement agencies, we believe the COPLINK approach to information integration and analysis could also help streamline and improve federal, state, and local civil services including: corrections, litigations, social services, transportations, voting, and so on. For more information, please visit the COPLINK project web site at http://ai.bpa.arizona.edu/COPLINK.

The Future

With common threads and unique challenges facing digital library, digital government, and e-commerce, we foresee many active and high-impact research opportunities for researchers in information science, library science, computer science, public policy, and management information systems. Digital library, digital government, and e-commerce researchers are well positioned to become the “agents of transformation” for the new Net of the 21st century.

Acknowledgements. This project has primarily been funded by:

- NSF, Digital Government Program, “COPLINK Center: Information and Knowledge Management for Law Enforcement,” EAI-9983304, July, 2000-June, 2003.
- National Institute of Justice, “COPLINK: Database Integration and Access for a Law Enforcement Intranet,” July 1997-January 2000.

We would also like to thank the Digital Equipment Corporation External Technology Grants Program (agreement #US-1998004) for its award of an equipment grant of a DEC Alpha Server for the COPLINK Project. We would like to thank the Tucson Police Department and the Phoenix Police Department for their significant contributions to the project.

References

1. J. L. Ambrite, Y. Arens, E. Hovy, A. Philpot, L. Gravano, V. Hatzivassiloglou, and J. Klavans, “Simplying Data Access: The Energy Data Collection Project,” *IEEE Computer*, 34(2), Pages 32-38, February 2001.
2. H. Atabakhsh, J. Schroeder, H. Chen, M. Chau, J. Xu, J. Zhang, and H. Bi, “COPLINK Knowledge Management for Law Enforcement: Text Analysis, Visualization and Collaboration,” *National Conference on Digital Government*, May 21-23, Los Angeles, CA, 2001.
3. V. Cerf, “Digital Government and the Internet,” *National Conference on Digital Library Research*, Los Angeles, CA, May 20, 2002.
4. H. Chen, B. R. Schatz, T. D. Ng, J. Martinez, A. Kirchhoff, and C. Lin, “A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Special Section on Digital Libraries: Representation and Retrieval, 18(8), 771-782, 1996.
5. H. Chen, “Knowledge Management Systems: A Text Mining Perspective,” University of Arizona, Tucson, Arizona, 2002.
6. H. Chen, J. Schroeder, R. V. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, C. Boarman, K. Rasmussen, and A. W. Clements, “COPLINK Connect: Information and Knowledge Management for Law Enforcement,” *Decision Support Systems,* forthcoming, 2002.
7. A. K. Elmagarmid & W. J. McIver Jr., “The Ongoing March Toward Digital Government,” *IEEE Computer*, 34(2), 32-38, February 2001.
8. ERCIM, ERCIM News, Special Theme: E-government, European Research Consortium for Information and Mathematics, Number 48, January 2002.
9. C. Fiorina, “The World Stands at the Threshold of a Digital Renaissance,” Aspen Summit 2000: Cyberspace and the American Dream VII, Aspen, CO, August 25, 2000.

10. T. F. Gordon, “Introduction to E-Government,” ERCIM News, Special Theme: E-government, *European Research Consortium for Information and Mathematics*, Number 48, 12-13, January 2002.
11. R. V. Hauck & H. Chen, “COPLINK: A Case of Intelligent Analysis and Knowledge Management,” *Proceedings of the 20th Annual International Conference on Information Systems ’99*, 15-28, 1999.
12. R. V. Hauck, H. Atabakhsh, P. Ongvasith, H. Gupta, and H. Chen, “Using Coplink to Analyze Criminal-Justice Data,” *IEEE Computer,* 35(3), 30-37, 2002.
13. J. Lingerfelt, “Technology As a Force Multiplier,” *Proceedings of the Conference in Technology Community Policing*, National Law Enforcement and Corrections Technology Center, 1997.
14. National Research Council, Computer Science and Telecommunications Board, “Information Technology Research, Innovation, and E-Government,” National Academy Press, Washington, DC, 2002.
15. National Science Foundation, Program Announcement – Digital Government, NSF98-121, National Science Foundation, 1998.
16. L. Pliant, “High-technology Solutions,” *The Police Chief*, 5(38), 38-51, 1996.
17. B. R. Schatz & H. Chen, “Building Large-scale Digital Libraries,” *IEEE COMPUTER*, 29(5), 22-27, 1996.
18. B. R. Schatz & H. Chen, “Digital Libraries: Technological Advancements and Social Impacts,” *IEEE COMPUTER*, 31(2), 45-50, 1999.
19. J. Taschek, “Egov Challenges Tech,” *eWEEK*, 19(14), April 8, 2002.
20. G. Wang, H. Chen, and H. Atabakhsh, “Automatically Detecting Deceptive Criminal Identities,” *Communications of the ACM*, conditionally accepted, 2002.
21. J. Xu & H. Chen, “Criminal Social Network Analysis: A Data Mining Approach,” *Communications of the ACM*, under review, 2002.

Progress on Educational Digital Libraries: Current Developments in the National Science Foundation's (NSF) National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program

Jane C. Prey and Lee L. Zia[1]

Division of Undergraduate Education
National Science Foundation
Arlington, VA 22230

1 Background

Technological advances in computing and communication continue to challenge the "why, where, and when" of the educational enterprise. Policymakers and advocates from organizations such as EDUCAUSE[2], the Coalition for Networked Information (CNI)[3], the European Schoolnet partnership[4], and UNESCO[5] have long argued for the potential that information technology has to transform education by improving access – both spatially ("anyplace") and temporally ("anytime"), by redefining the nature of the classroom and learning, and in some cases by changing cost structures. In recent years various consortia of higher education institutions have formed to explore international collaborations in distance learning, for example: Universitas 21 (http://www.universitas.edu.au/) and the Global University Alliance (www.gua.com) both of which involve institutions from Asia, Europe, and the United States. Many individual institutions have also embarked on distance education ventures and efforts in the K-12 sector are growing as well (see http://chronicle.com/indepth/distance/players.htm#virtual). Development of online course management systems continues to progress on the commercial front, e.g. WebCT (www.webct.com) and Blackboard (www.blackboard.com), and more recently an open-source effort called the Open Knowledge Initiative (http://web.mit.edu/oki/) has begun.

Also of particular interest are the opportunities digital libraries have to provide critical infrastructure in the form of content and services that can enable transformations in learning. Although the precise definition of digital library remains in flux and may eventually go the way of "horseless carriage", the following quote from Collier offers a reasonable characterization:

> *digital library:* "A managed environment of multimedia materials in digital form, designed for the benefit of its user population, structured to facilitate

[1] All views expressed in this article are solely those of the authors and do not represent an official NSF policy statement.

[2] Formed in 1998 from the merger of Educom and CAUSE (see http://www.educause.edu/).

[3] See http://www.cni.org.

[4] See http://www.eun.org/eun.org2/eun/en/About_eschoolnet/entry_page.cfm?id_area=101.

[5] See http://www.unesco.org/education/index.shtml.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 53-66, 2002.

> access to its contents, and equipped with aids to navigate the global network ... with users and holdings totally distributed, but managed as a coherent whole."
>
> --Mel Collier, International Symposium on Research,
> Development, and Practice in Digital Libraries 1997

To investigate the research and development challenges associated with this idea, the National Science Foundation (NSF) – along with the National Aeronautics and Space Administration (NASA) and the Department of Defense Advanced Research Projects Agency (ARPA) – initiated the Digital Libraries Initiative (DLI) in 1994. Subsequently, the Digital Libraries Initiative – Phase 2 (DLI-2) began in 1998 with several additional federal agency partners and continues its support of digital library research (see http://www.dli2.nsf.gov).

In late 1995 an internal concept paper for the NSF Division of Undergraduate Education outlined the capabilities and issues that digital libraries pose for improving science, technology, engineering, and mathematics education. The idea was explored and developed further through a series of workshops and planning meetings [1-6]. Papers followed that considered evaluation and dissemination issues [7] and the organization and architecture of a digital library [8]. In 1998 as a precursor to the current program, the *Special Emphasis: Planning Testbeds and Applications for Undergraduate Education* program was begun under the auspices of the multi-agency DLI-2 program. Although most of these initial projects focused primarily on collection development, others began to explore organizational and managerial functions of a digital library; and they have helped to lay the foundation for the current program.

2 Vision

NSF formally launched the National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) program in fiscal year (FY) 2000. The program resides formally in the Division of Undergraduate Education, however its ultimate user base is comprehensive. Indeed, through this program NSF seeks to stimulate and sustain continual improvements in the quality of science, technology, engineering, and mathematics (STEM) education at all levels. The resulting digital library targets a broad audience – pre-K to 12, undergraduate, graduate, and life-long learners – in formal and informal settings, and individual and collaborative modes. NSDL will be a virtual facility, and as a "national treasure" it will enable seamless access to a rich array of interactive learning resources and learning environments, and services not bound by place or time. Furthermore, it will be valued for its comprehensiveness, authenticity, and reliability.

The fundamental operational mode of the library consists of an interaction of users, content, and tools via the network that connects these three elements. Users comprise students in formal settings, educators, and life-long learners. Content describes a rich and diverse set of materials in multiple media, including structured learning materials; large real-time or archived data sets; audio, video, images, and animations; "born" digital learning objects (e.g. simulations and applets); interactive (virtual or remote) laboratories; and primary source material. Tools encompass a wide range of functionality to support users including searching, referring, validation, integration, annotation, creation, customization, sharing, publishing, notification, and collaboration. Ultimately NSDL looks to support “learning communities”, “customizable collections”, and “application services”. For more details on the vision and goals for the NSDL program see [9].

3 Current Status

The NSDL program (http://www.ehr.nsf.gov/ehr/due/programs/nsdl/) held its third formal funding cycle during fiscal year 2002 with a proposal deadline in mid-April 2002. Proposals were accepted in three tracks, *Collections*, *Services*, and *Targeted Research*. One hundred fifty-six proposals were received in response to the solicitation, seeking approximately $92 M in total funding. Forty-one new projects were recently announced with a cumulative budget of approximately $25M. These include 25 in the *Collections* track, 10 in the *Services* track, and 6 in the *Targeted Research* track. Several NSF sister directorates to EHR, the Directorate for Geosciences (GEO), the Directorate for Mathematical and Physical Sciences (MPS), and the Directorate for Biological Sciences are providing significant co-funding on nearly ten projects. These jointly funded projects illustrate the NSDL program’s facilitation of the integration of research and education that is an important strategic objective of NSF. For information about projects from the previous two funding cycles, see [10 and 11].

Current projects feature good coverage in the subject domains of the life sciences, geosciences, various areas of engineering, the mathematical sciences, and several cross-disciplinary areas. New projects from FY 2002 expand this coverage into areas of chemistry and physics. New also are projects that focus on collections of video materials, as well as continued work focusing on collections and services targeting the pre-K to12 educational enterprise. Indeed nearly half of all existing and new projects in the program have explicit linkages with the pre-K to12 sector or strong secondary efforts in that area. All projects feature interdisciplinary teams of principal investigators representing a variety of backgrounds including expertise in the library and information sciences, computer science, digital library research, disciplinary content, and instructional design.

The past year of the program has also seen the emergence of an information architecture framework for the library resulting from the joint efforts of the Core Integration project and many of the Collections, Services, and Targeted Research projects. An NSDL community produced white paper, “Pathways to Progress,” [12] informed much of this work and has also stimulated the creation of a community governance structure that is jointly establishing guidelines, practices, and policies for participation in the larger NSDL building effort. Complete information on technical and organizational progress including links to the community workspaces may be found at the NSDL Communications Portal, see http://comm.nsdlib.org. All workspaces are open to the public and interested organizations and individuals are encouraged to learn more about NSDL and join in its development.

The Appendix lists all new awards displaying the title, the grantee institution, the name of the Principal Investigator (PI), and the official NSF award number. Short descriptions of the projects are also included. Full abstracts are available from the

Awards Section at the NSDL Program site http://www.ehr.nsf.gov/ehr/due/programs/nsdl/. Projects with shared titles are formal collaborations and are grouped together.

4 Future Directions

The NSDL program expects to have another funding cycle in fiscal year 2003 with an anticipated deadline for proposals in mid-April, 2003. Optional letters of intent would be due in mid-March, 2003. Support for *Collections***,** *Services***,** and *Targeted Research* will be available again. The new solicitation should be available online in early January, see http://www.ehr.nsf.gov/ehr/due/programs/nsdl/.

Looking beyond the current programmatic emphases two interesting opportunities bear examination. The first recognizes the international dimension of efforts to improve education, particularly in STEM areas of inquiry. Though the languages used to describe a scientific phenomenon may be varied, the underlying phenomenon itself is common. Thus there is great opportunity to share high quality educational resources particularly those of a digital nature across international boundaries. As globalization continues, government agencies should continue to seek ways to define areas of common interest and jointly support projects to exploit the power of digital library technology in the service of education.

Secondly, there should be serious consideration of how to expand the scope of the NSDL program beyond its science, technology, engineering, and mathematics content domains. Notable digital library projects in other domains do exist, for example the Perseus Project in the Humanities (see http://www.perseus.tufts.edu/), but there has not yet emerged a strong advocacy for programmed funding of digital library projects in “non-STEM” areas. However, if efforts are launched in this arena, there should be coordination with existing programs such as DLI-2 and NSDL and other agencies such as the Institute for Museum and Library Services (www.imls.gov). This will enable advances in research and development, and progress towards the implementation of distributed organizational and technological infrastructure to be leveraged. A recent PITAC report does address some of these possibilities, Digital Libraries: Universal Access to Human Knowledge (see http://www.itrd.gov/pubs/pitac/pitac-dl-9feb01.pdf).

References

1. Information Technology: Its Impact on Undergraduate Education in Science, Mathematics, Engineering, and Technology (NSF 98-82), April 18-20, 1996. http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf9882.
2. Developing a Digital National Library for Undergraduate Science, Mathematics, Engineering, and Technology Education, NRC workshop, August 7-8, 1997. http://www.nap.edu/catalog/5952.html
3. Report of the SMETE Library Workshop (NSF 99-112), July 21-23, 1998. http://www.dlib.org/smete/public/report.html

4. Serving the Needs of Pre-College Science and Mathematics Education: Impact of a Digital National Library on Teacher Education and Practice, NRC workshop, September 24-25, 1998. http://www.nap.edu/books/NI000781/html/
5. Digital Libraries and Education Working Meeting, January 4-6, 1999. http://www.dli2.nsf.gov/dljanmtg.pdf
6. Portal to the Future: A Digital Library for Earth System Education, workshop report, August 8-11, 1999. http://www.dlese.org/documents/reports/panelreports/reports.html
7. Mogk, David W. and Zia, Lee L. "Addressing Opportunities and Challenges in Evaluation and Dissemination through Creation of a National Library for Undergraduate Science Education." Invited Symposium in Proceedings of the 31st Annual Meeting of the Geoscience Information Society, October 28-31, 1996, Denver, CO (1996). Available at http://gdl.ou.edu/rp1.html
8. Wattenberg, Frank. A National Digital Library for Science, Mathematics, Engineering, and Technology Education, D-Lib Magazine, October 1998. http://www.dlib.org/dlib/october98/wattenberg/10wattenberg.html
9. Zia, Lee L. Growing a National Learning Environments and Resources Network for Science, Mathematics, Engineering, and Technology Education: Current Issues and Opportunities for the NSDL Program. http://www.dlib.org/dlib/march01/zia/03zia.html
10. Zia, Lee L. The NSF National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) Program: A Progress Report. http://www.dlib.org/dlib/october00/zia/10zia.html
11. Zia, Lee L. The NSF National Science, Technology, Engineering, and Mathematics Education Digital Library (NSDL) Program: New Projects and a Progress Report. Available at http://www.dlib.org/dlib/november01/zia/11zia.html
12. Pathways to Progress: Challenges and Goals of Building the Social and Technical Infrastructure of the NSDL, http://doclib.comm.nsdlib.org/PathwaysToProgress.pdf.

Appendix

Collections Track

Projects are expected to aggregate and manage a subset of the library's content within a coherent theme or specialty.

BioSciEd Net (BEN) Collaborative: Cycle 2. Institution: American Association For Advancement Science. PI: Yolanda George. DUE-0226185. The Biosci Ed Net (BEN) Collaborative (http://www.biosciednet.org/portal/) comprises twelve professional societies and coalitions for biology education with an interest in transforming the teaching and learning of undergraduate biology. Six collections from BEN partners are currently featured, including the American Physiological Society (APS), the Ecological Society of America (ESA), the American Society for Microbiology (ASM), the American Society for Biochemistry and Molecular Biology (ASBMB), the Human Anatomy and Physiology Society (HAPS), and Science's Signal Transduction Knowledge Environment (STKE).

www.eSkeletons.org: An Interactive Digital Library of Human and Primate Anatomy. Institution: University of Texas at Austin. PI: John Kappelman. DUE-0226040. This project is expanding the range of content at the www.eSkeletons.org, the degree of learner interactivity with the materials, and the amount of interaction among users.

New laser scanning equipment and improvements in high resolution X-ray computed tomography technologies allow the inclusion of species of a much smaller body size than before, with the completion of the scans accomplished on a much faster timetable. The collection provides students with a more complete understanding of the range of primate diversity and facilitates a great diversity of lab exercises.

Ceph School: A Pedagogic Portal for Teaching Biological Principles with Cephalopod Molluscs. Institution: University of Texas Medical Branch at Galveston. PI: Phillip Lee. DUE-0226334. Using cephalopods as model organisms, Ceph School strives to have students understand basic principles in biology, observe the methodology of scientific research and become familiar with cephalopods, and also provide student and teacher-specific support. Collection modules are enhanced with web cameras, videos, and links to additional data. From remote sites students observe living animals in real time, use interactive maps to explore their geographical distribution and habitats, learn about their anatomy, physiology and behavior, search appropriate bibliographies, locate world experts, and participate in the scientific process.

Collaborative Research: Health Education Assets Library. Institutions: University of Oklahoma Health Sciences Center, University of Utah, and University of California-Los Angeles. PI: Chris Candler, DUE-0226102; PI: Sharon Dennis, DUE-0226132; and PI: Sebastian Uijtdehaage, DUE-0226314. The Health Education Assets Library (HEAL) supports K-12, undergraduate, and professional health science education as well as patient and consumer education. HEAL represents a collaboration of three institutions: the University of Utah, UCLA, and the University of Oklahoma Health Sciences Center, together with the National Library of Medicine and the Association of American Medical colleges. These partnerships have facilitated the formation of a rapidly growing network of institutions that seek either to contribute teaching resources directly to the HEAL collection or enable seamless bridging to their collections.

Advanced Placement Digital Library for Biology, Physics and Chemistry. Institution: William Marsh Rice University. PI: Siva Kumari. DUE-0226317. Rice University, in collaboration with the College Board, is creating the Advanced Placement Digital Library (APDL), an online digital library for high school Advanced Placement (AP) students and teachers of Biology, Physics and Chemistry (BPC).

Marine Mammal Commission Digital Library of International Environmental and Ecosystem Policy Documents. Institution: Ohio State University Research Foundation. PI: Paul Berkman. DUE-0226195. This project is developing a sustainable single-source digital collection of international environmental and ecosystem policy documents that facilitates knowledge discovery, supports a “rich learning environment” and benefits researchers, teachers, students, diplomats and decision-makers throughout society from global to local levels. Materials for this collection are gleaned from the Marine Mammal Commission’s five-volume Compendium of Selected Treaties, International Agreements, and Other Relevant Documents on Marine Resources, Wildlife, and the Environment.

Collaborative Research: TeachEngineering - Hands-on Engineering. Institutions: Tufts University, University of Oklahoma Norman Campus, University of Colorado at Boulder. PI: Martha Cyr, DUE-0226191; PI: Michael Mooney, DUE-0226236; and PI: Jackie Sullivan, DUE-0226322. This project builds on extensive K-12 engineering curriculum developments funded by the NSF GK-12 program with several engineering colleges collaborating to create an on-line digital library of engineering resources (the TeachEngineering Collection) for use by K-12 teachers and engineering college faculty conducting outreach in their communities. Lead institutions are the University of Colorado, the University of Oklahoma, and Tufts University, and each is partnered with numerous local school districts to promote engineering as a vehicle for math and science integration.

A Digital Library Collection for Computer Vision Education. Institution: Swarthmore College. PI: Bruce Maxwell. DUE-0226273. This project is gathering high-quality material into a comprehensive digital library collection for computer vision education. The resource links to assignments at a variety of institutions and hosts a set of vetted assignments, complete with data sets and solutions. In addition, it contains educational resources such as lecture notes, links to other computer vision courses, and reviews of textbooks, software, and hardware.

Advanced Technology Environmental Education Library (ATEEL). Institution: Eastern Iowa Community College District. PI: Ellen Kabat Lensch. DUE-0226116. ATEEL serves the environmental information needs of environmental technology students, educators, and technicians as well as individuals with an interest in or a knowledge requirement of environmental issues. Project partners include the Advanced Technology Environmental Education Center (ATEEC), Massachusetts Institute of Technology (MIT), Partnership for Environmental Technology Education (PETE), and the Davenport Public Library.

Harvard-Smithsonian Digital Video Library. Institution: Smithsonian Institution Astrophysical Observatory. PI: Matthew Schneps. DUE-0226354. The Harvard-Smithsonian Center for Astrophysics is assembling and managing an extensive collection of STEM digital video materials for education including programs such as “A Private Universe” and professional development materials created for the Annenberg/Corporation for Public Broadcasting project. Materials span a variety of topics and formats and include high-quality, case-study footage showing teaching in action, rare and difficult to create materials documenting children’s ideas in science and mathematics, interviews with internationally prominent researchers in STEM learning, and computer animations and other costly visualizations of STEM concepts.

The Moving Image Gateway. Institution: Rutgers University New Brunswick. PI: Grace Agnew. DUE-0226140. The Moving Image Gateway is developing a web portal for moving images that combines an archives directory database with a union catalog to provide access to the world’s moving image collections. Initially archives of moving images in the Library of Congress, Cable News Network, National Geographic Television, the National Library of Medicine, the Oregon Health Sciences University, ResearchChannel, and the Smithsonian Institution are being made available to STEM students, educators, and researchers, as well as the general public.

Teachers'Domain - Physical Science and Engineering. Institution: WGBH Educational Foundation. PI: Michele Korf. DUE-0226184. The Teachers Domain Digital Library is harnessing the extensive broadcast, video, and interactive programming resources of WGBH to support standards-based teaching and learning at the K-12 level. Initially focused on the life sciences, the collection is now expanding its coverage to include the physical sciences and engineering. Through a searchable, web-based repository of contextualized multimedia units teachers are easily accessing materials for their own professional development as well as to enrich classroom activities for students.

Viewing the Future: Aligning Internet2 Video to K-12 Curriculum. Institution: Merit Network, Inc.. PI: Marcia Mardis. DUE-0226323. The University of Washington and the ResearchChannel are using high quality Internet2 video resources to build a collection of STEM materials for the K-12 community. These resources are being aligned to state and national curriculum standards, including relevant classroom assessments.

An Active Object-Based Digital Library for Microeconomics Education. Institution: University of Arizona. PI: James Cox. DUE-0226344. Using an Open Archives Initiative (OAI)-compliant approach, this project is developing an extensible and scalable collection of Microeconomics related content that incorporates experimental software and automated e-commerce agents that simulate markets and enable the exploration of intelligent trading systems.

Earth Exploration Toolbook: A Collection of Examples of Educational Uses of Earth System Science Tools, Datasets and Resources. Institution: TERC Inc. PI: Tamara Ledley. DUE-0226199. The Earth Exploration Toolbook (EET) guides educators at both the pre-college and college levels on how to use, in an educational context, various Earth system tools and datasets developed and archived by and for scientists. Examples provide educators experience with and in-depth knowledge of these resources to be able to use them in other contexts, and to help their students use them to explore and investigate issues in Earth system science.

Linking Pedogogy, Resources and Community Interaction to Support Entry-Level Undergraduate Geoscience Courses. Institution: Carleton College. PI: Cathryn Manduca. DUE-0226243. This project targets faculty teaching entry-level geoscience to college students and is exploring ways that the NSDL can catalyze improvement in undergraduate teaching and learning. The collection contains the full suite of resources needed to support faculty teaching at the entry level including teaching resources (e.g., visualizations; field, lab, and classroom activities; problem sets), information on effective teaching methods, and examples of successful teaching in the geosciences.

The Journal of Chemical Education Digital Library. Institution: University of Wisconsin-Madison. PI: John Moore. DUE-0226244. The Journal of Chemical Education (JCE) is creating the JCE Digital Library with four new collections joining the considerable digital material already available at JCE Online. DigiDemos comprises digitized text, graphics, sound, and video of chemical demonstrations. Computer Algebra Systems contains documents for Mathcad, Mathematica, Maple, or

MATLAB to help students learn mathematically intensive aspects of chemistry. JCE WebWare delivers animations, simulations, calculations, and other pedagogically useful items to promote discussion and interaction among students, and to provide new insights through graphic and other non-traditional means. Resources for Student Assessment include homework, quiz, and examination questions with anytime anywhere, feedback and tutoring based on student responses, and new approaches to student assessment.

Collaborative Research: To Enhance the Depth, Breadth, and Quality of the Collections of the Digital Library of Earth System Education (DLESE). Institutions: American Geological Institute, Dartmouth College, Foothill College, and Columbia University. PI: Sharon Tahirkheli, DUE-0226196; PI: Barbara DeFelice, DUE-0226233; PI: Christopher DiLeonardo, DUE-0226289; and PI: Kim Kastens, DUE-0226292. This project is improving the breadth, depth and quality of the Digital Library for Earth System Education (DLESE) through a set of four integrated tasks. The first is a systematic comparison of the scope and balance of the existing resources with those desired by the geosciences community. The second and third focus on filling identified gaps or thin spots in the collection and cataloging of resources. Finally, the fourth continues development and implementation of a Community Review System

Collaborative Research: A Digital Library Collection for Visually Exploring United States Demographic and Social Change. Institutions: CUNY Queens College and University of California-Los Angeles. PI: Andrew Beveridge, DUE-0226279 and PI: David Halle, DUE-0226295. This collaborative project between CUNY and UCLA is developing a collection of web-based materials that depict and explore growth and social change in the United States based on US census data at the county, tract, and city levels, stretching back in some cases to the late 1700s. It is accessible to students from elementary through postgraduate school, library users, the media, and all others interested in relating a variety of demographic and other data to one another. In addition users may visualize data in the form of maps and charts, download data for further analysis, and relate data to specific issues in the social sciences.

ComPADRE: Communities for Physics and Astronomy Digital Resources in Education. Institution: American Association of Physics Teachers. PI: Bruce Mason, DUE-0226129. The American Association of Physics Teachers (AAPT), the American Physical Society (APS), the American Institute of Physics/Society of Physics Students (AIP/SPS), and the American Astronomical Society (AAS) are creating an interconnected set of digital collections of educational materials and providing specific learning environments accessible to learners and teachers from elementary school through graduate school. ComPADRE's initial collections include resources for introductory astronomy, quantum physics, pre-college physical science teachers, undergraduate majors or prospective majors in physics and astronomy, and informal science education.

Collection Building and Capacity Development for K-12 Federally-Produced Mathematics and Science Education Digital Resources. Institution: Ohio State University Research Foundation. PI: Kimberly Roempler. DUE-0226228. The Eisenhower National Clearinghouse is aggregating science and mathematics education resources developed through Federal funding, to make them easily and meaningfully accessible by teachers, parents, and students. It is also laying the groundwork for ensuring that future resource and materials development adheres to standard metadata schema and tagging practices, on which efficient and relevant search, navigation, and access to content rests.

Kinematic Models for Design Digital Library (K-MODDL). Institution: Cornell University - Endowed. PI: John Saylor. DUE-0226238. A team of faculty and librarians are aggregating educational materials associated with the 220 late 19th-century model machine elements designed for research and teaching by the founder of modern kinematics, Franz Reuleaux (1829-1905). Resources of the collection include still and navigable moving images of these kinematic teaching models; systematic descriptions, and historical and contemporary documents related to the collection of the mechanisms; computer simulations of mathematical relationships associated with the movements of the mechanisms, and sample teaching modules that employ the models and simulations in the classroom at the undergraduate, secondary and middle school levels.

Second Generation Digital Mathematics Resources with Innovative Content for Metadata Harvesting and Courseware Development. Institution: University of Illinois at Urbana-Champaign. PI: Bill Mischo. DUE-0226327. This project is developing second-generation capabilities for two mathematical digital collections that support mathematics, engineering, physics, and applied sciences education and research: the "MathWorld" site at http://mathworld.wolfram.com and the "Functions" site at http://functions.wolfram.com. Functional enhancements are being added to these sites, and collection and item-level metadata from these resources are being integrated into the NSDL Metadata Repository framework via maintenance of an Open Archives Initiative (OAI) server.

Math Tools Project. Institution: Drexel University. PI: Eugene Klotz. DUE-0226284. The Math Forum (www.mathforum.org) is aggregating mathematical software critical to the learning of school mathematics, including software for handheld devices, small interactive web-based tools such as applets, and other small modules based on software application packages.

Collaborative Project: Physics Teaching Web Advisory (Pathway) -- A Digital Video Library for Enhancement and Preparation of Physics Teachers. Institutions: Carnegie-Mellon University and Kansas State University. PI: Scott Stevens, DUE-0226219 and PI: Dean Zollman, DUE-0226157. Kansas State University and Carnegie-Mellon University are creating a proof-of-concept demonstration of a new type of digital library for physics teaching. The project brings together several long-standing research projects in digital video libraries, advanced distance learning technologies, and collaboration technologies, and nationally known experts in physics pedagogy and high-quality content. The project builds on Carnegie Mellon University's Informedia Digital Video Library, which is addressing the problem of information extraction from video and audio content, and "synthetic interviews" of master physics teachers.

Services Track

Projects are expected to develop services that support users, collection providers, or the core integration capabilities, and enhance the impact, efficiency, and value of the library.

Digital Library Service Integration. Institution: Foundation @ NJIT, New Jersey Institute of Technology. PI: Michael Bieber. DUE-0226075. This project is developing a Digital Library Service Integration infrastructure that enables a systematic approach to sharing relevant information services within a seamless, integrated interface. In addition to integrating relatively simple services, the project is also exploring the sharing of services that require customization, such as peer review, to a particular collection or community, incorporating collaborative filtering for customizing large sets of links, and developing advanced lexical analysis tools.

The Development and Use of Digital Collections to Support Interdisciplinary Education. Institution: Washington and Lee University. PI: Frank Settle. DUE-0226152. This project is using the Alsos Digital Library (http://alsos.wlu.edu) as a model for educators wishing to develop digital collections for educational uses, especially those that are multidisciplinary and integrate science and technology with the humanities. A series of workshops targets faculty who want to connect digital collections to their courses through credible, digital, searchable, annotated references. Topics include collection development, software systems, processes for editing materials, integration of collections into courses and curricula, evaluation, and dissemination. Additional discussion centers on assessment, maintenance, culling, technology migration, security, collaboration, and integration into larger digital libraries.

Access NSDL. Institution: WGBH Educational Foundation. PI: Madeleine Rothberg. DUE-0226214. The National Center for Accessible Media (NCAM), a joint effort of WGBH and the Corporation for Public Broadcasting, is building the capacity of the NSDL to serve learners with disabilities. Through this project the NSDL is benefiting from and contributing to the national and international dialogue on access specifications for online learning resources to meet the needs of users with disabilities and to ensure interoperability of accessible content.

Optimizing Workflow and Integration in NSDL Collections. Institution: University of Wisconsin-Madison. PI: John Strikwerda. DUE-0226332. This project is creating turn-key software for collection developers to manage their organizing and cataloging tasks and to create and share item-level metadata. In addition a digital library workflow management knowledge archive is under development.

The NSDL Collaboration Finder: Connecting Projects for Effective and Efficient NSDL Development. Institution: California State University, Trustees. PI: Brandon Muramatsu. DUE-0226277. This project is developing a web-based, searchable and browsable database tool to capture information about the goals, ongoing activities, deliverables, schedules, development stages and discipline areas covered by NSDL projects. Cooperatively populated by the NSDL community and organized by its shared vision, the Collaboration Finder works with the NSDL Community Services

Standing Committee to facilitate its use by the existing collections, services, and targeted research track projects.

Unleashing Supply: Services for Collaborative Content Development. Institution: Wayne State University. PI: Robert Stephenson. DUE-0226367. This project is exploring how to increase the number and quality of STEM Education learning objects by facilitating virtual communities of content developers. An open course application service provider hosts the collaboration tools needed for each community, including a directory of open course projects, a collection of Web-based tools for student use, a consultant database to connect projects with skilled experts, and a set of licenses suitable for open course learning objects. This project is enabling the NSDL to explore an infrastructure to support future content acquisition for its collections.

A Digital IdeaKeeper For K-12: NSDL Scaffolded Portal Services for Information Analysis and Synthesis. Institution: University of Michigan Ann Arbor. PI: Chris Quintana. DUE-0226241. The "IdeaKeeper" is a specialized scaffolded NSDL portal for K-12 science learners with services that support students in analyzing library resources and synthesizing the information into arguments addressing their questions. IdeaKeeper is being deployed in Detroit middle school classrooms to assess the impact of such supportive digital library services on how well they support students in doing information analysis/synthesis and how much students learn about information analysis/synthesis.

Strand Maps as an Interactive Interface to NSDL Resources. Institution: University of Colorado at Boulder. PI: Tamara Sumner. DUE-0226286. The Strand Map Service provides an interactive and flexible interface to the NSDL's educational resources by mapping them to interrelated science learning goals based on the AAAS Benchmarks for Science Literacy and the NRC National Science Education Standards. The service enables educators and learners 1) to discover educational resources that support the learning goals articulated in the strand maps; 2) to browse the interconnected learning goals in the strand maps; and 3) to enhance their own content knowledge by exploring important background information such as prior research on student misconceptions.

Scaling the Peer Review Process for National STEM Education Digital Library Collections. Institution: California State University, Trustees. PI: Gerard Hanley. DUE-0226269. This project is addressing a key challenge to implementing and scaling peer review systems for NSDL collections, namely high costs associated with face-to-face training, retention of reviewers, and the difficulty reviewers face in keeping pace with the rapidly increasing size of digital collections of educational material. A professional development module for training peer reviewers targets NSDL collection developers (and others) to effectively and efficiently implement and sustain peer review of their collections. To facilitate adaptation and usage by other collections, the module is designed as a channel adaptable to the uPortal framework being implemented by the NSDL Core Integration development team.

Collaborative Proposal: Managing Authority Lists for Customized Linking and Visualization: A Service for the National STEM Education Digital Library. Institutions: Johns Hopkins University and Tufts University. PI: Golam Choudhury, DUE-0226234 and PI: Gregory Colati, DUE-0226304. The Services for a

Customizable Authority Linking Environment (SCALE) project supports two broad classes of service for the NSDL. First are linking services that automatically bind key words and phrases to supplementary information to help students, professionals outside a particular discipline, and the interested public to read documents full of unfamiliar technical terms and concepts. A second class of service bases automatic linking on authority control of names and terms and on links among different authority lists such as thesauri, glossaries, encyclopedias, subject hierarchies, and object catalogues.

Targeted Research Track

Projects are expected to explore specific topics that have immediate applicability to the Collections or Services track or to the NSDL as a whole.

Using Spatial Hypertext as a Workspace for Digital Library Providers and Patrons. Institution: Texas Engineering Experiment Station. PI: Frank Shipman. DUE-0226321. This project is investigating the use of spatial hypertext by digital library patrons to build personal and shared annotated digital information spaces, and by digital library providers to organize, annotate, and maintain collections of digital information. Spatial hypertext enables users to collect source materials as information objects in a set of two-dimensional spaces and imply attributes of and relationships between the materials via visual and spatial cues.

ReMarkable Texts: A Digital Notepad for the NSDL. Institution: Brown University. PI: Andy van Dam. DUE-0226216. Faculty and students are investigating capabilities for an innovative pen-based digital notebook to enable users, particularly students, to work and interact with NSDL's digital materials in a personalized manner. The main features include viewing, taking notes on, annotating (e.g. freehand ink, post-it notes, and bidirectional fine-grained hyperlinks), organizing, and collaborating on multimedia documents, all with the ability to replay the temporal sequence of one's notes in the contexts in which they were made. Whiteboarding and audio facilities are also supported.

Effective Access: Using Digital Libraries to Enhance High School Teaching in STEM. Institution: Education Development Center. PI: Katherine Hanson. DUE-0226483. This project is studying the use of networked digital resources by secondary school science, technology, engineering, and mathematics (STEM) teachers as they seek and select quality materials and tailor them to fit the learning environment of their classroom. Areas of interest include: characterizing the use of digital resources for preparing lessons and for incorporating materials directly into student projects; and investigating types of software tools, lesson templates, and support that enable teachers to integrate digital library resources into general classroom settings.

Question Triage for Experts and Documents: Expanding the Information Retrieval Function of the NSDL. Institution: University of Massachusetts Amherst. PI: W. Bruce Croft. DUE-0226144. This project is investigating the merger of the information retrieval (IR) and digital reference components of the National STEM Education Digital Library (NSDL). Combining these functions enables users to find answers to questions regardless if those answers come from documents in NSDL collections or experts accessible through the NSDL's virtual reference desk.

MetaTest: Evaluating the Quality and Utility of Metadata. Institution: Syracuse University. PI: Elizabeth Liddy. DUE-0226312. Researchers are evaluating the use and utility of metadata from multiple perspectives. These include the comparison of the subjective quality of metadata that is assigned both manually and automatically to learning resources; the comparison of the retrieval effectiveness due to metadata that is assigned manually versus automatically to learning resources; determination of searching and browsing behaviors of users when engaged in information seeking in the digital library; and an analysis of the relative contribution of individual elements of the GEM + Dublin Core metadata scheme to users' searching and browsing behavior.

Integrating Digital Library Resources into Online Courses. Institution: Texas Engineering Experiment Station. PI: Connie McKinzie. DUE-0226217. This project is investigating a model for integrating content from NSDL into online and web-enhanced courses in higher education, using as a testbed community the students and faculty at the Texas A&M Electronic Teachers College (ETC). Services under investigation include mapping of modular learning object content to teacher certification competencies in high area needs such as mathematics; defining competency-based metadata vocabularies for these learning objects that reflects the teacher certification requirements; and creating a Content-Packaging Tool Suite to enable faculty to select learning objects from a repository, build them into learning modules, and connect them into course-management systems or other delivery platforms.

Examples of Practical Digital Libraries: Collections Built Internationally Using Greenstone

Ian H. Witten

New Zealand Digital Library Project
Department of Computer Science
University of Waikato, New Zealand
ihw@cs.waikato.ac.nz

Although the field of digital libraries is still young, digital library collections have been built around the world and are being deployed on numerous public web sites. But what is a digital library, exactly? In many respects the best way to characterize the notion is by extension, in terms of actual examples, rather than by intension as in a conventional definition. In a very real sense, digital libraries are whatever people choose to call by the term "digital library."

The Greenstone Digital Library Software[1] provides a way of building and distributing digital library collections, opening up new possibilities for organizing information and making it available over the Internet or on CD-ROM (Witten and Bainbridge, 2003). Produced by the New Zealand Digital Library Project, Greenstone is intended to lower the bar to the construction of practical digital libraries, yet at the same time leave a great deal of flexibility in the hands of the user.

In accordance with the maxim "simple things should be easy, complex things should be possible" new users can quickly put together standard-looking collections from a set of source documents that may be HTML, Word, PDF, or many other formats (Witten *et al*, 2001). Given an existing collection, it is easy to clone its structure and populate an identical copy with entirely new documents, provided they are in the same formats as those in the existing collection. A more committed user who studies the options that Greenstone offers can personalize the digital library system and create new kinds of collection that take advantage of available metadata to provide different kinds of browsing facilities, which are akin to different perspectives on the collection. Users with programming skills can extend the system by adding modular units called "plugins" that accommodate new document and metadata formats, and new browsing and document access facilities.

Greenstone has been used to make many digital library collections. Some were created within NZDL as demonstration collections. However, the use of Greenstone internationally is growing rapidly, and several web sites show collections created by external users. Most contain unusual and interesting material, and present it in novel and imaginative ways. This paper briefly reviews a selection of Greenstone digital library sites to give a feeling for what public digital libraries are being used for. Examples are given from different countries: China, Germany, India, Russia, UK, US;

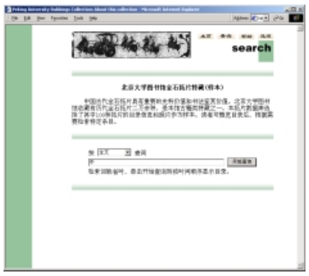
[1] Available from www.greenstone.org

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 67-74, 2002.

different kinds of library: historical, educational, cultural, research; different sorts of source material: text, document images, pictures, voice. There are many more examples of Greenstone in other places: this is just a small sampling.

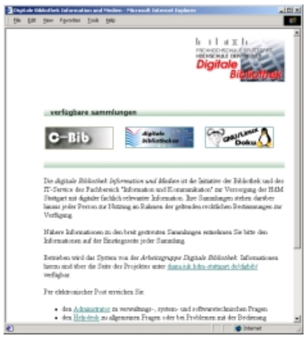
Examples of Digital Libraries

China *Peking University digital library* http://162.105.138.23/tapian/tp.htm



An experimental collection was created by the Chinese Department at Peking University with the assistance of a NZDL project member who visited there some years ago. It contains rubbings of Tang Dynasty poetry, whose originals were carved into wood or stone. These are collections of images, but the text has been hand-entered into electronic form. The entire interface is in Chinese, and, like all Greenstone collections, is fully searchable.

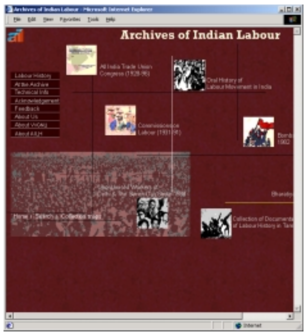
Germany *Digitale Bibliothek Information und Medien* http://diana.iuk.hdm-stuttgart.de/digbib/gsdl/cgi-bin/library



This “Information and Media” digital library created by the University of Applied Sciences, Stuttgart, Germany, includes three collections: a bibliography, full-text documents about digital libraries and related topics, and technical documentation on the Linux operating system.

While most digital libraries give every appearance of being unmanned, this one advertises an administrator and helpdesk, both reachable by email.

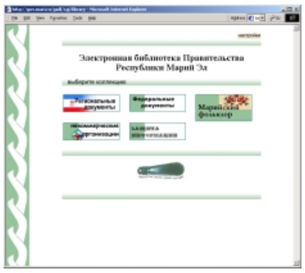
India *Archives of Indian Labour* http://www.indialabourarchives.org/


Archives of Indian Labour
Labour History
At the Archive
Technical Info
Acknowledgement
Feedback
About Us
About VVGNLI
About AILH
All India Trade Union Congress (1928-96)
Oral History of Labour Movement in India

A collaborative project between the National Labour Institute and the Association of Indian Labour Historians, the Archives of Indian Labour are dedicated to preserving and making accessible historic documents on the Indian working class. This library contains collections on the All India Trade Union Congress (1928-96), the Commission on Labour (1930-1991), an Oral History of the Labour Movement in India, and special collections on key events in India’s labour history.

Russia *Mari El Republic government information* http://gov.mari.ru/gsdl/cgi/library

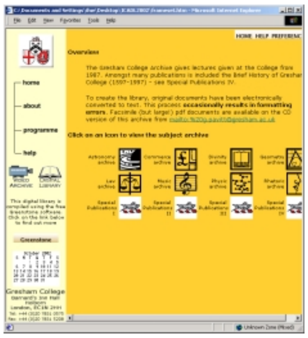
The regional government department in the Mari El Republic of the Russian Federation has built several Russian-language collections. This site is interesting because, by themselves and on their own initiative, they added a Russian-language interface to Greenstone, which at the time offered several other different user interface languages. Since then, interfaces in languages such as Hebrew and Indonesian have been added to the standard list, which includes most European languages, Chinese, Arabic, and Maori.



Электронная библиотека Правительства Республики Марий Эл

UK *Gresham College Archives* http://www.gresham.ac.uk/greenstone/frameset.html

This collection includes all lectures given at the Gresham College, London, from 1987, among with many other special publications, such as the *Brief History of Gresham College (1597-1997)*. It is divided manually into the various subjects covered by the College. The collection is also issued on a standalone Greenstone CD-ROM that can self-installs on any Windows computer and is accessed through a Web browser in exactly the same way as the online version.



Overview

The Gresham College Archive gives lectures given at the College from 1987. Amongst many publications is included the Brief History of Gresham College (1597-1997) - see Special Publications IV.

To create the library, original documents have been electronically converted to text. This process **occasionally results in formatting errors**. Facsimile (but larger) pdf documents are available on the CD version of this archive from mailto:%20g.pavitt@gresham.ac.uk

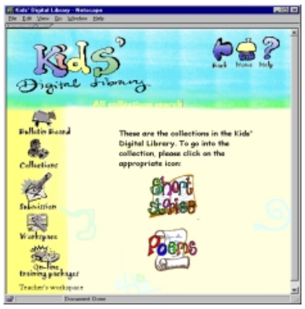
Click on an icon to view the subject archive

This digital library is compiled using the free Greenstone software. Click on the link below to find out more

Gresham College
Barnard's Inn Hall
Holborn
London, EC1N 2HH
Tel: +44 (0)20 7831 0575
Fax: +44 (0)20 7831 5208

UK *Kids' digital library* http://kidsdl.mdx.ac.uk/kidslibrary

A project at Middlesex University has been experimenting with a "Kids' digital library," deployed in a school in North London. Children can submit stories and poems to the library, which contains a collection of their own work. Teachers can monitor submissions before they are incorporated. This project has involved significant changes to Greenstone at the coding level, which is made possible because Greenstone is open source software.

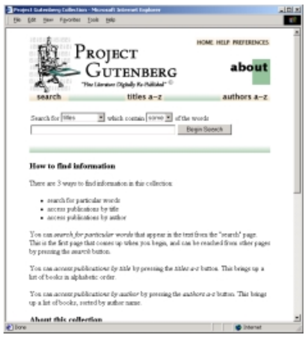


These are the collections in the Kids' Digital Library. To go into the collection, please click on the appropriate icon:

US *Project Gutenberg*

http://public.ibiblio.org/gsdl/cgi-bin/library.cgi?a=p&p=about&c=gberg

An on-going project to produce and distribute free electronic editions of literature, Gutenberg now contains more than 3,700 titles from Shakespeare to Dickens to the Bronte sisters. This site, maintained by Ibiblio, one of the original Gutenberg mirror sites, uses Greenstone to make the entire collection available in fully searchable form. Access to large collections through full-text search is simple, fast, requires no metadata, and scales easily to massive amounts of text.


Project Gutenberg Collection - Microsoft Internet Explorer
File Edit View Favorites Tools Help
HOME HELP PREFERENCES
PROJECT GUTENBERG
"Fine Literature Digitally Re-Published"
about
search
titles a-z
authors a-z
Begin Search
Done
Internet

Search for titles which contain some of the words

How to find information

There are 3 ways to find information in this collection:

- search for particular words
- access publications by title
- access publications by author

You can *search for particular words* that appear in the text from the "search" page. This is the first page that comes up when you begin, and can be reached from other pages by pressing the search button.

You can *access publications by title* by pressing the *titles a-z* button. This brings up a list of books in alphabetic order.

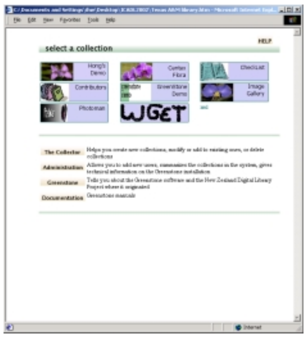
You can *access publications by author* by pressing the *authors a-z* button. This brings up a list of books, sorted by author name.

About this collection

US *Center for the Study of Digital Libraries*

http://botany.cs.tamu.edu/greenstone

This digital libraries research site at Texas A&M University has an emphasis on digital floras—collections of digital images of plants. There are several prototype Greenstone collections containing numerous plant images, classified according to a family tree, and a separate collection with detailed biological descriptions of plants. Good use is made of Greenstone's hierarchical browsing facilities to allow access through standard biological taxonomic structures.

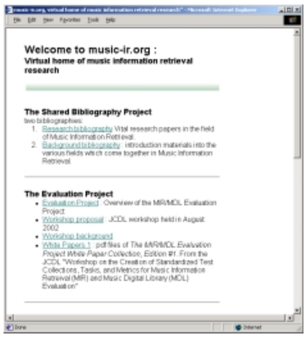


The Collector	Helps you create new collections, modify or add to existing ones, or delete collections
Administration	Allows you to add new users, summarise the collections in the system, gives technical information on the Greenstone installation
Greenstone	Tells you about the Greenstone software and the New Zealand Digital Library Project where it originated
Documentation	Greenstone manuals

US *Music information retrieval research*

http://www.music-ir.org/

This site calls itself the "virtual home of music information retrieval research." Its main content comprises two fully-searchable and browsable bibliographic collections: a research bibliography comprising vital research papers in the field of Music Information Retrieval; and a background bibliography with introductory materials to the various fields which come together in Music Information Retrieval. However, it does not yet use Greenstone's extensive music retrieval capabilities (see below, under New Zealand Digital Library).



Welcome to music-ir.org :
Virtual home of music information retrieval research

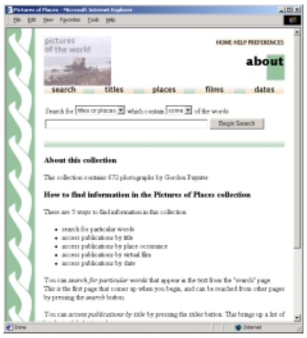
The Shared Bibliography Project
two bibliographies:

1. Research bibliography Vital research papers in the field of Music Information Retrieval.
2. Background bibliography : introduction materials into the various fields which come together in Music Information Retrieval.

The Evaluation Project

- Evaluation Project : Overview of the MIR/MDL Evaluation Project
- Workshop proposal : JCDL workshop held in August 2002
- Workshop background
- White Papers 1 : pdf files of *The MIR/MDL Evaluation Project White Paper Collection, Edition #1.* From the JCDL "Workshop on the Creation of Standardized Test Collections, Tasks, and Metrics for Music Information Retrieval (MIR) and Music Digital Library (MDL) Evaluation"

US *Pictures of the world* http://tuatara.ucr.edu/gsdl-bin/library?a=p&p=about&c=pictures



Search for titles or places which contain some of the words

About this collection

This collection contains 672 photographs by Gordon Paynter

How to find information in the Pictures of Places collection

There are 5 ways to find information in this collection:

- search for particular words
- access publications by title
- access publications by place occurrence
- access publications by virtual film
- access publications by date

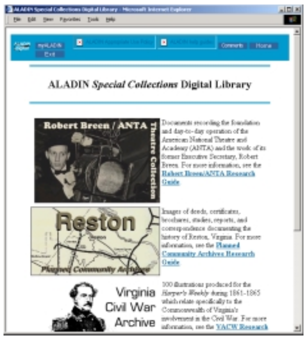
You can *search for particular words* that appear in the text from the "search" page. This is the first page that comes up when you begin, and can be reached from other pages by pressing the *search* button.

You can *access publications by title* by pressing the titles button. This brings up a list of

This is a personal collection of photographs, due to Gordon Paynter, which presents a rich set of searching and browsing options—by date, place, title, and reel of film. Although quite small at present, there are virtually no limits to the amount that it can grow, because all the structure is based on metadata, which is quite small in volume. During early testing, Greenstone was used to build collections of over 10 million relatively lengthy metadata items (in the form of MARC records) without any problems arising.

The metadata for these photographs was entered in a succinct XML format that allows multiple assignments of metadata to the same item, and a single metadata assignment to apply to several items. The directory hierarchy containing the source files, or the filename conventions, can be used to allow assignment of the same metadata value simultaneously to large numbers of files. When new metadata values are assigned to an item, any previous values can be added to, or ignored.

US *Aladin digital library* http://www.aladin.wrlc.org/gsdl/


ALADIN Special Collections Digital Library - Microsoft Internet Explorer
File Edit View Favorites Tools Help
ALADIN digital
myALADIN
Exit
ALADIN Appropriate Use Policy
ALADIN help guides
Comments
Home
Robert Breen / ANTA
Theatre Collection
Reston
Planned Community Archives
Virginia
Civil War
Archive

ALADIN *Special Collections* Digital Library

Documents recording the foundation and day-to-day operation of the American National Theatre and Academy (ANTA) and the work of its former Executive Secretary, Robert Breen. For more information, see the Robert Breen/ANTA Research Guide.

Images of deeds, certificates, brochures, studies, reports, and correspondence documenting the history of Reston, Virginia. For more information, see the Planned Community Archives Research Guide.

100 illustrations produced for the *Harper's Weekly* during 1861-1865 which relate specifically to the Commonwealth of Virginia's involvement in the Civil War. For more information, see the YACW Research

This site contains digital material from the special collections of the seven universities of the Washington Research Library Consortium in Washington D.C. There are presently four collections. The first contains documents recording the foundation and day-to-day operation of the American National Theatre and Academy. The second has images of deeds, certificates, brochures, studies, reports, and correspondence documenting the history of Reston, Virginia. The third has one hundred illustrations produced for the *Harper's Weekly* during 1861–1865 which relate specifically to the Commonwealth of Virginia's involvement in the Civil War. The fourth is a predominantly audio collection that gives recordings of interviews conducted by Felix Grant of jazz and blues artists. For copyright reasons, access to the digital audio files is restricted to members of the Washington Research Library Consortium.

US *New York Botanical Garden* not yet released

The LuEsther T. Mertz Library
Rare Book Digitization Project

This project has been made possible by a generous grant from
The Andrew W. Mellon Foundation

[illegible]

Click here or on the photo above to view and search online titles.

About the Collection

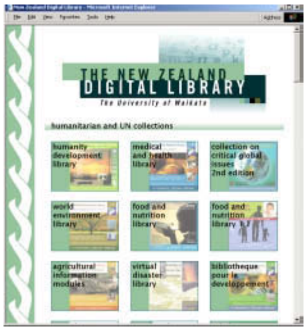
Through collaboration with the Andrew W. Mellon Foundation, The LuEsther T. Mertz Library has begun a pilot project to digitize and make Web-accessible three rare 19th century works on American trees by French botanists André and François André Michaux. The eight-volume collection represents three important illustrated botanical

The LuEsther T. Mertz Library has begun to digitize and make Web-accessible three rare 19th century works on American trees by French botanists André and François André Michaux. This eight-volume collection of three important illustrated botanical books reflects the early investigation of the flora of North America by botanists who were seeking new plants for commerce and horticulture. It contains many gorgeous full color plates.

US *Mercy Corps*

The Mercy Corps, centered in Portland Oregon and with operations in about thirty of the world's most unstable countries, is using Greenstone to organize its extensive collection of in-house documents, manuals, forms, and memos. This is not a public site. However, it is especially noteworthy because Mercy Corps have made significant enhancements to Greenstone to support a workflow for new acquisitions to the library. Field offices submit new documents by filling out metadata on a simple web-based form and attaching the document. It arrives in the in-tray of a central librarian who checks it for correctness and integrity before finally including it in the appropriate collection. Collections, rebuilt automatically every night, are available on the web for in-house use, and are written at regular intervals to CD-ROM for physical distribution.

NZ *New Zealand Digital Library* http://www.nzdl.org



THE NEW ZEALAND
DIGITAL LIBRARY
The University of Waikato

humanitarian and UN collections

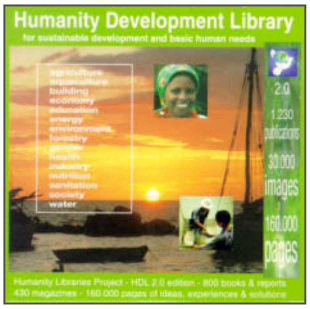
humanity development library
medical and health library
collection on critical global issues 2nd edition
world environment library
food and nutrition library
food and nutrition library
agricultural information modules
virtual disaster library
bibliotheque pour le developpement

The New Zealand Digital Library website shows several dozen demonstration collections built by project staff. Some highlight particularly unusual capabilities: here are two.

The Musical Digital Library subsection offers several innovative collections that involve music retrieval by singing or humming a snatch of the desired tune. In some collections, text search can be combined with melody matching to yield a more comprehensive search technique.

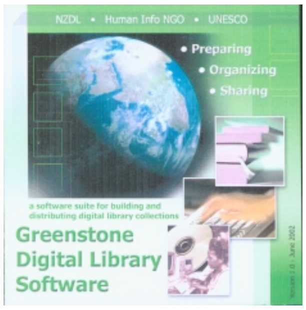
First Aid in Pictures is a collection designed for illiterate users: it presents purely pictorial, diagrammatic, information on basic First Aid. All the indexing mechanisms are also purely visually based. Explanatory text can be displayed at the bottom of each page, and spoken by a voice synthesizer.

Intl *Humanitarian collections* http://www.globalprojects.org


Humanity Development Library
for sustainable development and basic human needs
agriculture
aquaculture
building
economy
education
energy
environment
forestry
gender
health
industry
nutrition
sanitation
society
water
2.0
1,230 publications
30,000 images
160,000 pages
Humanity Libraries Project - HDL 2.0 edition - 800 books & reports
430 magazines - 160,000 pages of ideas, experiences & solutions

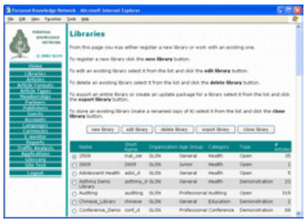
Greenstone is being used to deliver humanitarian and related information in developing countries on CD-ROM. There are about twenty different collections from organizations such as the United Nations University, UNESCO, Food and Agri-culture Organization, the World Health Organization, the Pan-American Health organization, GTZ, the United Nations Development Programme, and UNAIDS.
Many of these are produced by Human Info, a small NGO in Belgium, in conjunction with an OCR service bureau in Romania.

Intl *UNESCO project* http://www.unesco.org/webworld/build_info/gct/bestpractices/anthologies.shtml


NZDL • Human Info NGO • UNESCO
• Preparing
• Organizing
• Sharing
a software suite for building and distributing digital library collections
Greenstone
Digital Library
Software
Version 2.4 · June 2002

UNESCO is participating in developing and distributing Greenstone. Digital libraries are radically reforming how information is disseminated and acquired in UNESCO's partner communities and institutions in the fields of education, science and culture around the world, particularly in developing countries. The Greenstone project is an international cooperative effort with UNESCO established in August 2000. This initiative will encourage the effective deployment of digital libraries to share information and place it in the public domain.

Intl *Global Library Services Network* http://www.glsn.com



Libraries

From this page you may either register a new library or work with an existing one.

To register a new library click the **new library** button.

To edit an existing library select it from the list and click the **edit library** button.

To delete an existing library select it from the list and click the **delete library** button.

To export an entire library or create an update package for a library select it from the list and click the **export library** button.

To clone an existing library (make a renamed copy of it) select it from the list and click the **clone library** button.

	Name	Short Name	Organization	Age Group	Category	Type	# Articles
●	2509	mal_ser	GLSN	General	Health	Open	35
○	2609	2609	GLSN	Junior	Health	Open	1
○	Adolescent Health	adol_d	GLSN	General	Health	Open	5
○	Asthma Demo Library	asthma_d	GLSN	General	Health	Demonstration	23
○	Auditing	auditing	GLSN	Professional	Auditing	Open	119
○	Chinese_Library	chinese	GLSN	General	Education	Demonstration	2
○	Conference_Demo	conf_d	GLSN	Professional	Conference	Demonstration	68

GLSN provides remote communities with access to digital libraries for use offline. It implements an architecture and infrastructure to allow large-scale non-networked digital libraries in remote places to be acquired, installed, and updated, on a commercial (but low-cost) basis. GLSN makes arrangements with information providers, both commercial and non-commercial, to populate the collections. There are many freely available demonstration collections, mostly on health topics (Adolescent Health, Asthma, Chinese medicine, to name just a few).

GLSN uses Greenstone at its core. Like the Peace Corporation (see above), it has extended the basic facilities of Greenstone with an interactive web-based interface for selecting documents and gathering metadata.

Conclusion

The examples above show a wide variety of different types of digital library. And they are by no means exhaustive: we know of many other Greenstone collections in countries from Canada to South Africa, some of which have unusual features such as collections optimized for viewing on small-screen handheld devices.

The last four examples represent institutional rather than individual users. Each has large numbers of different collections. The New Zealand Digital Library, which originated Greenstone, offers scores of collections and represents the cutting edge of digital library research using Greenstone as a vehicle for dissemination. The Humanitarian collections involve a huge ancillary effort in digitizing thousands of books, reports, and other documents for inclusion on Greenstone CD-ROMs, and a vast distribution mechanism—fifty thousand copies are distributed annually, of which 60% are provided free. The UNESCO project distributes not information collections themselves but the capacity to build new information collections, which is a more effective strategy for sustained long-term human development. Finally, Global Library Services Network is a large-scale commercial application of Greenstone which is aimed at the educational and health sectors, particularly in remote regions.

Virtually every new collection involves its own idiosyncratic requirements. Consequently, those building digital libraries need constant access to advice and assistance from others, in order to continue to learn how to tailor the software to meet ever-changing new requirements. There is a lively email discussion group for assistance with Greenstone; it contains participants from over 40 different countries. The software itself is being downloaded around 1500 times per month, on average.

Truly the world of practical digital libraries is burgeoning. The time has come to stop talking about digital libraries and get on with building them!

Acknowledgements. The Greenstone Digital Library software has grown out of the stimulating research environment of the New Zealand Digital Library project, and I would like to acknowledge the profound influence of all project members.

References

1. Witten, I.H., Bainbridge, D. and Boddie, S. (2001) Power to the people: end-user building of digital library collections. *Proc ACM Digital Libraries*, Roanoke, VA.
2. Witten, I.H. and Bainbridge, D. (2003) *How to build a digital library*. Morgan Kaufmann, San Francisco, CA.

Data Mining Technologies for Digital Libraries and Web Information Systems

Ramakrishnan Srikant

IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120, USA
srikant@us.ibm.com

In the first half of the talk, I will discuss data mining technologies that can result in better browsing and searching. Consider the problem of merging documents from different categorizations (taxonomies) into a single master categorization. Current classifiers ignore the implicit similarity information present in the source categorizations. I will show that by incorporating this information into the classification model, classification accuracy can be substantially improved [1]. Next, I will demonstrate novel search technology that treats numbers as first-class objects, and thus yields dramatically better results than current Web search engines when searching over product descriptions or other number-rich documents [2].

The second half of the talk will focus on privacy. I will give a brief introduction to the field of private information retrieval [3], which allows users to retrieve documents without the library identifying which document was retrieved. I will then cover the exciting new research area of privacy preserving data mining [4] [5], which allows us to build accurate data mining models without access to precise information in individual data records, thus finessing the potential conflict between privacy and data mining.

References

1. Rakesh Agrawal and Ramakrishnan Srikant. On catalog integration. In *Proc. of the Tenth Int'l World Wide Web Conference,* Hong Kong, May 2001.
2. Rakesh Agrawal and Ramakrishnan Srikant. Searching with numbers. In *Proc. of the Eleventh Int'l World Wide Web Conference,* Honolulu, Hawaii, May 2002.
3. Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In *IEEE Symposium on Foundations of Computer Science*, pp. 41-50, 1995.
4. Rakesh Agrawal and Ramakrishnan Srikant. Privacy preserving data mining. In *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 439-450, Dallas, Texas, May 2000.
5. Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. In *Proc. of the 8th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining,* Edmonton, Canada, July 2002.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 75, 2002.

Chinese Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis

Jen-Yuan Yeh[1], Hao-Ren Ke[2], and Wei-Pang Yang[1]

[1] Department of Computer & Information Science, National Chiao-Tung University,
1001 Ta Hsueh Rd., Hsinchu, Taiwan 30050, R.O.C.
{jyyeh, wpyang}@cis.nctu.edu.tw
[2] Digital Library & Information Section of Library, National Chiao-Tung University,
1001 Ta Hsueh Rd., Hsinchu, Taiwan 30050, R.O.C.
claven@lib.nctu.edu.tw

Abstract. In this paper, two novel approaches are proposed to extract important sentences from a document to create its summary. The first is a corpus-based approach using feature analysis. It brings up three new ideas: 1) to employ ranked position to emphasize the significance of sentence position, 2) to reshape word unit to achieve higher accuracy of keyword importance, and 3) to train a score function by the genetic algorithm for obtaining a suitable combination of feature weights. The second approach combines the ideas of latent semantic analysis and text relationship maps to interpret conceptual structures of a document. Both approaches are applied to Chinese text summarization. The two approaches were evaluated by using a data corpus composed of 100 articles about politics from New Taiwan Weekly, and when the compression ratio was 30%, average recalls of 52.0% and 45.6% were achieved respectively.

1 Introduction

With the advent of the Information Age, people are faced with the problem of finding relevant information efficiently and effectively. Today, text searching and summarization are two essential technologies to solve this problem. Text search engines serve as information filters that retrieves an initial set of relevant documents, and text summarizers play the role of information spotters to help users locate a final set of desired documents [7].

Automated text summarization extracts the essence of a text. In general, the process can be decomposed into three phases: analyzing the input text, transforming it into a summary representation, and synthesizing an appropriate output [16]. Text summarizers can be roughly classified into two categories: one is based on feature analysis [1] [6] [10] [13] [15] [18], and the other is based on the understanding of the meaning of the text [2] [3] [7] [10] [11] [17] [19]. The former estimates the possibility that a sentence belongs to the summary according to a statistical model. The latter identifies conceptual structures in the document by external resources, such as *WordNet* [21], and generates a summary according to the structure representations.

In this paper, we present two novel approaches: *Modified Corpus-based Approach*, and *LSA-based T.R.M. Approach*. The first is based on feature analysis, and the sec-

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 76-87, 2002.

ond uses latent semantic analysis (LSA) and a text relationship map (T.R.M) to extract salient semantic information from the document.

The remainder of this paper is organized into five sections. Section 2 describes related studies, Section 3 and Section 4 elaborate our two approaches, Section 5 presents the experimental results, and Section 6 gives conclusions.

2 Related Work

2.1 The History of Text Summarization

Research on text summarization started in the 1950s. Due to the lack of powerful computers and Nature Language Processing (NLP) techniques, early work focused on the study of text genres such as position and cue phrases [6]. From the 70s to the early 80s, Artificial Intelligence (AI) was applied. The idea was to employ knowledge representations, such as frames or templates, to identify conceptual structures of a text and find salient concepts by inference [2] [18]. But, the main drawback is that limited templates make conceptual structures incomplete.

Since the early 90s, Information Retrieval (IR) has been used [1] [10] [13] [18]. As with IR, text summarization can be regarded as how to find significant sentences in a document. However, since IR techniques applied to text summarization focus on symbolic-level analysis, it does not take semantic issues into account. Besides the above techniques, there are still two kinds of methods. One is based on cognitive psychology and the other is based on computational linguistics.

2.2 Corpus-Based Approaches

Recently, corpus-based approaches have played an important role in text summarization [1] [10] [13] [18]. By exploiting technologies of machine learning, it becomes possible to learn rules from a corpus of documents and the corresponding summaries. The process is decomposed into two phases: the training phase and the test phase. In the training phase, the system extracts particular features from the training corpus and generates rules by a learning algorithm. In the test phase, the system applies rules learned from the training phase to the test corpus to generate the corresponding summaries; furthermore, the performance is measured.

Kupiec et al. (1995) proposed a trainable summarizer based on Bayes's rule [13]. For each sentence s, the probability that it belongs to the summary S, given k features F_j, is computed. The probability is expressed as Equation 1, where $P(F_j|s\in S)$ is the probability that F_j appears in a summary sentence, $P(s\in S)$ is the ratio of the number of summary sentences to the total number of sentences, and $P(F_j)$ is the probability that F_j appears in the training corpus. All of them can be computed from the training corpus.

$$P\left(s \in S \mid F_1, F_2, \ldots, F_k\right) = \frac{\prod_{j=1}^{k} P\left(F_j \mid s \in S\right) P\left(s \in S\right)}{\prod_{j=1}^{k} P\left(F_j\right)} \ . \tag{1}$$

The features used in their experiments were “Sentence Length”, “Fixed-Phrase”, “Paragraph”, “Thematic Word”, and “Uppercase Word”. The results showed that the best combination of features was *paragraph+fixed-phrase+sentence length*.

2.3 Text Summarization Using a Text Relationship Map

Salton et al. (1997) introduced the construction of a *text relationship map* (T.R.M.) to link similar paragraphs [19]. In the map, each node stands for a paragraph and is represented by a vector of weighted terms. A link is created between two nodes if the two corresponding paragraphs have *strong* relevance. The relevance between two paragraphs is determined by their similarity which is typically computed as the inner product between the corresponding vectors. When the similarity between two paragraphs is larger than a predefined threshold, the link is constructed.

They defined *bushiness* to measure the significance of a paragraph. The bushiness of a paragraph is the number of links connecting it to other paragraphs. They also proposed three heuristic methods to generate a summary: global bushy path, depth-first path, and segmented bushy path. Since a highly bushy node is linked to many other nodes (i.e. it has many overlapping vocabularies with others), it is likely to discuss main topics covered in many other paragraphs.

3 Modified Corpus-Based Approach

In this section, we propose a trainable summarizer. Three new ideas are employed to improve corpus-based text summarization. First, sentence positions are ranked to emphasize the significance of sentence positions; second, word units are reshaped to achieve higher accuracy of keyword importance; third, the score function is trained by the genetic algorithm to obtain a suitable combination of feature weights.

For a sentence *s*, the belief (score) that it belongs to the summary is calculated given the following features:

f_1: *Position*—Important sentences are usually located at some particular positions. For example, the first sentence in the first paragraph always introduces main ideas; hence, it is much more important than other sentences. Additionally, we believe that even for two sentences in the summary, their positions give rise to a difference in their significance. To emphasize the significance of sentence position when creating summaries, each sentence in the summaries of the training corpus is given a rank ranging from 1 to 5 in our implementation. For a sentence *s*, this feature-score is defined as Equation 2, where *s* comes from $Position_i$.

$$Score_{f_1}(s) = P(s \in S \mid Position_i) \times \frac{Avg.\ rank\ of\ Position_i}{5.0} \quad . \tag{2}$$

f_2: *Positive Keyword*—Since words are the basic elements of a sentence, the more content-bearing keywords a sentence has, the more important it is. For a sentence *s*, assume that it contains $Keyword_1$, $Keyword_2$, …, $Keyword_n$. This feature-score is defined as Equation 3, where c_k is the number of $Keyword_k$ occurring in *s*.

$$Score_{f_2}(s) = \sum_{k=1,2,\ldots,n} c_k \cdot P(s \in S \mid Keyword_k) \ . \tag{3}$$

f_3: Negative Keyword—In contrast to f_2, *negative keywords* are those frequent keywords that are not included in the summary. For a sentence *s*, assume that it contains $Keyword_1$, $Keyword_2$, …, $Keyword_{n.}$ This feature-score is defined as Equation 4, where c_k is the number of $Keyword_k$ occurring in *s*.

$$Score_{f_3}(s) = \sum_{k=1,2,\ldots,n} c_k \cdot P(s \notin S \mid Keyword_k) \ . \tag{4}$$

f_4: Resemblance to the Title—It is obvious that the title always sums up the theme of a document. Therefore, the more overlapping keywords a sentence has with the title, the more important it is. For a sentence *s*, this feature-score is defined as Equation 5.

$$Score_{f_4}(s) = \frac{|keywords\ in\ s\ \cap\ keywords\ in\ the\ title|}{|keywords\ in\ s\ \cup\ keywords\ in\ the\ title|} \ . \tag{5}$$

f_5: Centrality—The centrality of a sentence is its similarity to other sentences, which is usually measured as the degree of vocabulary overlapping between it and other sentences. Generally, the sentence with the highest centrality is the centroid of the document. For a sentence *s*, this feature-score is defined as Equation 6.

$$Score_{f_5}(s) = \frac{|keywords\ in\ s\ \cap\ keywords\ in\ other\ sentences|}{|keywords\ in\ s\ \cup\ keywords\ in\ other\ sentences|} \ . \tag{6}$$

In our implementation, a dictionary is used to identify keywords. The major drawback is that it is difficult to find new keywords that are not in the dictionary. To overcome this shortcoming, *word co-occurrence* [12] is computed for adjacent Chinese characters (or keywords) to determine if they can constitute a new keyword. We also consider these new keywords when computing scores of keyword-based features (i.e. f_2, f_3, f_4, and f_5), to achieve higher accuracy of keyword importance.

For a sentence *s*, a weighted score function (Equation 7) is defined to integrate all of the above-mentioned feature scores, where w_i indicates the weight of f_i.

$$\begin{aligned} Score(s) = {} & w_1 \cdot Score_{f_1}(s) + w_2 \cdot Score_{f_2}(s) - w_3 \cdot Score_{f_3}(s) \\ & + w_4 \cdot Score_{f_4}(s) + w_5 \cdot Score_{f_5}(s) \end{aligned} \ . \tag{7}$$

The genetic algorithm is exploited for obtaining an appropriate score function. The chromosome is represented as the combination of feature weights such as (w_1, w_2, w_3, w_4, w_5). To measure the effectiveness of a genome, we define the fitness as the average recall obtained with the genome when applied to the training corpus. To perform the genetic algorithm, we produce 1,000 genomes, evaluate the fitness of each genome, retain the fittest 10 to mate and reproduce new genomes of the next generation. In our experiment, we evaluated 100 generations, and obtained the steady combinations of feature weights.

By using the genetic algorithm, a suitable combination of feature weights can be found. Although we cannot guarantee that the "appropriate" score function will perform well for the test corpus, we can assert that it will perform well if the genre of the test corpus is very close to that of the training corpus..

Before selecting sentences to construct a summary, all sentences are ranked according to their scores as calculated in Equation 7. A designated number of the top-scoring sentences are picked to form the summary.

4 LSA-Based T.R.M. Approach

Latent semantic analysis (LSA) is an automatic mathematical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse [14]. We apply LSA for deriving latent structures from the document. In the following sections, the method to derive semantic representations by LSA is presented, and a novel method to generate the summary according to semantic representations is proposed.

4.1 The Process of LSA-Based T.R.M.

The process shown as Figure 1 consists of four phases: 1) *Preprocessing*, 2) *Semantic Model Analysis*, 3) *Text Relationship Map Construction*, and 4) *Sentence Selection*.

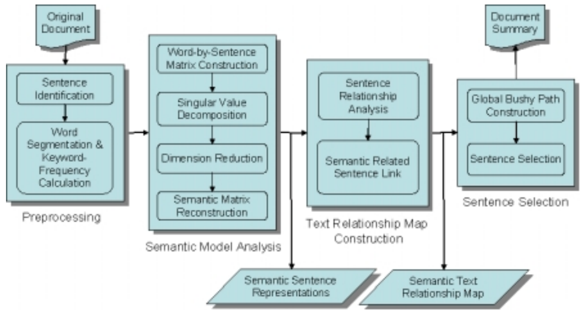

Original Document
Sentence Identification
Word Segmentation & Keyword-Frequency Calculation
Preprocessing
Word-by-Sentence Matrix Construction
Singular Value Decomposition
Dimension Reduction
Semantic Matrix Reconstruction
Semantic Model Analysis
Sentence Relationship Analysis
Semantic Related Sentence Link
Text Relationship Map Construction
Document Summary
Global Bushy Path Construction
Sentence Selection
Sentence Selection
Semantic Sentence Representations
Semantic Text Relationship Map

Fig. 1. Process of our LSA-based T.R.M. approach

Preprocessing delimits each sentence by punctuation such as “。”, “，”, “？”, and “！”. It also segments a sentence into keywords with a toolkit, named *AutoTag* [5]. *Semantic Model Analysis* represents the input document as a word-by-sentence matrix and reconstructs a semantic matrix via singular value decomposition (SVD) and dimension reduction. *Text Relationship Map Construction* constructs a text relationship map based on semantic sentence representations derived from the semantic matrix. *Sentence Selection* establishes a global bushy path according to the map and selects important sentences to compose a summary.

4.2 Semantic Sentence/Word Representations

Let D be a document, W ($|W|=M$) be a set of keywords, and S ($|S|=N$) be a set of the sentences in D. A *word-by-sentence* matrix, A, is constructed as follows, where S_i indicates a sentence and W_i indicates a keyword. In our implementation, only nouns and verbs are included in the word-by-sentence matrix.

$$A = \begin{array}{c|cccc} & S_1 & S_2 & \cdots & S_N \\ \hline W_1 & a_{11} & a_{12} & \cdots & a_{1N} \\ W_2 & a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ W_M & a_{M1} & a_{M2} & \cdots & a_{MN} \end{array}$$

In A, a_{ij} is defined as Equation 8, where L_{ij} is the local weight of W_i in S_j, and G_i is the global weight of W_i in D. L_{ij} is defined as $L_{ij}=log(1+\frac{c_{ij}}{n_j})$, and G_i is defined as $G_i=1-E_i$, where c_{ij} is the frequency of W_i occurring in S_j, n_j is the number of words in S_j, and E_i is the normalized entropy of W_i [4].

$$a_{ij} = L_{ij} \times G_i \ . \tag{8}$$

We then perform a *singular value decomposition* (SVD) to A. The SVD of A is defined as $A = USV^T$, where U is a $M \times N$ matrix of left singular vectors, S is a $N \times N$ diagonal matrix of singular values, and V is a $N \times N$ matrix of right singular vectors.

Then the process of *dimension reduction* is applied to S by deleting a few entries in it, and a new matrix, A', is reconstructed by multiplying the three component-matrixes. A' is defined as Equation 9, where S' is a semantic space that derives latent semantic structures from A, $U'=[u_i']$ is a matrix of left singular vectors whose ith vector u_i' represents W_i in S', and $V'=[v_j']$ is a matrix of right singular vectors whose jth vector v_j' represents S_j in S'.

$$A \approx A' = U'S'V'^T \ . \tag{9}$$

In A', each column denotes the *semantic sentence representation*, and each row denotes the *semantic word representation.*

4.3 Summary Generation

Salton et al. (1997) used a text relationship map to represent the document and then generated the corresponding summary according to the map [19]. One problem of their map is the lack of the *type* or the *context* of a link. To consider the context of a link, we combine the map and the above-mentioned semantic sentence representations to promote text summarization from keyword-level analysis to semantic-level analysis.

In our paper, a sentence S_k is represented by the corresponding semantic sentence representation mentioned in Section 4.2 instead of the original keyword-frequency vectors. The similarity between a pair of sentences S_i and S_j is evaluated to see if they are semantically related. The similarity is defined as Equation 10. The significance of a sentence is measured by counting the number of links that it has. A global bushy path [19] is established by arranging the *n* bushiest sentences in the order that they appear in the original document. Finally, a designated number of sentences are selected from the global bushy path to generate a summary.

$$Sim\left(\vec{S_i}, \vec{S_j}\right) = \frac{\vec{S_i} \cdot \vec{S_j}}{\left|\vec{S_i}\right|\left|\vec{S_j}\right|} . \tag{10}$$

5 Evaluation

In this section, we report our preliminary experimental results.

5.1 Data Corpus

We collected 100 articles on politics from the *New Taiwan Weekly* [20] and partitioned the collection into five sub-collections, named *Set 1*, *Set 2*, …, *Set 5* respectively. Table 1 shows the statistical information of the data corpus.

Table 1. Statistical information of the data corpus

Document Statistics	***Set 1***	***Set 2***	***Set 3***	***Set 4***	***Set 5***
Documents per collection	20	20	20	20	20
Sentences per document	27.5	24.8	26.7	31.5	26.4
Sentences per manual summary	8.8	8.0	8.5	9.8	8.4
Manual compression ratio per document	32%	32%	32%	31%	32%

5.2 Evaluation Methods

We used *recall* and *precision* to judge the coverage of both the manual and machine-generated summaries. Assuming that *T* is the manual summary and *S* is the machine-generated summary, the measurements are defined as follows:

$$Precision = \frac{|S \cap T|}{|S|} \qquad Recall = \frac{|S \cap T|}{|T|}$$

Since we set the length of the machine-generated summary to the length of the manual summary (i.e. compression ratio is about 30%), the values of precision and recall were the same; therefore, in the following, only recalls are listed.

5.3 Modified Corpus-Based Approach

In this experiment, we measured the performance of our first approach in the way called *K. Cross-Validation* [9]. In each step, one collection is chosen as the test corpus, and the other collections are the training corpus. To compare the performance of the original corpus-based approach and ours, each feature weight was set to be 1. Table 2 shows the performance of our approach.

Table 3 shows the effect of different features. (O. means *Original* and M. means *Modified*.) The result shows that for any two sentences, their significance differs because of their positions. It also shows that reshaping word units may achieve higher accuracy of keyword importance. In addition, the best combination of features for our modified approach was *Position+Positive Keyword+Resemblance to the Title+Centrality*; that is, without the “Negative Keyword” feature. Table 4 shows the performance when “*Negative Keyword*” is not considered. It can be seen that *Modified* outperforms *Original* by about 5.5% on average.

Table 5 lists the feature weights obtained by the genetic algorithm (GA). This table also lists the performance when the feature weights are applied to the training corpus. Table 6 shows the performance when the feature weights in Table 5 are applied to the test corpus (i.e. *Modified+GA*). It can be seen that *Modified+GA* outperforms *Modified* by about 7.4% on average. This illustrates the benefit of employing the genetic algorithm in training. The main advantage is to provide a preliminary analysis of the corpus and provide a means to fine tune the score function.

Table 2. Performance comparison of *Original* vs. *Modified* (All features considered)

	Set 1	***Set 2***	***Set 3***	***Set 4***	***Set 5***	**Avg.**
Original	0.2746	0.3700	0.2769	0.2633	0.2419	0.2853
Modified	0.2684	0.3772	0.2841	0.2574	0.2478	0.2870
Improvement	-2.3%	1.9%	2.6%	-2.2%	2.4%	0.6%

Table 3. Influence of each feature of *Original* vs. *Modified*

	f_1		f_2		f_3		F_4		f_5	
	O.	**M.**	**O.**	**M.**	**O.**	**M.**	**O.**	**M.**	**O.**	**M.**
Set 1	0.44	0.48	0.46	0.48	0.20	0.19	0.43	0.44	0.45	0.48
Set 2	0.46	0.49	0.36	0.39	0.30	0.28	0.42	0.45	0.39	0.40
Set 3	0.46	0.48	0.44	0.42	0.23	0.20	0.36	0.37	0.47	0.52
Set 4	0.48	0.50	0.49	0.50	0.17	0.18	0.46	0.46	0.48	0.50
Set 5	0.43	0.46	0.54	0.54	0.17	0.18	0.38	0.39	0.50	0.52
Avg.	0.45	0.48	0.46	0.47	0.21	0.21	0.41	0.42	0.46	0.48

Table 4. Performance comparison of *Original* vs. *Modified* (Minus *Negative Keyword*“)

	Set 1	***Set 2***	***Set 3***	***Set 4***	***Set 5***	**Avg.**
Original	0.4647	0.3799	0.4191	0.5142	0.5149	0.4586
Modified	0.4906	0.4028	0.4491	0.5348	0.5410	0.4837
Improvement	5.6%	6.0%	4.7%	4.0%	5.1%	5.5%

Table 5. Feature weights obtained by the genetic algorithm (Minus "*Negative Keyword*")

	f_1	F_2	f_3	f_4	F_5	**Fitness (Recall)**
Combination 1	0.926	0.013	0.000	0.359	0.002	0.7841
Combination 2	0.867	0.013	0.000	0.689	0.011	0.7875
Combination 3	0.996	0.013	0.000	0.401	0.025	0.7674
Combination 4	0.981	0.021	0.000	0.527	0.004	0.7782
Combination 5	0.875	0.012	0.000	0.581	0.022	0.7746

Table 6. Performance comparison of *Modified* vs. *Modified+GA*

	Set 1	***Set 2***	***Set 3***	***Set 4***	***Set 5***	**Avg.**
Modified	0.4906	0.4028	0.4491	0.5348	0.5410	0.4837
Modified+GA	0.5556	0.4790	0.4604	0.5376	0.5655	0.5196
Improvement	13.2%	18.9%	2.5%	0.5%	4.5%	7.4%

5.4 LSA-Based T.R.M. Approach

In this experiment, the feasibility of applying LSA to text summarization was evaluated. For different test corpuses, the dimension reduction ratios differ; for example, the best ratio for *Set 1* is 0.65 (i.e. if the rank of the singular value matrix is n, then only $0.65n$ is kept for semantic matrix reconstruction). The average ratio is about 0.64. Table 7 shows the performance of our approach. On average, *LSA-based T.R.M.* outperforms *Keyword-based T.R.M.* [19] by about 12.9%. To sum up, with the best reduction ratio (see Table 8 for the impact of different ratios), LSA can be used to promote text summarization from keyword-level analysis to semantic-level analysis.

Table 7. Performance comparison of *Keyword-based T.R.M.* vs. *LSA-based T.R.M.*

	LSA-based T.R.M.	***Keyword-based T.R.M.***	**Improvement**
Set 1	0.4616 (ratio = 0.65)	0.3425	34.8%
Set 2	0.4005 (ratio = 0.45)	0.3817	4.5%
Set 3	0.4567 (ratio = 0.80)	0.4469	2.2%
Set 4	0.4657 (ratio = 0.65)	0.4276	9.6%
Set 5	0.4943 (ratio = 0.65)	0.4201	17.7%
Avg.	0.4558 (ratio = 0.64)	0.4038	12.9%

Table 8. Influence of different dimension reduction ratios

Ratio	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
Set 1	0.323	0.314	0.334	0.391	0.403	0.424	0.432	0.427	0.378
Set 2	0.296	0.321	0.318	0.384	0.380	0.344	0.359	0.374	0.309
Set 3	0.369	0.360	0.366	0.412	0.399	0.431	0.451	0.455	0.431
Set 4	0.290	0.381	0.410	0.402	0.417	0.431	0.434	0.379	0.392
Set 5	0.353	0.381	0.404	0.444	0.486	0.474	0.450	0.479	0.380

Comparing our two proposed approaches, the performance of our LSA-based T.R.M. approach is very close to that of our corpus-based approach (Tables 6 and 7). The main advantage of the LSA-based approach over the modified corpus-based approach is that it involves a single document, and hence needs no preprocessing and is easy to implement.

The effect of LSA on text summarization is illustrated with an example. The precedent [1] means that the following sentence belongs to the manual summary, the precedent [2] means a summary sentence created by *Keyword-based T.R.M.*, and precedent [3] means a summary sentence created by *LSA-based T.R.M*. Table 9 shows text relationship maps created by *Keyword-based T.R.M.* and *LSA-based T.R.M.* respectively.

Topic

前總統夫人曾文惠出現在台北地方法庭

Content

[1,2,3]<P1S1>三月四日一大早約九點出頭，前總統夫人曾文惠在女兒李安妮與隨扈的護送下，出現在台北地方法庭。[2,3]<P1S2>在出發之前，前總統李登輝才對曾文惠表示了精神上的完全支持，但是她還是抵擋不住硬吞下眼淚的那種心情。

[1,2,3]<P2S1>台灣有史以來，第一次出現前第一夫人到法院出庭的情況，曾文惠臉上沒有面對群眾時慣有的那種溫暖笑容，而是勉強擠出淺淺的笑，低著頭快速地進入法庭。<P2S2>只有在步出法庭時，看到熱情的支持群眾，她才露出親切溫柔的笑臉。

<P3S1>許多人都還記得，當然，李登輝一家人也都深深地記得。[1,3]<P3S2>兩年前總統大選後的那幾天，許多「國民黨人士」包圍國民黨中央黨部，在民眾情緒激憤，要求李登輝下台的時候，謝啓大在宣傳車上，對著底下的群眾喊著「曾文惠帶了八千五百萬美金逃到美國」。

<P4S1>接下來，前立委馮滬祥以及前僑務委員戴錡更召開記者會，提出洋洋灑灑的「證據」，公開指稱曾文惠搭乘長榮航空，私運八千五百萬美元到美國，被美方拒絕入境，又緊急搭華航班機運回美元，於是引來了所謂的「八千五百萬元美金運送風波」。

[1]<P5S1>小女兒李安妮不甘曾文惠被如此惡意誹謗，建議曾文惠自訴謝啓大等三人涉嫌誹謗，並求償三億元賠償。[1,2,3]<P5S2>但是，法官出身的謝啓大深闇司法，第一次出庭就採取反擊，反控曾文惠誣告，也要求三億元賠償，並且要求曾文惠出庭，也使得曾文惠必須在三月四日出庭應訊。

<P6S1>當天，曾文惠進入台北地院的北大門時，離開庭時間還有約半個小時，她快速地走上樓梯進入休息室，並準時出現在位於二樓的第七法庭。[2]<P6S2>經過冗長的庭訊過程，從上午九點四十分開庭到中午一點休息，曾文惠完全沒有發言。<P6S3>經過短暫的休息之後，曾文惠才站在法庭前接受法官的詢問，否認運美金赴美。

[1,2]<P7S1>在經過身體與精神的雙重煎熬之下，下午三點多，曾文惠終於承受不住心裡的委屈，趴在桌上偷偷地落淚，並在李安妮的攙扶下暫時離開法庭。<P7S2>在庭訊的過程中，曾文惠也不禁用紙張寫下她的心情，「上帝創造人的眼淚是流下來的，我的眼淚卻是吞進去的」。

<P8S1>實際上，基於對司法的尊重，曾文惠與家人也完全不願意對這件官司發表談話。<P8S2>而儘管曾文惠的高中校友鄭玉麗，曾經在二○○○年三月二十二日下午打了通電話給她，並聊了將近半個小時，但基於自己沒有舉證責任的原則之下，曾文惠也不願鄭玉麗出面作證。

[1]<P9S1>對曾文惠而言，這場官司是一種捍衛自己尊嚴的官司。<P9S2>看著老妻受到這麼大的委屈，李登輝心底絕對是相當心疼的。

In this example, *LSA-based T.R.M.* had a recall of 67% and *Keyword-based T.R.M.* a recall of 50%. P3S2 and P4S1, two semantically similar sentences, are used here to demonstrate the superiority of *LSA-based T.R.M.*. The similarity of the two sentences computed by *Keyword-based T.R.M.* was 0.0831 since only a few keywords (“八千五百萬”, “美金”, “美國”, and “曾文惠”) overlap. However, their similarity computed by *LSA-based T.R.M.* was 0.8604. This explains why LSA can derive more precise semantic meanings from a text.

Table 9. Text relationship maps created by *LSA-based T.R.M.* and *Keyword-based T.R.M.*

	LSA-based T.R.M.		*Keyword-based T.R.M.*	
	Connected Sentences	**Link**	**Connected Sentences**	**Link**
P1S1	P2S1, P5S2, P6S2, P7S1	4	P1S1, P2S1, P4S1, P5S1, P6S2, P7S1	6
P1S2	P2S1, P3S2, P7S1, P7S2	4	P1S1, P2S1, P2S2, P7S1, P7S2	5
P2S1	P1S1, P1S2, P2S2, P3S2, P4S1, P5S2, P6S1	7	P1S1, P1S2, P2S2, P5S2, P6S1	5
P2S2	P2S1	1	P1S2, P2S1, P6S1	3
P3S1	P3S2	1	P3S2	1
P3S2	P1S1, P2S1, P3S1, P4S1, P5S1, P5S2, P9S2	7	P3S1, P4S1, P5S2	3
P4S1	P2S1, P3S2, P5S2, P6S3	4	P3S2, P6S3	2
P5S1	P3S2, P5S2, P7S1	3	P1S1, P5S2, P7S1	3
P5S2	P1S1, P2S1, P3S2, P4S1, P5S1	5	P1S1, P2S1, P3S2, P5S1	4
P6S1	P2S1, P8S2	2	P2S1, P2S2, P8S2	3
P6S2	P1S1	1	P1S1, P6S3, P7S1, P7S2	4
P6S3	P4S1, P7S1	2	P4S1, P6S2	2
P7S1	P1S1, P1S2, P5S1, P6S3, P8S2, P9S2	6	P1S1, P1S2, P5S1, P6S2, P8S2, P9S2	6
P7S2	P1S2	1	P1S2, P6S2	2
P8S1	P9S1	1	P9S1	1
P8S2	P6S1, P7S1	2	P6S1, P7S1	2
P9S1	P8S1	1	P8S1	1
P9S2	P3S2, P7S1	2	P7S1	1

6 Conclusion and Future Work

In this paper, we proposed two text summarization approaches: the *Modified Corpus-based Approach* and the *LSA-based T.R.M. Approach*. Three new ideas are introduced in the first approach: 1) to employ ranked position to emphasize the significance of sentence position, 2) to reshape word units to achieve higher accuracy of keyword importance, and 3) to train the score function (using the genetic algorithm) to take the properties of the data corpus into account. The second approach uses latent semantic analysis (LSA) to derive the semantic matrix of a document, and uses semantic sentence representations to construct a semantic text relationship map. The two novel approaches were measured on a data corpus composed of 100 political articles, and when the compression was 30%, average recalls of 52.0% and 45.6% were achieved respectively.

In future, we plan to study how to construct a knowledge model, such as *word chains* [3], via semantic word representations. In this manner, we expect to propose more useful methods to promote traditional text summarization from keyword-level analysis to semantic-level analysis.

References

1. Aone, C., Okurowski, M.E., Gorlinsky, J., Larsen, B.: A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In: Mani, I., Maybury, M. (eds.): Advances in Automated Text Summarization. MIT Press (1999) 71-80

2. Azzam, S., Humphreys, K., Gaizauskas, R.: Using Coreference Chains for Text Summarization. Processing of the ACL'99 Workshop on Coreference and its Applications. ACL, Baltimore (1999)
3. Barzilay, R., Elhadad, M.: Using Lexical Chains for Text Summarization. Processing of the Workshop on Intelligent Scalable Text Summarization. (1997)
4. Bellegarda, J.R., Butzberger, J.W., Chow, Y.L.: A Novel Word Clustering Algorithm Based on Latent Semantic Analysis. Conference on Acoustics, Speech, and Signal Processing, Vol. 1. IEEE (1996) 172-175
5. CKIP AutoTag. Available at http://godel.iis.sinica.edu.tw/CKIP
6. Edmundson, H.P.: New Methods in Automatic Extracting. In: Mani, I., Maybury, M. (eds.): Advances in Automated Text Summarization. MIT Press (1999) 23-42
7. Gong, Y., Liu, X.: Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. SIGIR. ACM, New Orleans Louisiana (2001)
8. Habn, U., Mani, I.: The Challenge of Automatic Summarization. Computer, Vol. 33, No. 2000. IEEE (2000) 29-36
9. Han, J., Kember, M.: In Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2001)
10. Hovy, E., Lin, C.Y.: Automated Text Summarization in SUMMARIST. In: Mani, I., Maybury, M. (eds.): Advances in Automated Text Summarization. MIT Press (1999) 81-94
11. Kim, J.H., Kim, J.H., Hwang, D.: Korean Text Summarization Using an Aggregative Similarity. Processing of the 5th International Workshop on Information Retrieval with Asian Languages. ACM (2000)
12. Kowalski, G. (ed.): Information Retrieval Systems: Theory and Implementation. Kluwer Academic Publishers (1997)
13. Kupiec, J., Pedersen, J., Chen, F.: A Trainable Document Summarizer. SIGIR. ACM, Seattle Washington (1995)
14. Landauer, T.K., Foltz, P.W., Laham, D.: An Introduction to Latent Semantic Analysis. Discourse Processes, Vol. 25. (1998) 259-284
15. Lin, C.Y.: Training a Selection Function for Extraction. CIKM. ACM, Kansas City (1999)
16. Mani, I., Maybury, M. (eds.): Advances in Automated Text Summarization. MIT Press (1999)
17. McKeown, K.R., Radev, D.R.: Generating Summaries of Multiple News Articles. SIGIR. ACM, Seattle Washington (1995) 74-82
18. Myaeng, S.H., Jang, D.: Development and Evaluation of a Statistical Based Document System. In: Mani, I., Maybury, M. (eds.): Advances in Automated Text Summarization. MIT Press (1999) 61-70
19. Salton, G., Singhal, A., Mitra, M., Buckley, C.: Automatic Text Structuring and Summarization. Information Processing & Management, Vol. 33, No. 2. Elsevier (1997) 193-207
20. New Taiwan Weekly. Available at http://www.newtaiwan.com.tw
21. WordNet (A Lexical Database for the English Language). Available at http://www.cogsci.princeton.edu/~wn/

A Linear Text Classification Algorithm Based on Category Relevance Factors

Zhi-Hong Deng[1], Shi-Wei Tang[2], Dong-Qing Yang[1], Ming Zhang[1], Xiao-Bin Wu[1], and Meng Yang[1]

[1]Department of Computer Science and Technology，Peking University，Beijing 100871
zhdeng@rdb01.pku.edu.cn
[2]Center for Information Sciences, National Laboratory on Machine Perception,
Peking University, Beijing 100871
tsw@pku.edu.cn

Abstract. In this paper, we present a linear text classification algorithm called CRF. By using category relevance factors, CRF computes the feature vectors of training documents belonging to the same category. Based on these feature vectors, CRF induces the profile vector of each category. For new unlabelled documents, CRF adopts a modified cosine measure to obtain similarities between these documents and categories and assigns them to categories that have the biggest similarity scores. In CRF, it is profile vectors not vectors of all training documents that join in computing the similarities between documents and categories. We evaluated our algorithm on a subset of Reuters-21578 and 20_newsgroups text collections and compared it against *k*-NN and SVM. Experimental results show that CRF outperforms *k*-NN and is competitive with SVM.

1 Introduction

In recent years we have seen an exponential growth in the volume of text documents available on the Internet. These Web documents contain rich textual information, but they are so numerous that users find it difficult to obtain useful information from them. This has led to a great deal of interest in developing efficient approaches to organizing these huge resources and assist users in searching the Web. Automatic text classification, which is the task of assigning natural language texts to predefined categories based on their content, is an important research field that can help both in organizing and in finding information in these resources.

Text classification presents many unique challenges and difficulties due to the large number of training cases and features present in the data set. This has led to the development of a number of text classification algorithms, which address these challenges to different degrees. These algorithms include k-NN [1], Naïve Bayes [2], decision tree [3], neural network [4], SVM [5], and Linear Least Squares Fit [6].

In this paper, we present a new linear text classification algorithm based on category relevance factors, which represent the discriminating power of features to categories. This algorithm is also called CRF algorithm. For each category, CRF algorithm first computes all category relevance factors of this category, and then uses

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 88-98, 2002.

these factors to obtain feature vectors of training documents belonging to this category. Based on these feature vectors, the algorithm induces the profile vector of each category, which is the summary of a category. For new unlabelled documents, this algorithm adopts a modified cosine measure to obtain the similarities of these documents to categories and assigns them to categories that have the biggest similarity scores. Experiments presented in Section 5.3 show that it outperforms k-NN and is competitive with SVM.

The remainder of the paper is organized as follows. Section 2 describes k-NN and SVM, which are state of the art of classification algorithms. Section 3 describes category relevance factors and profile vectors of categories. Section 4 presents the CRF algorithm in detail. Section 5 experimentally evaluates the CRF algorithm on some text collections. Section 6 summarizes the paper and points out future research.

2 State-of-the-Art Text Classification Algorithms

In the following sections, we briefly describe *k* nearest neighbor and support vector machines used in our study. These two methods have the best classification performance among current methods according to [7, 8, 9].

2.1 *K* Nearest Neighbor

K nearest neighbor (*k*-NN) classification is an instance-based learning algorithm that has been applied to text classification since the early days of research [1, 10, 11]. In this classification paradigm, the new unlabelled document *d* is assigned to the category c_i if c_i has the biggest similarity score to *d* among all categories. The similarity score of documents *d* to a category c_j is computed as:

$$s(d,c_j)=\sum_{d_i\in k-NN} sim(d,d_i)y(d_i,c_j)\ . \tag{1}$$

sim(*d*, d_i) is the similarity between the document *d* and the training document d_i; $d_i \in k$-NN stands for that d_i is one of the k nearest neighbors to *d* in the light of the function *sim*(); $y(d_i, c_j) \in \{0,1\}$ is the classification for document d_i with respect to category c_j ($y(d_i, c_j) =1$ for YES, and $y(d_i, c_j) = 0$ for NO). Finally, based on these similarity calculated from formula (1), the category of document is assigned by (2), where $c_1, \ldots, c_m$ are the predefined categories.

$$\arg\max_{j=1,\ldots,m}(s(d,c_j)) \tag{2}$$

For *k*-NN, we adopt *tf*idf* as feature weight scheme and use vector-space model [12] to stand for documents. The similarity of two documents is measured by cosine similarity instead of Euclidean distance. Given two documents d_1 and d_2, their corresponding weighted feature vectors are $V_1 = (w_{11}, \ldots, w_{1n})$ and $V_1 = (w_{21}, \ldots, w_{2n})$. The similarity between d_1 and d_2 is defined as:

$$sim(d_1, d_2) = \cos(V_1, V_2) = \frac{V_1 \bullet V_2}{\|V_1\|_2 \|V_2\|_2} = \frac{\sum_{i=1}^{n} w_{1i} \times w_{2i}}{\sqrt{\sum_{i=1}^{n} w_{1i}^2}\sqrt{\sum_{i=1}^{n} w_{2i}^2}} . \quad \textbf{(3)}$$

2.2 Support Vector Machines

Support vector machines (SVM) is a relatively new learning method initially introduced by Vapnik in 1995 for two-class pattern recognition problems using the Structural Risk Minimization principle [13]. Given a training set containing two kinds of data (one for positive examples, the other for negative examples), which is linearly separable in vector space, this method finds the decision hyper-plane that best separated positive and negative data points in the training set. The problem searching the best decision hyper-plane can be solved using quadratic programming techniques [14]. SVM can also extend its applicability to linearly non-separable data sets by either adopting soft margin hyper-planes, or by mapping the original data vectors into a higher dimensional space in which the data points are linearly separable.

Of course, SVM is suitable not only for two-class classification but also for *m*-class classification, where *m* is more than two. For *m*-class classification, a simple and natural way is to use *m* hyper-planes generated by SVM, each of these hyper-planes is a decision rule for one category. Given $C = \{c_1, \ldots, c_m\}$, $CS = \{cs_1, \ldots, cs_m\}$, where cs_i stands for a set of training documents belonging to c_i. For $c_i \in C$, we set positive set $c_i^+ = cs_i$ and negative set $c_i^- = \cup\, cs_j$ $(j \neq i)$. Using c_i^+ and c_i^- as input, we can generate a decision rule R_i for c_i by SVM. For all categories, we obtain *m* rules $R_1, R_2, \ldots, R_m$.

3 Profile Vectors Based on Category Relevance Factors

The essential problem of text classification is how to find profiles representing the predefined categories. In *k*-NN, all training documents belonging to a category are used to describe the category, and we predict the categories of new documents using the whole training set. In SVM, we find a hyper-plane for each category, and utilize these hyper-planes for deciding categories that new documents belong to. Although *k*-NN does not need the learning classifier phrase, it involves too much calculation due to utilizing all training documents in classifying new documents. Classifying new documents is easy to SVM classifier, but learning classifier phrase, which aims at finding hyper-planes, is hard and time-consuming. In this section, we propose a simple and efficient method for inducing category profiles based on category relevance factors.

3.1 Category Relevance Factor

Category relevance factor (*CRF*) stands for the discriminating power of features to categories. With no loss of generality, we suppose that $C = \{c_1, \ldots, c_m\}$ represents the set of predefined categories, $CS = \{cs_1, \ldots, cs_m\}$ represents the set of training document sets, and $F = \{f_1, \ldots, f_n\}$ represents the feature set selected from all training documents. The category relevance factor of f_i to c_j is defined as:

$$CRF(f_i, c_j) = \begin{cases} \max\left(\bigcup_{k=1}^{m} \{CRF(f_i, c_j), \left|CRF(f_i, c_j)\right| < \infty\} \cup \{\log \frac{\sum_{i=1}^{m} |cs_i|}{m}\} \right) \cdots X' = 0 \\ \log\left(\frac{X/Y}{X'/Y'} \right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots X' \neq 0 \cap X \neq 0 \\ \min\left(\bigcup_{k=1}^{m} \{CRF(f_i, c_j), \left|CRF(f_i, c_j)\right| < \infty\} \cup \{-\log \frac{\sum_{i=1}^{m} |cs_i|}{m}\} \right) \cdots X = 0 \end{cases} . \tag{4}$$

$|cs_i|$ is the number of documents belonging to c_i; X is the number of documents that contain feature f_i and belong to category c_j; Y is the number of documents that belong to category c_j; $X'$is the number of documents that contain feature f_i and don't belong to category c_j; $Y'$is the number of documents that don't belong to category c_j. The definition of category relevance factor is derived from *TERMREL* representing "term relevance factor". In information retrieval, the product of *tf* (term frequency) and *TERMREL* is the theoretically optimal feature weight scheme [15], and experimental evidence [16] confirms it. The main difference between *CRF* and *TERMREL* is that the values of *CRF* are always limited while the values of *TERMREL* may be infinite.

The bigger the value of *CRF* of a feature to a category, the higher is the discriminating power of the feature with respect to the category. If feature f_i only occurs in training documents belonging to c_j, we deem that f_i possesses the highest ability for discriminating c_j and should be assigned the biggest value. On the contrary, if feature f_i doesn't occur in any training documents belonging to c_j, we may consider that f_i plays no role or a negative role in discriminating c_j and should be assigned the smallest value. It is obvious that the definition of *CRF* is rational and corresponds to the theory of statistics.

3.2 Profile Vectors

For category c_j, we can obtain n category relevance factors in terms of n features. These n category relevance factors constitute a vector $(CRF(f_1, c_j), CRF(f_2, c_j), \ldots, CRF(f_n, c_j))$, which is called category relevance factor vector of c_j and is abbreviated to CRF_j. CRF_j is also written as $(crf_{1j}, crf_{2j}, \ldots, crf_{nj})$, where crf_{ij} is the abbreviation of $CRF(f_i, c_j)$.

For training document d belonging to c_j, the feature vector V of d is written as $V = (v_1, v_2, \ldots, v_n)$, where v_i is obtained as follows:

$$v_i = tf_i(d) \times crf_{ij}, 1 \leq i \leq n \ . \tag{5}$$

$tf_i(d)$ can be obtained by formula (6), where $TF(f_i, d)$ is the number of times that feature f_i occurs in document d.

$$tf_i(d) = 0.5 + 0.5 \times \frac{TF(f_i, d)}{\max\{TF(f_1, d), TF(f_2, d), ..., TF(f_n, d)\}} \quad \textbf{(6)}$$

The idea behind the profile vector based on category relevance factor is very simple. The profile vector of a category is just a vector that integrates all feature vectors of documents belonging to the category. Given category c_i and $cs_i = \{V \mid V$ is the feature vector of a training document and the category of the training document is $c_i\}$, the category profile P_i of c_i is defined as formula (7), where $|cs_i|$ is the number of documents that belong to c_i. Formula (7) is nothing more than the vector obtained by averaging the weights of the various features present in the documents belonging to c_i.

$$P_i = \frac{1}{|cs_i|} \sum_{V \in cs_i} V \quad \textbf{(7)}$$

4 CRF Algorithm

The idea behind the CRF algorithm is very simple. For each category, we first compute its category relevance factor vector according to formula (4); then we compute the feature vectors of training documents belonging to the category in terms of formulas (5) and (6); last we obtain the category profile vector of the category in the light of formula (7). The category of a new document *d* is determined by the similarities between *d* and these category profile vectors. We assign *d* to the category with the highest similarity score.

There are two problems for CRF algorithm. One is how to compute the feature vector of a new document *d*. The other is how to measure the similarity. The first problem results from the lack of knowledge of the exact category of *d* -- we need predict it. Formula (5) is used to compute the feature vectors mentioned in section 3.2 using the training documents, whose categories are known in advance. Therefore, it cannot be applied to compute the feature vectors of unlabelled documents directly. To solve the first problem, we adopt a technique called supposition-test. Given $C = \{c_1, \ldots, c_m\}$, $P = \{P_1, \ldots, P_m\}$ (P_i is the category profile of c_i), and a new document *d*, supposition-test method first assumes that *d* belongs to c_i ($c_i \in C$) and computes the feature vector V_i of *d* according to formula (5); then it computes the similarity score s_score_i of V_i and category profile P_i. For all categories in *C*, we can obtain s_score_1, s_score_2, …, and s_score_m. If s_score_j is the biggest one among the *m* similarity scores, *d* is classified as category c_j.

The second problem is caused by the negative elements contained in the feature vector. Because the product of two negatives is positive, we may get an incorrect result if we adopt cosine (formula (3)) for computing similarity. For example, given three feature vectors $V_1 = (1, -2, 1)$, $V_2 = (0, -3, 0)$, $V_3 = (1, 0, 1)$, the similarity of V_1 and V_2 is 0.816 and the similarity of V_1 and V_2 is 0.577 according formula (3), which means that V_2 is more similar to V_1 than V_3. It is obvious that V_3 is closer to V_1 than V_2 because positive numbers stand for more discriminating power than negatives as mentioned in section 3.1. Therefore, we modify formula (3) as follows.

Given $V_1 = (v_{11}, v_{12}, \ldots, v_{1n})$ and $V_2 = (v_{21}, v_{22}, \ldots, v_{2n})$, V_1 represents document d_1 and V_2 represents d_2, d_1 and d_2 belong to category c, the category relevance factor vector of c is $CRF_c = (crf_1, crf_2, \ldots, crf_n)$. The similarity of d_1 and d_2 is defined as:

$$s\_score(d_1, d_2) = \frac{\sum_{i=1}^{n} sign(crf_i) \times v_{1i} \times v_{2i}}{\sqrt{\sum_{i=1}^{n} v_{1i}^2}\sqrt{\sum_{i=1}^{n} v_{2i}^2}} . \qquad \mathbf{(8)}$$

sign() is defined as:

$$sign(x) = \begin{cases} 1 \cdots\cdots\cdot x > 0 \\ 0 \cdots\cdots\cdot x = 0 \\ -1 \cdots\cdots x < 0 \end{cases} . \qquad \mathbf{(9)}$$

Given $P_c = (p_{c1}, p_{c2}, \ldots, p_{cn})$ and $V_1 = (v_{11}, v_{12}, \ldots, v_{1n})$, where P_c is the profile vector of c and V_1 is the feature vector of d_1 belonging to c. Similar to formula (8) that calculates the similarity of two document, the similarity of d_1 and c is defined as:

$$s\_score(d_1, P_c) = \frac{\sum_{i=1}^{n} sign(crf_i) \times v_{1i} \times p_{ci}}{\sqrt{\sum_{i=1}^{n} v_{1i}^2}\sqrt{\sum_{i=1}^{n} p_{ci}^2}} \circ \qquad \mathbf{(10)}$$

The CRF algorithm has two parts: the phase for learning classifiers and the phase for classifying new documents. For the sake of description, we label the former *CRF_Training* and the latter *CRF_Classifying*.

CRF_Training:

Input: training documents set $D = \cup D_i$, $1 \le i \le m$, D_i = {document $d \mid d$ belongs to category c_i }; feature set $F = \{f_1, f_2, \ldots, f_n\}$.
Output: category relevance factor vectors set $CRF = \{CRF_1, CRF_2, \ldots, CRF_m\}$, $P = \{P_1, P_2, \ldots, P_m\}$, where $CRF_i = (crf_{1i}, crf_{2i}, \ldots, crf_{ni})$ represents the category relevance factor vector of category c_i; P_i represents the category profile of c_i.
Step1. Set $CRF = \varnothing$, $P = \varnothing$.
Step2. For each c_i, compute the number of documents belonging to c_i. We label the number sum_i and the number of all training documents *SUM*. $SUM = \sum sum_i$, $1 \le i \le m$.
Step3. For each f_j, compute the number of documents that contain feature f_j and belong to c_i. The number is labeled as df_{ji}. Based on df_{ji}, we can obtain DF_j ($DF_j = \sum df_{ji}$, $1 \le i \le m$), which is the number of documents that contain feature f_j.
Step4. For i = 1 to m, do:

1. For each f_j, compute $CRF(f_j, c_i)$ according to formula (4) and generate $CRF_i = (crf_{1i}, crf_{2i}, \ldots, crf_{ni})$, which represents the category relevance factor vector of category c_i. Set $CRF = CRF \cup \{CRF_i\}$.

```
2. For each $d \in c_i$, compute its feature vector $V = (v_1, v_2, \ldots, v_n)$ according to formula (5) and formula (6).
3. According to formula (7), compute $P_i$, the category profile of $c_i$. Set $P = P \cup \{P_i\}$.
Step5. Output $CRF$ and $P$.
```

CRF_Classifying:

```
Input: $CRF = \{CRF_1, CRF_2, \ldots, CRF_m\}$, $P = \{P_1, P_2, \ldots, P_m\}$, feature set $F = \{f_1, f_2, \ldots, f_n\}$ and unlabelled document $d_{proc}$.
Output: the category of $d_{proc}$.
Step1. Set $simvalset = \varnothing$, where $simvalset$ holds similarity scores of $d_{proc}$ and categories. For each $f_j$, compute the $tf_j = tf_j(d_{proc})$ according to formula (6).
Step2. For $i$ = 1 to $m$, do:
    1. According to formula (5), compute feature vector $V_{proc}$ of $d_{proc}$ by using $\{tf_j\}$ and $CRF_i$.
    2. According to formula (10), compute similarity score $s\_score_i$ of $V_{proc}$ and $P_i$. Set $simvalset = simvalset \cup \{s\_score_i\}$.
Step3. find $s\_score_x$, which is the biggest one in $simvalset$, and output $c_x$ as the category of $d_{proc}$.
```

The computational complexity of the training phase of the CRF algorithm is linear on the number of documents and the number of features in the training documents set. The amount of time required to classify a new documents d is at most $O(mn)$, where m is the number of categories and n is the number of features. Thus, the overall computational complexity of this algorithm is very low, and is identical to fast document classifications such as Naïve Bayesian.

5 Experimental Results

We evaluated the performance of the CRF text classification algorithm by comparing it against k-NN algorithm and SVM algorithm on three text collections. The k-NN classifier constructed by us and SVM^{light} [17] classifier adopt the *tf*idf* scheme as feature weighting scheme[1] and vectors-space as representation of the documents. To obtain the best performance, we run the k-NN classifier repeatedly using the number of neighbor k to a variety of values in the light of text collections. For SVM, we tested the linear and non-linear models provided by SVM^{light}, and obtained a better result with the linear SVM models. This replicates the findings of Yang [9]. Therefore, we use only the result from linear SVM in the paper.

[1] Joachims used *tf*idf* scheme for SVM classifier constructed by SVM^{light} in [7]. Therefore, we also adopt *tf*idf* scheme.

5.1 Text Collections

Text Collections used to evaluate the classification algorithms are taken from 20_newsgroups [18] and Reuters-21578 text categorization test collection Distribution 1.1 [19]. 20_newsgroups contains about 19,997 articles evenly divided among 20 UseNet discussion groups. Each document in 20_newsgroups contains a header section which has a line labeled Newsgroup. The line describes what category the document belongs to. To test algorithms rationally, we skipped the UseNet headers in documents except for the subject line. Because 20_newsgroups text collection is too huge to be dealt with, we divided it into two text collections randomly without overlapping. For the sake of discussion, we labelled one Newsgroup_1 and the other Newsgroup_2. Each of these collections contains 10 categories and each category contains 1000 documents except for soc.religion.Christian, which has 997 documents and belongs to text collection Newsgroup_1. The first 700 documents (697 for soc.religion.Christian) in each category were used as training examples and the last 300 documents were used as test examples. For this paper, we extracted documents from Reuters-21578, with assigning each document to only one category and adopted categories that contain more than 100 training documents. This extracted text collection is named Reuters_100. It is well known that it is not effective to use a small number of training documents for learning classification models.

Note that for all test collections, we used a stop-list to remove common words, and the words were stemmed using Porter's suffix-stripping algorithm [20]. Furthermore, according to [21] we also skipped rare frequency words that occur in fewer than three documents. The characteristics of these document collections used in our experiments are summarized in Table 1.

Table 1. Summary of text collections

	Newsgroup_1	Newsgroup_2	Reuters_100
Training documents	6997	7000	6246
Test documents	3000	3000	2540
Categories	10	10	10
Features	17138	16954	6020

5.2 Evaluation Techniques

To evaluate the performance of the classification, we use the standard recall, precision and F_1 measures. Recall (r) is defined to be the ratio of correct positive predictions by the system divided by the total number of positive examples. Precision (p) is defined as the ratio of correct positive predictions by the system divided by the total number of positive predictions by the system. Recall and precision reflect different aspects of classification performance. Usually, if one of the two measures is increasing, the other will decrease. To obtain a better measure describing performance. F_1, which was first introduced by van Rijsbergen [22], is adopted. It combines recall and precision as follows:

$$F_1 = \frac{2rp}{r + p} \, . \tag{11}$$

To obtain an overview of the performance on all categories, we adopt *micro-averaging* F_1, which is widely used in cross-method comparisons, as standard measure for classification performance according to [8, 23].

5.3 Classification Performance

We applied statistical feature selection at a preprocessing stage for each classifier using *information gain* (IG) criterion. According to the value of *information gain* of features, we chose the top $p\%$ of all features listed in Table 1 as feature set, while p belongs to {10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. These 10 different feature sets were tested. For each feature set, k ∈ {30, 45, 60, 75, 90} for the *k*-NN classifier were tried.

Table 2. The best classification performance achieved by *k*-NN, SVM, CRF

	Reuters_100	Newsgroup_1	Newsgroup_2
CRF	0.892	0.881	0.859
SVM	0.917	0.849	0.829
k-NN	0.888	0.811	0.64

Table 2 summarizes the best classification performance achieved by the three algorithms, and shows that CRF outperformed *k*-NN on all text collections, and outperformed SVM on Newsgroup_1 and Newsgroup_2. On Reuters_100, CRF achieved the highest performance with p ∈ {20, 90, 100}; SVM achieved the highest performance with p = 20; *k*-NN achieved the highest performance with p ∈ {80, 90, 100} and k = 90. On Newsgroup_1, CRF achieved the highest performance with p = 100; SVM achieved the highest performance with p = 20; *k*-NN achieved the highest performance on with p = 100 and k = 90. On Newsgroup_2, CRF achieved the highest performance with p ∈ {80, 90, 100}; SVM achieved the highest performance with p = 10; *k*-NN achieved the highest performance with p = 100 and k = 75.

6 Conclusions

In this paper, we presented a linear text classification algorithm that learns the discriminating power of features and utilizes them in inducing the profile vectors of categories and classifying new documents. As our experimental results on three text collections have shown, our algorithm is very effective in text classification. It has significant advantage over *k*-NN and is competitive with SVM. Moreover, the complexity of this algorithm is very low and it is easy to implement. All these make CRF a very promising method for text classification.

In future work, we will investigate other formulas, which are suitable for computing the discriminating power of features, such as mutual information (MI) [24], χ^2

statistic (CHI) [24], and odds ratio [25], and apply them to text classification. Further more, we will explore calculation models of the discriminating power of features under insufficient number of training documents.

Acknowledgements. This research was supported by the National Grand Fundamental Research 973 Program of China under grant No. G1999032705. Opinions, findings, conclusions and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsor. We are also grateful to Dr. Liang-Huai Yang and three anonymous reviewers for their comments.

References

1. Y. Yang. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'94*), pages 13-22, 1994.
2. A. McCallum and K. Nigam. A comparison of event models for naïve bayes text classification. In *AAA-98 Workshop on Learning for Text Categorization*, 1998.
3. C. Apte, F. Damerau, and S. Weiss. Text mining with decision rules and decision trees. In *proceedings of Conference on Automated Learning and Discovery, Workshop 6: Learning from Text and the Web*, 1998.
4. H.T. Ng, W.B. Goh, and K.L. Low. Feature selection, perceptron learning, and a usability case study for text categorization. In *20th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'97*), pages 67-73, 1997.
5. S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management*, pages 148-155, 1998.
6. Y. Yang and C.G. Chute. An example-based mapping method for text categorization and retrieval. *ACM Transaction on Information Systems* (*TOIS*), 12(3): 252-277, 1994.
7. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machines Learning* (*ECML*), pages 137-142, 1998.
8. Y. Yang. An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, 1(1/2): 67-88, 1999.
9. Y. Yang, X. Liu. A re-examination of text categorization methods. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'99*), pages 42-49, 1999.
10. B. Masand, G. Linoff, and D. Waltz. Classifying News Stories using Memory Based Reasoning. In *15th Annul International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'92*), pages 59-64, 1992.
11. M. Iwayama, T. Tokunaga. Cluster-Based Text Categorization: A Comparison of Category Search Strategies. In *18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'95*), pages 273-280, 1995.
12. G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, MA, 1989.
13. V. Vapnic. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.
14. C. Cortes and V. Vapnik. Support Vector networks. *Machine Learning*, 20: 273-297, 1995.
15. C. T. Yu, K. Lam, G. Salton. Term weighting in information retrieval using the term precision model. *Journal of the ACM*, 29(1): 152-170, 1982.
16. G. Salton, M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

17. T. Joachims, Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*, MIT-Press, 1999.
18. 20_newsgroups data set. http://www-2.cs.cmu.edu/~mccallum/bow/rainbow/.
19. D.D. Lewis. Reuters_21578 text categorization test collection. http://www.research.att.com /~lewis/reuters21578.html.
20. M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3): 130-137, 1980.
21. F. Sebastiani. A Tutorial on Automated Text Categorisation. In *Proceedings of the First Argentinean Symposium on Artificial Intelligence*, 7-35, 1999.
22. C.J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.
23. D.D. Lewis. Representation and Learning in Information Retrieval. Ph.D. dissertation, University of Massachusetts, USA, 1992.
24. Y. Yang, J.P. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of 14th International Conference on Machine Learning*, 412-420, 1997.
25. D. Mladenic, M. Grobelnik. Feature Selection for Classification Based on Text Hierarchy. In *Working notes of Learning from Text and the Web*, *Conference on Automated Learning and Discovery* (*CONALD'98*), 1998.

A Hierarchical Framework for Multi-document Summarization of Dissertation Abstracts

Christopher S.G. Khoo, Shiyan Ou, and Dion Hoe-Lian Goh

Division of Information Studies
School of Communication & Information
Nanyang Technological University
Singapore, 637718
{assgkhoo,pg00096125,ashlgoh}@ntu.edu.sg

Abstract. This paper reports initial work on developing methods for automatic generation of multi-document summaries of dissertation abstracts in a digital library. The focus is on automatically generating a summary of a set of dissertation abstracts retrieved in response to user query, and presenting the summary using a visualization method. A hierarchical variable-based framework for multi-document summarization of dissertation abstracts in sociology and psychology is presented. The framework makes use of macro-level and micro-level discourse structure of dissertation abstracts as well as cross-document structure. The micro-level structure of problem statements found in a sample of 50 dissertation abstracts was analyzed, and the common features found are described in the paper. A demonstration prototype with a tree-view interface for presenting multi-document abstracts has been implemented.

1 Introduction

This paper reports initial work on developing methods for automatic generation of multi-document summaries of dissertation abstracts in a digital library. Automatic summarization is an important function that should be available in large digital library systems, where the retrieval of too many documents will be a major problem for users. Information retrieval systems and Web search engines attempt to address the problem of information overload by ranking documents retrieved by their likelihood of relevance, and displaying their titles and short abstracts to give the user some indication of the document content. The abstract may be constructed by humans or automatically generated by extracting the first few lines of the document text or identifying the more important sentences in the document.

However, the user has patience to scan only a small number of document titles and abstracts, usually in the range of 10 to 30 [2,7]. To help the user identify relevant documents from a larger number of records, some search engines group the retrieved records into folders or categories (e.g. the Northern Light search engine at http://www.northernlight.com). These categories may be pre-created, and the records assigned to them by human indexers or by automatic categorization techniques. The categories may also be constructed dynamically by clustering the records retrieved (e.g. Grouper [9]).

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 99-110, 2002.

A related approach is to dynamically construct a multi-document summary of the documents retrieved. A multi-document summary has several advantages over single-document summaries. It provides a domain overview of the subject area - indicating common information across many documents, unique information in each document, and cross-document relationships (relationships between pieces of information in different documents), and can allow users to zoom in for more details on aspects of interest.

Research on multi-document summarization has focused mainly on using sentence extraction, identification of similarities and differences between documents, and discourse analysis. Most studies attempt to construct multi-document summaries by extracting important text units using statistical approaches, and analysis of discourse structure to distinguish central or nucleus information from satellite or peripheral information. Other studies extract information that is common or repeated in several documents plus selected unique information in individual documents. However, these existing summarization approaches do not pay much attention to the semantic content and semantic relations expressed within and across documents. Another problem is that different users have different information needs. Thus, an ideal multi-document summarization should provide different levels of detail for different aspects of the topic according to users' interests.

One promising approach is to outline the overall structure of a group of related documents to give users an overview of the topic covered by the documents first, and subsequently to allow users to zoom in on different sub-topics according to their interests. This is performed by analyzing the cross-document structure, which is a kind of discourse structure covering a group of related documents rather than a single document. A discourse structure is a representation of a text from a linguistic/language perspective, and includes all aspects of the internal organizational structure of a discourse ranging from microstructures (such as lexical items and grammatical structure) to macrostructures (such as topics or themes expressed indirectly in larger stretches of text or whole discourses) [1]. In a document set, the documents are related to each other and address the same or related topics. This allows us to extend the scope of discourse from one document to a document set and to analyze structure beyond the limitations of document boundaries.

One of the most popular discourse theories is the Rhetorical Structure Theory (RST) [3]. The rhetorical parsing algorithm proposed by Marcu [4] is one way to derive the rhetorical structure of unrestricted texts. This algorithm relies on cue phrases and word co-occurrence to derive a valid rhetorical structure and present it as a rhetorical structure tree.

Because Rhetorical Structure Theory is limited to single documents, Radev [5] introduced a theory of cross-document structure (CST), which is used to describe the rhetorical structure of sets of related documents and present a reasonable equivalent to RST for multi-document summaries. CST makes use of a multi-document graph to represent text simultaneously at different levels of granularity (words, phrase, sentences, paragraphs, and documents). Each graph consists of smaller subgraphs for each individual document. It has two types of links: one type represents inheritance relationships among elements within a single document such as words –> sentences –> paragraphs –> documents; a second type of link represents cross-document semantic relationships among text units, such as *equivalence* (two text spans have the same information content), *cross-reference* (the same entity is mentioned), *contradiction* (conflicting information) and *historical background* (information that puts current information in context). Finally, different summaries are generated by preserving some links while removing others, according to user-specified constraints.

However, Radev's approach does not take into consideration discourse structures peculiar to a particular domain or document type. We seek to identify the discourse structure peculiar to dissertation abstracts in sociology and psychology, both at macro-level (between sentences and sections) and micro-level (within sentences). Our focus is more on meaning and semantics expressed in the text, whereas Radev's approach focuses more on structure.

Our project seeks to develop a method for the automatic construction of multi-document summaries of sets of dissertation abstracts that may be retrieved by a digital library system or search engine in response to a user's query. Dissertation abstracts were selected for this study because:

- There is substantial interest in constructing digital libraries of dissertations in universities because there is a ready supply of dissertations, there is no copyright problem, and they contain a wealth of high-quality information
- Dissertation abstracts have a relatively clear and standard discourse structure
- The language is more formal and standardized than in other corpora, e.g. news articles

We will focus on sociology and psychology dissertations in the initial study because of our greater familiarity with dissertations in these subjects.

To develop an effective multi-document summarization system, techniques have to be developed to accomplish the following tasks:

- Identify similar information found in multiple documents
- Identify unique information in each document
- Identify relationships between documents and pieces of information in different documents (information synthesis)
- Identify and extract the appropriate unit or level of information (word-level, clause-level, sentence-level, etc.)
- Organize the extracted information and identify their relative importance
- Identify and construct the domain knowledge and inference rules needed for accurate information extraction and synthesis
- Present the summary in a way that is easily understandable and useful to the user.

The framework for multi-document summarization taken in this study is to make use of the discourse structure of dissertation abstracts as well as their cross-document structure. Dissertation abstracts are typically informative abstracts (that include a summary of the research results) with a well-known discourse structure taught to all research students. They usually have the following sections which represent the macro-level structure: background/context, problem statement, research method, research results, and concluding remarks. Our study will, however, investigate both the macro-level discourse structure as well as micro-level structures within sentences.

This paper proposes a hierarchical framework for multi-document summarization of dissertation abstracts with five levels:

- Dependent variable level: All of the dependent variables identified in a group of related dissertation abstracts;
- Independent variable level: For each dependent variable, the independent variables investigated in the dissertation projects;
- Relationship level: The relationship between a specific pair of independent and dependent variables;
- Other dependent variable level: For each independent variable, the other dependent variables which are also associated with it besides the current one;
- Document level: The structured and summarized document that describes a specific relationship between a pair of variables.

We also report the results of an analysis of the micro-level structure of the problem statement section of dissertation abstracts based on a sample of 50 abstracts. A list of indicative phrases that denote different aspects of the problem statements is provided. These indicative phrases can be used later for automatic extraction of different pieces of information from the abstracts.

Visualization techniques for presenting the synthesized information will also be investigated in the study. The multi-document summary will be presented in a diagrammatic or visual form, rather than through a natural language abstract. We have implemented a tree-view interface for presenting a multi-document overview of the abstracts in a hierarchical structure. This is an interactive interface that allows users to drill-down for information according to their interests, providing a flexible way of generating multi-document summaries.

2 Discourse Analysis

We downloaded 50 Ph.D and masters' dissertation abstracts on sociology and psychology from the Dissertation Abstracts International and analyzed their structure at the macro-level and micro-level. These dissertations covered five topics: school crime, juvenile crime, domestic violence on children, women's studies and recreation.

2.1 Macro-Level Discourse

We explored the macro-level structure of 50 dissertation abstracts and found that they typically have the following sections: background/context, problem statement, research method, research results and concluding remarks. These five categories of information reflect different aspects of a research study. All the information in the abstracts analyzed can be subsumed under these five categories, although not every abstract contains all of the categories.

1. *Background/ Context.* This section may include previous work and rationale for carrying out the study, or the broader context in which the study was carried out. It may come before or after the problem statement or may be absent.

   - *Background:* Introduces the general area of the research problem and gives the background on why it is an important or interesting problem;
   - *Previous studies:* Present previous studies that are directly related to the present study or form the basis of the present study.
2. *Problem Statement.* This section includes research objectives, research questions, hypotheses, and theoretical framework adopted to investigate the problem. The expected results are sometimes indicated as well. In some cases, this section also provides definitions or explanations of concepts.
3. *Research Method.* This section states how the study was carried out. It can be decomposed into three subparts: *design*, *sampling* and *data analysis.* The Design section is used to clarify what type of study this is: experimental, survey, interview, field research, or data analysis. The Sampling section describes who participated in the study, how many cases were in the sample, how they were selected, and whom or what they represented. In addition, the authors may also indicate what types of statistical analyses were performed.
4. *Research Results.* The problems, questions or hypotheses that framed the research are answered in this section. Results of the data analysis are found in the *statistical results* subsection while conclusions are found in a subsection which we refer to as *research findings*.
5. *Concluding Remarks.* This section includes *recommendations*, *future work*, or *implications* of the research.

2.2 Micro-Level Structure of the Problem Statement

In addition to the observable macro-level discourse structure, we analyzed the micro-level structure of the problem statement in the 50 abstracts.

Problem statements are usually focused on variables. A study may investigate only one variable or the relationship between two or more variables. Social research can be divided into three types, depending on what kinds of relationships a study aims to explore [8]:

- Descriptive research: variables are measured
- Relational research: two or more variables are measured at the same time to see if there is any relationship between them
- Causal research: variables are manipulated by the researcher to see how they affect other variables

A variable may be an object or event (i.e. noun), process or action (i.e. verb) or an attribute (i.e. adjective). In causal research, one or more variables are designated as the dependent variable (DV) while another group of variables are designated the independent variables (IVs). DVs are the variables the researchers are interested in explaining or predicting, while IVs are variables that affect or are used to predict the DVs. In relational research however, variables are not distinguished as such.

To understand the micro-level structure of problem statements in dissertation abstracts, we analyzed the abstracts to identify the main types of semantic relations found in them. These are represented in conceptual graph notation [6] as follows:

- Relational research: [variable 1] -> (relation with. *) -> [variable 2]
- Causal research: [IV] -> (effect on. *) -> [DV]

In conceptual graph notation, concepts are represented in square brackets and relations are represented in round brackets with arrows indicating the direction of the relation. The asterisk indicates that the subtype of *relation* or *effect* is unknown and is to be determined in the research study.

Descriptive studies can be represented as follows:

- [IV] -> (attribute. *) -> [*]

This indicates that the study seeks to identify attributes of the variables investigated in the study.

While many studies aim to explore relationships directly, some explore relationships on the basis of perception, a framework, or a model. For example, "*The purpose of this qualitative, descriptive study was to examine mother's perception of how their children are affected by exposure to domestic violence.*" We call this the contextual relation. The types of contextual relations we identified are given in Table 1.

Table 1. Types of contextual relations found in problem statements

No.	Types of Contextual Relation	Example
1	perception	The purpose of this research is to assess the perception of public school districts' human resources directors regarding the effects of …
2	model	This research posits a theoretical model of school crime based on family, home, economics and demography, in association with …
3	hypothesis	My hypothesis was "as structural strain increase in the United States, the cultural moves from sensate to ideational."
4	attitude	The study also wished to discover teachers' attitudes about the impact that domestic violence has on children.
5	theory	The purpose of this study was to test Barkley's theory and …
6	framework	…, my dissertation develops a new conceptual framework for investigating social information processing…

The variables themselves, including IVs and DVs, may have other relations that qualify them or modify their meaning. A variable concept can be an object (such as *school, family*), abstract object (such as *training program, distance*), event (such as *violence*, *crime*), agent (subject who executes an action), patient (object of an action), process (such as *change*, *increase*, *growth*, *variation*), or phenomenon (such as *young Israelis journeyed to South America or Asia during the past two decades*). All of these may be related to other concepts which may specify an attribute of the variable concepts or qualify the variable concept.

An attribute is a quality that is part of the nature of the entity denoted by the concept. It may be unknown and needs to be investigated in the study. For example, in "*school size*", "*size*" is the attribute of an object "*school*". A qualifier, on the other hand, restricts or narrows the meaning of the concept. For example, in "*serious school crime*", the variable is an event *"crime"* whose scope is narrowed down by a location qualifier *"school"* and a degree qualifier *"serious*". Table 2 gives examples of different types of relations linked to different types of variable concepts. Tables 3 and 4 list the concept attributes and qualifiers found in the sample abstracts.

Table 2. Types of relations found in variables

No.	Variable Type	Example	Interpretation
1	[object]->(qualifier)->[*] [object]->(attribute.*)->[*]	public school size	object: school qualifier of object: public attribute of object: size
2	[object]->(qualifier)->[*] [attribute]->(qualifier)->[*] [object]->(attribute.*)->[*]->(attribute.*)->[*]	parameters of legal authority of officers employed by these programs	object: officers qualifier of object: employed by these programs attribute of object: authority qualifier of attribute: legal attribute of attribute: parameters
3	[event]->(qualifier)->[*]	serious school violence	event: violence location qualifier of event: school degree qualifier of event: serious
4	[event]->(qualifier)->[*] [event]->(attribute.*)->[*]	juvenile crime rate	event: crime qualifier of event: juvenile attribute of event: rate
5	[action]->(qualifier)->[*] [agent]->(action)->[*]	student extracurricular activity participation	agent: student agent's action: participation qualifier of action: extracurricular activity
6	[patient]->(action)->[*]	teacher recruitment	patient of action : teacher action to patient: recruitment
7	[process]->(qualifier)->[*]	changes in myosin heavy chain protein	process: changes qualifier of process : in myosin heavy chain protein

3 Indicator Phrases for Information Extraction

To parse the discourse structure of documents and extract them for multi-document structure summarization, indicator phrases are used. These are phrases that are associated with and that often signal the type of information in the text**.** From indicator phrases, patterns can be constructed to extract the information from the text. In this study, we focus on problem statements and analyze their characteristics to derive patterns from them.

In the 50 sample abstracts, almost all the problem statements contained an indicator phrase in the beginning of the sentence such as "*The purpose of this study was to investigate …*", "*This research was designed to examine…*" Although these phrases do not carry any significant content, they are very useful for identifying whether the sentence is a problem statement. Other indicator phrases are given in Table 5.

Table 6 lists some synonyms for the keywords given in brackets in Table 5. Table 7 lists indicator phrases that indicate the type of relation between variables.

Table 3. Attributes found in sample abstracts

No.	Attribute	Example
1	nature	the nature of sibling relationships in families with a history of violence
2	characteristic	characteristics of single African-American welfare recipient mothers
3	size	school district size
4	density	school district density
5	rate	rate of school crime
6	pattern	the patterns of fathers' touch with low birth weight infants
7	quality	the quality of fathers' touch with low birth weight infants
8	parameters	parameters of the legal authority of the officers
9	type	types of violence
10	prevalence	prevalence of violence
11	function	if father visitation is more helpful or harmful in domestic violence families
12	frequency	frequencies of criminal incidences
13	facilitator	facilitators to collaboration
14	barrier	barriers to collaboration
15	length	length of (children's) exposure to violence
16	category	categories of crimes incidences
17	ability	children's ability to manage emotional expressions
18	possibility	the possibility of measuring internalizing behaviors in preschool children
19	likelihood	the likelihood of engaging in delinquent behavior

Table 4. Qualifiers found in sample abstracts

No.	Qualifier	Example
1	degree	serious school violence, juvenile violent crime rate
2	location	school violence, violence in three high schools in the Camden City Public School District
3	scope	officers employed by these programs, risk factors of children under the age of twelve who had engaged in delinquent acts.
4	time	performance parameters during 4-h of prolonged steady state cycling followed by a maximal performance time trial
5	purpose	a curriculum for leisure transition planning
6	aspect	football fans' attitudes toward televised football

4 Variable-Based Framework with Hierarchical Structure

We developed a variable-based framework with a hierarchical structure to analyze the cross-document structure of multiple dissertation abstracts on a specific topic. The hierarchy has five levels: IV level, DV level, relationship level, other DV level and document level.

Table 5. Indicator phrases for identifying problem statements

No.	Indicator Phrase	Example
1	The -> [purpose] -> of -> this -> [study] -> be to -> [investigate] -> […]	The purpose of this study was to investigate the relationship between school size and the variables of school dropout rate, rate of school crime and violence and student extracurricular activity participation.
2	My -> [purpose] -> be to -> [answer] -> […]	My purpose here is to answer the following question: What are the effects of restructuring on school delinquency?
3	This -> [study] -> [examine] -> […]	This study examined school administrator differences in reporting school crime based upon selected school and administrator variables.
4	This -> [study] -> [aim to] -> [fill] -> […]	The present study aimed to fill this gap by reviewing 108 files of at risk children.
5	This -> [study] -> be -> [designed to] -> [understand] -> […]	This study was designed to understand the subjective experience of women who had been abused as teens, and generate ideas about what might make it difficult to leave an abuser.
6	[…], this -> [study] -> [discuss] -> […]	Using primary and secondary sources, the thesis discusses the impact of China's political, legal and social changes on juvenile crime.

Table 6. Synonyms for indicator phrases

No.	Indicator Phrase	Synonyms
1	purpose	aim, goal, objective, intent
2	study	research, investigation, research question, research paper, project, thesis, report, dissertation
3	investigate, examine, answer, understand, discuss, fill	explore, address, deal with, determine, identify, increase, test, answer, compare, develop, ……
4	aim to	seek to , propose to, wish to, go on to ……
5	design to	conduct to

Our study analyzed the cross-document structure of five topics: school crime, juvenile crime, domestic violence on children, women's studies and recreation, each of which contains 10 documents. The following example is from the topic of "school crime". We present this example with our prototype system (see Figure 1), which is implemented in Microsoft's ASP scripting language associated with a TreeView client-side ActiveX component. The TreeView component encapsulates specific functionality or behavior on a page and can be used in the same way as regular HTML elements. The database for the prototype was manually constructed. In the future, an information extraction program will be developed to identify and extract IVs, DVs and other pieces of information defined in the framework.

The prototype displays 14 DVs regarding school crime and presents them in an expandable tree structure. Each variable is indicated with a short phrase such as "school crime" and "serious school violence".

Table 7. Relations between variables and selected indicator phrases

Relation Type	Indicator Phrase	Example
Descriptive research [V]-> (attribute.*) -> [*]		The purpose of this study was to determine the types and prevalence of violence ……
Relational research [variable 1]-> (relation with.*)-> [variable 2]	relation with	The purpose of this study was to investigate the relationship between school size and the variables of school dropout rate, rate of school crime and violence and student extra-curricular activity participation.
	association with	This research focused on the identification of variables associated with variance in school crime rates …
Causal research [IV] -> (effect on. *) -> [DV]	effect on	The purpose of this research is to assess the perception of public school districts' human resources directors regarding the effects of serious school violence on teacher retention and recruitment.
	difference in DV based on IV	This study examined school administrator differences in reporting school crime based upon selected school and administrator variables.

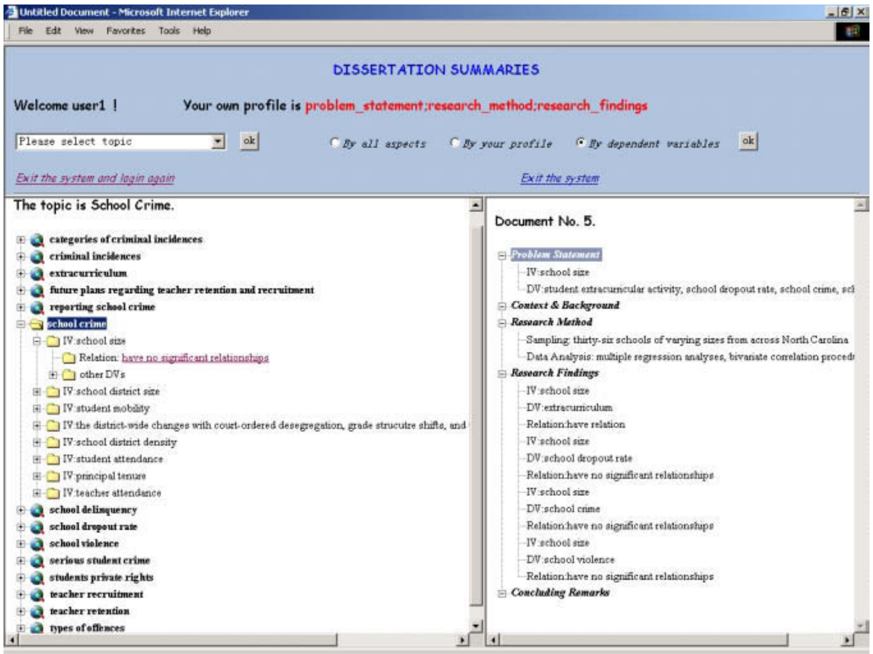
** indicates that the relation is unknown.*

Users can select any DV to explore in greater detail. After the IV level is expanded, all the IVs which influence the selected DV are listed. This presents users with an overview of which factors affect or are associated with a variable of interest.

When an IV is selected, the relationship between it and the selected DV is shown. The relationship, which is expressed with a concise word or phrase, such as "*related with*", "*have no relationship*", "*cause*", is linked with the document which describes it. In addition, the other DVs which the IV can influence are also displayed. A DV is often influenced by many IVs, and an IV often can influence more than one DV. The display thus gives users an overview of how the variables are related to one another in a group of documents.

Documents selected for viewing are shown in the right frame of the window with an expandable tree structure containing two levels: the first contains the five parts/sections of the abstract discussed in Section 2.1, while the second contains the content of each part/section summarized by extracting salient information.

In summary, this framework provides an entire map on a specific topic by identifying the variables from a group of related dissertation abstracts and grouping the information according to dependent variables from different sources. Furthermore, this approach not only outlines the overall structure of the document set, but also allows users to explore further according to their information needs. For example, users can explore the contents of specific abstracts that describe relationships between a pair of dependent and independent variables of interest. Each abstract is displayed with its five sections summarized, so users can locate information easily. This approach thus provides a way to summarize multiple documents that is different from conventional summarization methods, which do not consider important semantic relations expressed within a document or inferred across documents.


Untitled Document - Microsoft Internet Explorer
File Edit View Favorites Tools Help
ok
ok

DISSERTATION SUMMARIES

Welcome user1 ! Your own profile is problem_statement;research_method;research_findings

Please select topic ○ By all aspects ○ By your profile ● By dependent variables

Exit the system and login again Exit the system

The topic is School Crime.

- categories of criminal incidences
- criminal incidences
- extracurriculum
- future plans regarding teacher retention and recruitment
- reporting school crime
- school crime
  - IV:school size
    - Relation: have no significant relationships
    - other DVs
  - IV:school district size
  - IV:student mobility
  - IV:the district-wide changes with court-ordered desegregation, grade strucutre shifts, and
  - IV:school district density
  - IV:student attendance
  - IV:principal tenure
  - IV:teacher attendance
- school delinquency
- school dropout rate
- school violence
- serious student crime
- students private rights
- teacher recruitment
- teacher retention
- types of offences

Document No. 5.

- *Problem Statement*
  - IV:school size
  - DV:student extracurricular activity, school dropout rate, school crime, sc
- *Context & Background*
- *Research Method*
  - Sampling: thirty-six schools of varying sizes from across North Carolina
  - Data Analysis: multiple regression analyses, bivariate correlation proced
- *Research Findings*
  - IV:school size
  - DV:extracurriculum
  - Relation:have relation
  - IV:school size
  - DV:school dropout rate
  - Relation:have no significant relationships
  - IV:school size
  - DV:school crime
  - Relation:have no significant relationships
  - IV:school size
  - DV:school violence
  - Relation:have no significant relationships
- *Concluding Remarks*

Fig. 1. Prototype system of the multi-document summary on the topic of "school crime"

5 Conclusion and Future Work

This paper proposes a framework for summarizing multiple dissertation abstracts through analyzing the discourse structure across the documents on the basis of their variables. In this framework, we identify the dependent and independent variables underlying a given document collection on a particular topic. Starting with dependent variables, the framework presents the relationships among variables from different sources using a hierarchical structure. This is achieved through our Web-based tree-view interface. The framework has two advantages: it gives users a map or an overview of a particular topic and it allows them to explore according to their interests.

As a next step, we intend to analyze more samples of sociology and psychology abstracts to further develop and validate the macro-level and micro-level discourse structure of dissertation abstracts. We plan to extend the analysis of micro-level structure to include the research findings section. We also plan to construct indicator phrase patterns for automatic parsing and extraction of information.

User studies will also be conducted to evaluate the tree-view interface, to determine whether this technique is an effective way to present multi-document summaries of dissertation abstracts. Other visualization methods will also be explored to identify how best to present abstract summaries.

References

1. Bell, A., & Garrett, P. (1998). Media and discourse: A critical overview. *Approaches to Media Discourse*, 65-103. Oxford: Malden.
2. Jansen, B., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the web. *SIGIR Forum, 32*(1), 5-17.
3. Mann, W., & Thompson, S. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text, 8*(3), 243-281.
4. Marcu, D. (1997). The rhetorical parsing, summarization, and generation of natural language texts. Ph.D. Dissertation, Department of Computer Science, University of Toronto.
5. Radev, D. (2000). A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue*. Available at: http://www.sigdial.org/sigdialworkshop/proceedings/ radev.pdf
6. Sowa, J. (1984). *Conceptual structures: Information processing in mind and machine.* Reading, MA: Addison-Wesley.
7. Spink, A., & Xu, J.L. (2000). Selected results from a large study of Web searching: the Excite study. *Information Research* [Online], 6(1). Available at: http://informationr.net/ir/6-1/paper90.html
8. Trochim, W. (1999). *The research methods knowledge base*. Cincinnati, OH: Atomic Dog Publishing.
9. Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to web search results. In *Proceedings of the Eighth International World Wide Web Conference.* Available at: http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html

Generality of Texts

R.B. Allen and Yejun Wu

College of Information Studies
University of Maryland
College Park, MD 20742, U.S.A.
{rba,wuyj}@glue.umd.edu

Abstract. When searching or browsing, a user may be looking for either very general information or very specific information. We explored predictors for characterizing the generality of six encyclopedia texts. We had human subjects rank-order the generality of the texts. We also developed statistics from analysis of word frequency and from comparison to a set of reference terms. We found a statistically significant relationship between the human ratings of text generality and our automatic measure.

1 Introduction

In collection selection, it would be helpful to have some measures of the characteristics of a collection. As part of our ongoing research to define the properties of documents and collections, we had demonstrated a measure of the "scope" of collection [1], where scope is considered to be the "breadth" of a collection.

We believe that it would also be of interest to determine the scope of individual documents. However, we were not able to get consistent inter-subject ratings of the "scope" of documents in pilot studies. One difficulty concerned the subjects' understanding of what constituted the "scope". The subjects seemed particularly confused about how to compensate for document structure.

In this work, we explore a related property of document "generality". By "generality", we mean that a document "addresses general things or concepts". We believe that document generality is easier for individuals to understand. The documents we examined did not have distinct sections. We believe the searchers will find it useful to know whether a retrieved document is general or concrete. For instance, a general chemistry textbook would cover many areas of chemistry and would be distinguished from texts that focused on sub-specialities in chemistry such as organic chemistry and biochemistry.

2 Generality of Texts

2.1 Rating the Generality of Documents

We selected six one-page articles from www.encyclopedia.com which we judged covered a range from "general" to "concrete". Eight subjects were asked to rate

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 111–116, 2002.

these documents using the instructions in Figure 1. To encourage consistency, the subjects first made pair-wise comparisons between the documents and they used these judgements to rank-order the documents' generality. The subjects were paid $12 for the approximately one-hour task.

> You are asked to judge (have a sense of) the generality/concreteness of each document. If a document addresses mostly general things/concepts, we consider it has high generality; if a document addresses mostly concrete or specific things, it has low generality. Read through each article in 4-5 minutes and have a sense of its generality. Please feel free to take notes on the documents. Please plan to read through all the 6 documents in about 30 minutes. Take 2 documents at a time and compare their generality, assign either "more than" (>), or "equal to" (=), or "less than" (<) between the two documents on the attached work sheet. There are totally 15 comparisons. Please plan to finish all the comparisons in about 20-25 minutes. Finally, rank/sort the 6 documents in terms of generality (please indicate ties if any).

Fig. 1. Instructions to subjects.

Beyond the instructions, the subjects developed their own criteria. During the debriefing at the end of the experiment, they reported basing their ratings of concreteness on factors such as: "more numbers, places, enumeration of the aspects of an entity and detailed things". Figure 2 shows a selection from a general document while Figure 3 shows a selection from a concrete document.

> In many countries extensive government programs control the planning, financing, and regulation of agriculture. Agriculture is still the occupation of almost 50% of the world's population, but the numbers vary from less than 3% in industrialized countries to over 60% in Third World countries.

Fig. 2. Selection for the text on "modern agriculture" from encyclopedia.com which was rated as most "general".

The average of the inter-subject correlation was 0.12. However, one subject was very different from the others and without that subject the inter-observer correlation would have been much higher.

3 General/Concrete Terms

Following our earlier work on scope [1], in which we were able to predict the distance between documents (and hence the scope of the collection) based on

> In the 1960s and 1970s, a combination of 2,4-D and 2,4,5-T was widely used in Vietnam as a defoliant under the name Agent Orange . As a result of questions concerning the possible health effects of the use of Agent Orange, heightened awareness of possible ecological and health dangers attributable to herbicides has resulted in reevaluation of many compounds and has called indiscriminate use into question.

Fig. 3. Selection for the text on "herbicide" which was rated as most "concrete".

the distance between individual words, we attempted to predict the generality of documents from the generality of the terms in those documents. Therefore, we developed a measure for the generality of documents based on the terms in the documents.

4 Using Cooccurrence to Tell Whether a Target Word Is General or Concrete

We believe that general terms should appear across a wide variety of contexts. A set of 64 terms was chosen to serve as a reference collection. An attempt was made to choose a wide range of terms from general to specific. Our technique allows all parts of speech to be measured as concrete or specific. For instance, we were able to judge adjectives as specific or general.

4.1 Measuring Co-occurrence

Search engines such as Google and Altavista provide counts of "pages found" for their searches. As developed in Allen and Wu [1], these measures can be used to provide a measure of the relatedness of two terms.

Figure 4 shows the (asymmetric) relatedness of several pairs of words obtained with this method. See [1] for more discussion on the interpretation of these values.

Word1	**Word2**	**Relatedness**	
		W1-W2	**W2-W1**
truck	automobile	0.0740	0.0776
gun	shoot	0.1097	0.1917
flower	beautiful	0.1242	0.0481
mouse	keyboard	0.2542	0.2960

Fig. 4. Sample co-occurrence relatedness values for word pairs (from [1]).

We define A as the number of hits for word1, B for word2, and C for word1 AND word2. The relatedness between word1 and word2 can be easily calculated as $\frac{C}{A}$ or $\frac{C}{B}$ or other variant expressions. A joint entropy measure (Equation 1) was most successful.

$$Rel\ Joint\ Entropy = -(\frac{C}{A}Log\frac{C}{A}) - (\frac{C}{B}Log\frac{C}{B}) \tag{1}$$

We wrote a C-language program that used the UNIX utility "expect" for managing the interactive telnet with the Google research server, extracted the frequency counts from the return file, and stored the results using the database utility "gdb".

4.2 Predicting the Generality of Words

Figure 5 shows examples of words from our reference list of 64 terms that demonstrated high-generality and low-generality.

High-Generality Terms	**Low-Generality (Concrete) Terms**
allow	cognition
take	hawk
approach	hat
nature	atom
change	fluoride
process	minnow

Fig. 5. Examples of high-generality and low-generality terms.

The joint entropy measure was used to confirm that general terms were more related to each other than were concrete terms related to each other. We computed the relatedness of the 32 general terms with themselves, the 32 concrete terms with themselves, and the general terms with the concrete terms (see Figure 6). The difference between 0.168 ($\sigma = 0.017$) and 0.137 ($\sigma = 0.025$) is significant, $t = 10.48, p < 0.01, df = 31$ (two-tailed test). The difference between 0.137 and 0.110 ($\sigma = 0.028$) is also significant, $t = 4.98, p < 0.01, df = 31$.

	General	**Concrete**
General	0.168 (hi)	0.137 (mid)
Concrete	0.137 (mid)	0.110 (lo)

Fig. 6. Average relatedness of the two sub-lists. Predictions are shown in parentheses.

4.3 Validation with WordNet

As another validation of our approach, we hypothesized that general words would be closer to the root of the WordNet hierarchies [3]. We counted the levels between a word and its root in the WordNet hierarchy. If a word has more than one meaning in WordNet, we take the most common one. The average level of the top 32 general words versus the 32 concrete words is 3.30 ($\sigma = 1.31$) and 6.96 ($\sigma = 2.51$) respectively ($t = 8.87, two - tailed, p < 0.01, df = 31$).

4.4 Relationship between Term Generality and Document Generality

We took all the terms for each document and filtered them through a stop list of 300 common words. We then took words that were repeated at least twice in the remaining word lists. The generality of each word was computed as its relatedness to the set of 64 reference words using the joint entropy measure. We took the mean of word generality to obtain the generality of the entire document. Finally, we ranked the six documents according to their generality and compared them with the averaged human ratings of the documents' generality. A linear trend test was performed with ANOVA [4]. A specific comparison for the linear component was a significant effect, $F(1, 42) = 6.24, p < 0.05$. This confirmed that the generality of the terms was an effective predictions of the generality of the texts.

4.5 Word Frequency

Word frequency has been proposed as a measure of document readability, although more common measures of readability (e.g., [6]) are based on the number of syllables and the length of words. General words tend to be of higher frequency. The correlation of generality and word frequency was $r = 0.62, df = 5, n.s.$.

We believe that our generality measure has more face validity than word frequency for predicting the generality of concepts in the articles. For instance, note that there are several low-frequency terms with high generality such as "approach" and "handle". There are also high-frequency terms with low-generality terms such as "hat" and "game". We found that the generality measure predicted mean ratings better than the word frequencies using a stepwise multiple linear regression, although the difference of the measures was not significant.

4.6 Hierarchy and Categories

Because of the relationship between these measures and the WordNet hierarchy reported above, we explored whether these measures fully predicted hierarchical relationships. We found that in the hierarchies such as ANIMAL >> MAMMAL >> DOG that mammal was less frequent than required to support the hierarchical order. Rather, the results seem to support the view that DOG is a basic-level category [5], while MAMMAL is not.

5 Conclusion

5.1 Applications and Implications

In the traditional approach to document searching, a search return list simply presents the documents ranked according to how well they match a retrieval algorithm. We go further and propose that the notion of relevance ranking should be replaced by indicators of document generality. Indeed, we would go even further and suggest that information retrieval systems should increasingly attempt to educate their users to understand the concepts used in the documents in the return list. This may mean that general documents returned from a search should be presented before concrete documents. More ambitiously, the search return lists could adopt the principles of adaptive hypermedia [2].

5.2 Other Properties of Documents

We have now explored the notions of the "scope" of a collection and the "generality" of a document. There are several other related concepts that we would like to operationalize. These include the "depth" and "coverage" of a document. However, our initial attempts to do this have encountered difficulties. For instance, we would like to measure the "number" of concepts in a document. However, thus far we have had difficulty in defining an unambiguous measure for identifying concepts. Nonetheless, we believe that eventually a full range of metrics will be worked out and that these will help both human beings and machines interact with those documents.

Acknowledgements. We thank Google for access to the research port on their service during the Spring of 2002. We also thank encyclopedia.com for the use of their articles.

References

1. Allen, R. B., and Wu, Y. Measuring the scope of collections. *Journal of the American Society for Information Science* (submitted).
2. Brusilovsky, P. Adaptive hypermedia. *User Modeling and User Adapted Interaction 11* (2001), 87–110.
3. Fellbaum, C., Ed. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.
4. Keppel, G. *Design and analysis: A researcher's handbook*, 3^{rd} ed. Prentice Hall, Englewood Cliffs, NJ, 1991.
5. Rosch, E. Principles of categorization. In *Cognition and Categorization*, E. Rosch and B. B. Lloyd, Eds. L. Erlbaum Associates, 1978, pp. 27–78. Hillsdale, NJ.
6. Sherman, L. A. *Analytics of Literature: A Manual for the Objective Study of English Prose and Poetry.* Ginn and Co., Boston, 1983.

The Effectiveness of a Graph-Based Algorithm for Stemming

Michela Bacchin, Nicola Ferro, and Massimo Melucci

Department of Information Engineering
University of Padua,
Via Gradenigo, 6/a – 35031 Padova – Italy
{michela.bacchin, nicola.ferro, massimo.melucci}@unipd.it

Abstract. In Information Retrieval (IR), stemming enables a matching of query and document terms which are related to a same meaning but which can appear in different morphological variants. In this paper we will propose and evaluate a statistical graph-based algorithm for stemming. Considering that a word is formed by a stem (prefix) and a derivation (suffix), the key idea is that strongly interlinked prefixes and suffixes form a community of sub-strings. Discovering these communities means searching for the best word splits which give the best word stems. We conducted some experiments on CLEF 2001 test sub-collections for Italian language. The results show that stemming improve the IR effectiveness. They also show that effectiveness level of our algorithm is comparable to that of an algorithm based on a-priori linguistic knowledge. This is an encouraging result, particularly in a multi-lingual context.

1 Introduction

In an Information Retrieval (IR) system that manages text resources, indexing is the process that assigns a set of best content describing index terms to each document or query. Usually, both documents and queries are written in natural language, so the words may often occur with many morphological variants, even if they are referred to as a common concept. The basic idea which exists in stemming is that words which are similar in morphology are likely to be similar in meaning, so they can be considered as equivalent from an IR point of view. The goal of a stemming algorithm is to reduce variant word forms to a common morphological root, called "stem".

The effectiveness of stemming is a debated issue, and there are different results and conclusions. If effectiveness is measured by the traditional precision and recall measures,[1] it seems that for a language with a relatively simple morphology, like English, stemming influences the overall performance little. [8] In contrast, stemming can significantly increase the retrieval effectiveness and can

[1] *Recall* is the fraction of relevant documents that has been retrieved, and *precision* as the fraction of retrieved documents that are relevant.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 117–128, 2002.

also increase precision for short queries or for languages with a more complex morphology, like the romance languages. [10,15] Finally, as the system performance must reflect user's expectations it has to be considered that the use of a stemmer is intuitive to many users, who can express the query to the system using a specific term without keeping in mind that only a variant of this term can appear in a relevant document. [8] Hence, stemming can be viewed also as a sort of feature related to the user-interaction interface of an IR service.

To design a stemming algorithm, it is possible to follow a linguistic approach, using prior knowledge of the morphology of the specific language, or a statistical approach using some methods based on statistical principles to infer from the corpus of documents the word formation rules in the language studied. The former kind of algorithms imply manual labor which has to be done by experts in linguistics – as matter of the fact, it is necessary to formalize the word formation rules, the latter being hard work, especially for those languages whose morphology is complex. Stemming algorithms based on statistical methods ensure no costs for inserting new languages on the system, and this is an advantage that becomes crucial especially for applications to Digital Libraries which are often constructed for a particular institution or nation, and can manage a great amount of non-English documents as well as documents written in more than one different languages.

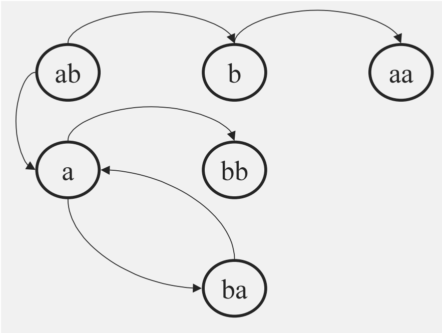
2 Methodological Approach

We will consider a special case of stemming, which belongs to the category known as *affix removal stemming.* [5] In particular our approach stays on a suffix stripping paradigm which is adopted by most stemmers currently in use by IR, like those reported in [11,14,18]. This stemming process splits each word into two parts, prefix and suffix, and considers the stem as the sub-string corresponding to the obtained prefix.

By exploiting a sort of mutual reinforcement between prefix and suffix of a word, we compute the best stem and the best derivation, i.e. the best split, of the word. The rationale of using mutual reinforcement is based on the idea that stems extracted from a finite collection of unique words are those prefixes that are very frequent and form words together with very frequent suffixes. The key idea is that interlinked good prefixes and suffixes form a community of sub-strings whose links correspond to words, i.e. to splits. Discovering these communities is like searching for the best splits. Note that very frequent prefixes are candidate to be stems, but they are discarded if they are not followed by very frequent suffixes; for example, all initials are very frequent prefixes but they are unlikely stems because the corresponding suffixes are rather rare, if not unique – the same holds for suffixes corresponding to ending vowels or consonants. Thus, there are prefixes being less frequent than initials, but followed by frequent suffixes, yet less frequent than ending characters – these suffixes and prefixes correspond to candidate correct word splits and we label them as "good".

2.1 Mutual Reinforcement as a Method for Discovering the Best Stems

Let us consider a finite collection of unique words $W = \{w_1, ..., w_N\}$ and a word $w \in W$ of length $|w|$, then w can be written as $w = xy$ where x is a prefix and y is a suffix, provided $|x| > 0$ and $|y| > 0$. If we split each word w into all the $|w| - 1$ possible pairs of sub-strings, we build a collection of sub-strings, and each sub-string may be either a prefix, a suffix or both of at least an element $w \in W$. Using a graphical notation, the set of prefixes and suffixes can be written as a graph $g = (V, E)$ such that V is the set of sub-strings and $w = (x, y) \in E$ is an edge w that occurs between nodes x, y if $w = xy$ is a word in W. By definition of g, no vertex is isolated. As an example, let us consider the following toy set of words: W={aba, abb, baa}; splitting these into all the possible prefixes and suffixes produces a graph, reported in Figure 2.1a, with vertex set V={a, b, aa, ba, ab, bb} and edge set {(ab,a), (ba,a), (b,aa), (ab,b), (a,ba), (a,bb)}.



sub-string	prefix score	suffix score
a	0.250	0.333
aa	0.000	0.167
ab	0.375	0.000
b	0.125	0.167
ba	0.250	0.167
bb	0.000	0.167

(a) (b)

Fig. 1. (a) The graph obtained from W. (b) The prefix and suffix scores from W

The method used to compute the best split of each word employs the notion of mutual reinforcement and the criteria based on frequencies of sub-strings to decide the goodness of prefixes and suffixes, often used in statistical morphological analysis, [13,6] and in the pioneer work. [7] The contribution of this paper is the use of mutual reinforcement notion applied to prefix frequencies and suffix frequencies, to compute the best word splits which give the best word stems.

If a directed edge exists between x and y, the mutual reinforcement notion can be stated as follows:

> good prefixes point to good suffixes, and good suffixes are pointed to by good prefixes.

In mathematical form, let us define $P(y) = \{x : \exists w, w = xy\}$ and $S(x) = \{y : \exists w, w = xy\}$ that are, respectively, the set of all prefixes of a given suffix y and the set of all suffixes of a given prefix x. If p_x is the prefix score, i.e. the degree to which the prefix x is a stem, and s_y is the suffix score, i.e. the degree to which the suffix y is a derivation, then the mutual reinforcing relationship can be expressed as:

$$s_y = \sum_{x \in P(y)} p_x \qquad p_x = \sum_{y \in S(x)} s_y \tag{1}$$

under the assumption that scores are expressed as sums of scores and splits are equally weighed.

2.2 The Estimation of Prefix Scores

To estimate the prefix score, we used the quite well-known algorithm called HITS (Hyper-link Induced Topic Search) reported in [9] and often discussed in many research papers as a paradigmatic algorithm for Web page retrieval. It considers a mutually reinforcing relationship among good authorities and good hubs, where an authority is a web page pointed to by many hubs and a hub is a web page which points to many authorities. The parallel with our context will be clear when we associate the concept of a hub to a prefix and that of authority to a suffix.

Using the matrix notation, the graph g can be described with a $|V| \times |V|$ matrix $\mathbf{M}$ such that

$$m_{ij} = \begin{cases} 1 & \text{if prefix i and suffix j form a word} \\ 0 & \text{otherwise} \end{cases}$$

As explained in [9], the algorithm computes two matrices after the first iteration: $\mathbf{A} = \mathbf{M}^T\mathbf{M}$ and $\mathbf{B} = \mathbf{M}\mathbf{M}^T$, where the generic element a_{ij} of $\mathbf{A}$ is the number of vertices that are pointed by both i and j, whereas the generic element b_{ij} of $\mathbf{B}$ is the number of vertices that point to both i and j. The k-step iteration of the algorithm corresponds to computing $\mathbf{A}^k$ and $\mathbf{B}^k$. In the same paper, it has been argued that $\mathbf{s} = [s_j]$ and $\mathbf{p} = [p_i]$ converge to the eigenvectors of $\mathbf{A}$ and $\mathbf{B}$, respectively.

Here we map HITS in our study context, as follows:

```
Compute suffix scores and prefix scores from $W$
$V$: the set of sub-strings extracted from all the words in $W$
$N$: the number of all sub-strings in $V$
$n$: the number of iterations
$\mathbf{1}$: the vector $(1, ..., 1) \in \mathcal{R}^{|V|}$
$\mathbf{0}$: the vector $(0, ..., 0) \in \mathcal{R}^{|V|}$
```

```
$\mathbf{s}^{(k)}$: suffix score vector at step $k$
$\mathbf{p}^{(k)}$: prefix score vector at step $k$
$\mathbf{s}^{(0)} = \mathbf{1}$
$\mathbf{p}^{(0)} = \mathbf{1}$
for each $k = 1, ..., n$
        $\mathbf{s}^{(k)} = \mathbf{0}$
        $\mathbf{p}^{(k)} = \mathbf{0}$
        for each $y$
                $s_y^{(k)} = \sum_{x \in P(y)} p_x^{(k-1)}$;
        for each $x$
                $p_x^{(k)} = \sum_{y \in S(x)} s_y^{(k)}$;
        normalize $p^{(k)}$ and $s^{(k)}$ so that $1 = \sum_i p_i^{(k)} = \sum_j s_j^{(k)}$
for each $x$
        $p_x^{(n)} = p_x^{(n)}/|S(x)|$
end.
```

Differently from HITS, each prefix score p_x is divided after the n-th iteration by the number of words with the prefix x, i.e. the number of out-links of the node corresponding to the prefix x. The latter arithmetic operation provides an estimation of the probability that x is a stem of a given word. This probability is a component of a probabilistic framework, see [1] for a more detailed discussion, since the illustration of this framework is out of the scope of this paper. However, we explain why the scores can be modeled within a probabilistic framework. In a recent work, it has been proved that HITS scores can be considered as a stationary distribution of a random walk. [2] In particular, it has been proved the existence of a Markov chain $M^{(k)}$, which has the stationary distribution equal to the hub vector after the k^{th} iteration of the Kleinberg's algorithm, which is, in our context, the prefix score vector $\mathbf{p} = [p_j]$. The generic element $q_{ij}^{(k)}$ of the transition matrix referred to $M^{(k)}$ is the probability that, starting from i, one reaches j after k "bouncing" to one of the suffixes which begins to be associated with i and j. To interpret the result in a linguistic framework, p_i can be seen as the probability that i is judged as a stem by the same community of sub-strings (suffixes) being resulted by the process of splitting words of a language. Considering scores as probabilities permits us to model our graph-based stemming algorithm within a probabilistic framework [1].

In Table 1, all the possible splits for all the words are reported and measured using the estimated probability. For each word we choose as stem the prefix with the highest probability.

3 Experiments

The aim of the experiments is to compare the retrieval effectiveness of the link analysis-based algorithm illustrated in the previous Section with that of an algorithm based on a-priori linguistic knowledge, because the hypothesis is that a

Table 1. The candidate splits from W={aba, baa, abb}.

word	prefix	suffix	words beginning by prefix	probability	choice
baa	b	aa	1	0.1250	
baa	ba	a	1	0.2500	*
aba	a	ba	2	0.1250	
aba	ab	a	2	0.1875	*
abb	a	bb	2	0.1250	
abb	ab	b	2	0.1875	*

language-independent algorithm, such as the one we propose, might effectively replace one developed on the basis of manually coded derivational rules. Before comparing the algorithms, we assessed the impact of both stemming algorithms by comparing their effectiveness with that reached without any stemmer. In fact, we did also want to test if the system performance is not significantly hurt by the application of stemming, as hypothesized in [8]. To evaluate stemming, we decided to compare the performance of an IR system changing only the stemming algorithms for different runs, all other things being equal. We conducted the evaluation procedure following the trusted Cranfield methodology, [4] which requires us to evaluate an IR system on a test collection consisting of a set of documents, a set of queries and a list of relevance judgments – each judgment states whether a judged document is relevant or not for each query.

3.1 Experimental Setting

We carried out the retrieval experiments by using a test collection, an experimental prototype system, a suite of effectiveness measures for reflecting the search quality, and statistical methods for judging whether differences between runs can be considered statistically significant.

Test Collection. We carried out the retrieval experiments on the Italian subcollections of the Cross-Language Evaluation Forum (CLEF) 2001 test collection. CLEF is a series of evaluation campaigns which has been held once a year since 1999. [3,16] It offers an infrastructure for testing, tuning and evaluating IR systems operating on European languages. The test documents consist of two distinct subsets of articles both referring to year 1994:

- *La Stampa*, which is an Italian national newspaper;
- Italian SDA, which is the Italian portion of the news-wire articles of SDA (Swiss Press Agency).

We want to concentrate on a morphologically complex European language, as it is the Italian language, because it poses new challenges to stemming which is what we want to investigate. Main features of the test collection are reported in Table 2. After a simple case normalization, the Italian sub-collection has a vocabulary of 333,828 unique words. The query set consists of 50 topics, each one described by a *Title*, a *Description* and a body called *Narrative*.

Table 2. Main features of the collection used in the experiments.

	La Stampa	SDA	Total
Size in KB	198,112	87,592	285,704
Number of documents	58,051	50,527	108,578

Experimental System. For indexing and retrieval, we used an experimental IR system, called IRON, which has been realized by our research group with the aim of having a robust tool for carrying out IR experiments. IRON is built on top of the Lucene 1.2 RC4 library, which is an open-source library for IR written in Java and publicly available in [12]. The system implements the vector space model, [19] and a (tf · idf)–based weighting scheme. [20] The stop-list which was used consists of 409 Italian frequent words and it is publicly available in [21].

As regards the realization of the statistical stemming algorithm, we built a suite of tools, called Stemming Program for Language Independent Tasks (SPLIT), which implements the graph-based algorithm described in Section 2. Using the vocabulary extracted from the Italian CLEF sub-collection, SPLIT spawns a 2,277,297-node and 1,215,326-edge graph, which is processed to compute prefix and suffix scores – SPLIT took 2.5 hours for 100 iterations on a personal computer equipped with Linux, an 800 MHz Intel CPU and 256MB RAM.

Effectiveness Measures. We used R-precision, which is the precision after R relevant retrieved documents, and Average Precision, computed by the trusted evaluation program `trec_eval` developed as part of the experimental SMART system at Cornell University and freely available from [22].

To test the statistical significance of the difference between the compared runs, we carried out a statistical analysis considering each query as a statistical unit and applying the paired Wilcoxon test, which is a non-parametric statistical test working as follows: given two lists X, Y of measures – one list of each run observed – that test replaces each difference D_i between a pair of measures X_i, Y_i with the rank of its absolute value multiplied by the difference sign. The statistics is then $\sum R_i \;/\; \sqrt{\sum R_i^2}$, where $R_i = sign(D_i) \times rank|D_i|$ and is compared to its expected value under the null hypothesis that lists are equal.

3.2 Runs

We tested four different stemming algorithms:

1. NoStem: No stemming algorithm was applied.
2. Porter-like: We used the stemming algorithm for the Italian language, which is freely available in the Snowball Web Site edited by M. Porter. [17] Besides being publicly available for research purposes, we have chosen this algorithm because it uses a kind of a-priori knowledge of the Italian language.
3. SPLIT: We implemented our first version of the stemming algorithm based on a link-analysis with 100 iterations.
4. SPLIT-L3: We included in our stemming algorithm a little ignition of linguistic knowledge, inserting a heuristic rule which forces the length of the stem to be at least 3.

3.3 A Global Evaluation

We carried out a macro evaluation by averaging the results over all the queries of the test collection. Out the 50 queries of the test collection, 47 queries have relevant documents in the collection, so only these queries were evaluated in the analysis. Table 3 shows a summary of the figures related to the macro analysis of the stemming algorithm.

Table 3. Macro comparison among runs.

	N. Relevant Retrieved	Av. Precision	R-Precision
NoStem	1093	0.3387	0.3437
Porter-like	1169	0.3753	0.3619
SPLIT	1143	0.3519	0.3594
SPLIT-L3	1149	0.3589	0.3668

Note that all the considered stemming algorithms improve recall, since the number of retrieved relevant documents is larger than the number of retrieved relevant documents observed in the case of retrieval without any stemmer; the increase has been observed for all the stemming algorithms. It is interesting to note that precision increases as well, and then the overall performance is higher thanks to the application of stemming than when it is without any stemmer. As previous studies on other non-English languages showed, [15,10] this interesting result suggests that stemming does not cause any trade-off between recall and precision, i.e. both precision and recall can be increased. To confirm the increase of effectiveness, Figure 2 shows the Averaged Recall-Precision curve at different levels of recall.

As regards the use of link-based stemming algorithms, it is worth noting that SPLIT can attain levels of effectiveness being comparable to one based on linguistic knowledge. This is surprising if you know that SPLIT was built without

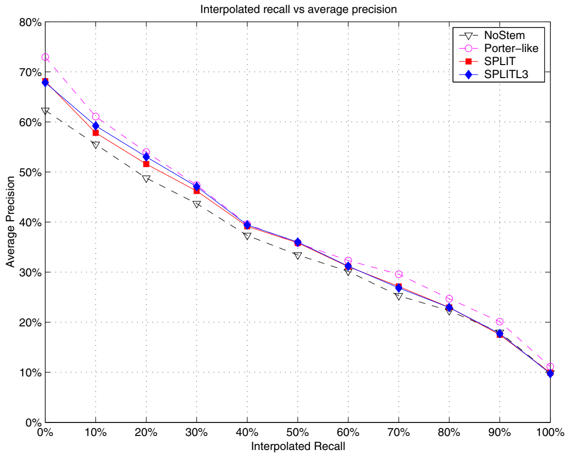


Fig. 2. Average Precision Curve for four stemming algorithms.

any sophisticated extension to HITS and that neither heuristics nor linguistic knowledge was used to improve effectiveness. It should also be considered as a good result, if you consider that it has also been obtained for the Italian language, which is morphologically more complex than English.

After analyzing the results obtained at macro level, we realized that performance varied with query and that `SPLIT` performed better than `Porter-like` for a subset of queries. This variation led us to carry out the analysis reported in the next Section.

3.4 Query-by-Query Evaluation

We conducted a more specific analysis based on the evaluation of the stemming effects on each query of the test collection, by calculating the R-Precision and Average-Precision figures for each query and for each run. We carried out the analysis for `Porter-like` and `SPLIT-L3`; the latter was chosen because it performed a little better than `SPLIT`, yet the difference was not statistically significant. Table 4 reports the number of queries in which a stemming algorithm improved, decreased or kept as equivalent R-precision and Average Precision with respects to the "no-stemming" case.

As Table 4 shows, the number of queries showing improvements in performance after the stemming process is greater than the number of queries for which precision decreased. However, the improvement is not strong enough to be considered statistically significant. Moreover, all the stemming algorithms yield comparable results in terms of R-Precision and Average Precision, as the Wilcoxon test suggests for $\alpha = 0.05$. This means that the number of queries for which `Porter-like` performed better than `SPLIT` is comparable to, i.e. not statistically different from, the number of queries for which `SPLIT` performed better

Table 4. Behavior of the algorithms compared with non-stemming.

	R-Precision		Avg-Precision	
	SPLIT-L3	Porter-like	SPLIT-L3	Porter-like
Improved	19	17	26	26
Decreased	13	15	19	19
Equivalent	15	15	2	2

than Porter-like. In other words, SPLIT and Porter-like are equivalently effective.

The latter way of considering improvements corresponds to assess performance from a more user-oriented than system-oriented point of view. If stemming is applied in an interactive context, such as that of Digital Libraries applications, the ranking used to display the results to the user acquire a great importance – from a practical rather than theoretical point of view at least. In fact, it would more interesting to know if the end user finds the relevant document after 10 or 20 retrieved documents instead of knowing if successful retrieval is reached after 50% retrieved documents. To assess stemming performance from a more user-oriented point of view, we were interested in evaluating how the observed improvement of effectiveness thanks to stemming can change the ranking of retrieved documents. Hence, we compared precision at 10, 20, 30 document cutoff, as suggested in [8]. The paired Wilcoxon test suggests to reject the null hypothesis that stemming has no effect on performance; on the contrary, we can confirm the hypothesis that stemming improves the results. Table 5 reports the number of queries in which a stemming algorithm improved, decreased or kept as equivalent the Precision figure computed at 10, 20 and 30 retrieved relevant documents.

Table 5. Behavior of the algorithms compared with the baseline of non-stemming.

	SPLIT-L3			Porter-like		
	10 docs	20 docs	30 docs	10 docs	20 docs	30 docs
Improved	14	19	18	20	23	19
Decreased	7	12	14	8	9	13
Equivalent	26	16	15	19	19	15

The test gives the same results both for the SPLIT-L3 and the Porter-like algorithm at all document cutoff values. To confirm that a statistical and link-based stemming algorithm can be successfully used instead of a-priori linguistic knowledge, we compared the SPLIT-L3 with the Porter-like algorithm for the document cutoff values selected above. Then, we computed the p-value, which is a measure of the probability that the observed difference between the two algorithms could have occurred by chance. We noted that the performances of such algorithms are so close that the p-value of Wilcoxon test for 20 and 30 document cutoff values are over 90%. This means that it is almost certain that the observed difference between the two algorithms occurred by chance, i.e. that there is not any "structural" reason so that the two algorithms are different. Table 6 reports the effects on the queries of SPLIT-L3 algorithm on Porter-Like baseline. For precision, in 10 relevant documents retrieved, the Porter-like algorithm performs better than SPLIT-L3 algorithm, this means that Porter-like is more effective if very few documents are seen; if more than a few documents are seen, SPLIT performs similarly.

Table 6. Behavior of SPLIT-L3 on Porter-like baseline

	10 docs	20 docs	30 docs
Improved	8	16	13
Decreased	13	19	15
Equivalent	26	12	19

4 Conclusions and Future Work

The objective of this research was to investigate a stemming algorithm based on link analysis procedures. The idea has been that prefixes and suffixes, that are stems and derivations, form communities once extracted from words. We tested this hypothesis by comparing the retrieval effectiveness of SPLIT, a graph-based algorithm derived from HITS, with a linguistic knowledge based algorithm, on a quite morphologically complex language as it is the Italian language.

The results are encouraging because effectiveness level of SPLIT is comparable to that developed by Porter. The results should be considered even better since SPLIT does not incorporate any heuristics nor linguistic knowledge. Moreover, stemming, and then SPLIT, showed to improve effectiveness with respects to not using any stemmer.

We are carrying out further analysis at a micro level to understand the conditions under which SPLIT performs better or worse compared to other algorithms. In parallel, theoretical work will disclose properties that permit us to improve SPLIT. Finally, further experimental work is in progress with other languages.

Acknowledgments. The authors thank Carol Peters for the support and the colleagues of the Information Management Systems group of the Department of Information Engineering for the useful discussions and suggestions.

Massimo Melucci's work has been partially supported by the University of Padua under the 2000 research program for young researchers (*Progetti di ricerca per giovani ricercatori*).

References

1. M. Agosti, M. Bacchin, N. Ferro and M. Melucci. University of Padua at CLEF 2002: Experiments to evaluate a statistical stemming algorithm. In *Cross-Language Information Retrieval and Evaluation: Proceedings of the CLEF 2002 workshop*, Lecture Notes in Computer Science series, Springer Verlag (forthcoming).
2. A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web. In *Proceedings of the World Wide Web Conference*, pages 415–429, Hong Kong, 2001. ACM Press.
3. CLEF Consortium. CLEF: Cross-Language Evaluation Forum. `http://www.clef-campaign.org`, 2002.
4. C. Cleverdon. The Cranfield Tests on Index Language Devices. In K. Sparck Jones and P. Willett (Eds.). *Readings in Information Retrieval*, pages 47-59, Morgan Kaufmann, 1997.
5. W.B. Frakes and R. Baeza-Yates. *Information Retrieval: data structures and algorithms.* Prentice Hall, 1992.
6. J. Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):154–198, 2001.
7. M. Hafer and S. Weiss. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10:371–385, 1994.
8. D. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–15, 1991.
9. J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, September 1999.
10. R. Krovetz. Viewing Morphology as an Inference Process,. In *Proceedings of the ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, 1993.
11. J. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.
12. The Jakarta Project. Lucene. `http://jakarta.apache.org/lucene/docs/index.html`, 2002.
13. C.D. Manning and H. Schütze. *Foundations of statistical natural language processing.* The MIT Press, 1999.
14. C.D. Paice. Another Stemmer. In *ACM SIGIR Forum*, 24, 56–61, 1990.
15. M. Popovic and P. Willett. The effectiveness of stemming for natural-language access to sloven textual data. *Journal of the American Society for Information Science*, 43(5):383–390, 1992.
16. C. Peters and M. Braschler. Cross-Language System Evaluation: the CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, 52(12):1067–1072, 2001.
17. M. Porter. Snowball: A language for stemming algorithms. `http://snowball.sourceforge.net`, 2001.
18. M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
19. G. Salton and M. McGill. *Introduction to modern Information Retrieval.* McGraw-Hill, New York, NY, 1983.
20. G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
21. Institut interfacultaire d'informatique. CLEF and Multilingual information retrieval. University of Neuchatel. `http://www.unine.ch/info/clef/`, 2002.
22. C. Buckley. Trec_eval. `ftp://ftp.cs.cornell.edu/pub/smart/`, 2002.

Searching Digital Music Libraries

David Bainbridge, Michael Dewsnip, and Ian Witten

Department of Computer Science
University of Waikato
Hamilton
New Zealand

Abstract. There has been a recent explosion of interest in digital music libraries. In particular, interactive melody retrieval is a striking example of a search paradigm that differs radically from the standard full-text search. Many different techniques have been proposed for melody matching, but the area lacks standard databases that allow them to be compared on common grounds—and copyright issues have stymied attempts to develop such a corpus. This paper focuses on methods for evaluating different symbolic music matching strategies, and describes a series of experiments that compare and contrast results obtained using three dominant paradigms.

1 Introduction

There has been a recent explosion of interest in digital music libraries—indeed, Apple's iPod has been called the world's first consumer-oriented digital library. In all human societies, music is an expression of popular culture. Different generations identify strongly with different musical styles. People's taste in music reflects their personality. Teenagers, in particular, feel that their musical preferences are strongly bound up with who they are. Many researchers seek to capitalize on this natural interest by building digital music libraries [BD00,DB01].

Digital music libraries are an attractive area of study because they present interesting and challenging technical problems, solutions to which are likely to be highly valued by enthusiastic end-users. This paper addresses the problem of searching a music library for a known melody. In other words, given a fragment of an unknown melody, typically played or sung by a user, return a list of possible matches to a large digital library collection. This operation is not supported by the information structures provided by traditional libraries, except insofar as knowledgeable music librarians are able to provide human assistance. For scholarly work on melody, there is a book that provides a paper index of themes [Par75], but its scope is restricted to the older classical repertoire and it does not provide flexible searching options. And yet the problem of melody retrieval is of great interest to a wide range of potential users—so much so that there are popular radio programs that feature human abilities to "guess that tune".

A practical scheme for searching digital music libraries requires robust implementations of several supporting components. First, it is necessary to assemble

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 129–140, 2002.

a large database of music in searchable form—which implies some kind of symbolic representation. Normally this is accomplished by manually entering a large number of melodies on a keyboard, a highly labor-intensive process. The alternative is to automatically infer notated information from an analog form of the music, such as a recording of a performance, or a paper score. The latter possibility, OMR for "optical music recognition," is a well-advanced technology (e.g. [BB01]) but is not addressed in this paper. Second, the audio query, generated by the user singing, whistling, or humming it—or playing it on a keyboard—must first be transcribed into the same symbolic representation. This is a far easier proposition than inferring a musical score from a recording of a performance, because the input is monophonic—only one voice is present. However, again we do not address the problem in this paper: the transformation is accomplished by standard signal-processing techniques of pitch detection [GR69], followed by quantization of the pitch track in both frequency (placing the notes on a standard musical scale) and time (placing the notes within a standard rhythmical framework). Third, the music searching operation must take place within a context that allows the user to examine search results and request that generalized "documents" be presented. Such documents might include stored musical performances, performances synthesized on the fly from some symbolic representation, facsimile images of the score, scores created on demand from a symbolic representation by musical typesetting techniques, and so on. Suitable general contexts for browsing and document presentation exist (e.g. [BNMW+99]); again, they are not addressed by this paper.

We focus here on the central problem of music matching. We assume that the material to be searched is stored in symbolic form in terms of the notated music. We assume that the audio query has been transcribed into the same symbolic form. We assume that a suitable infrastructure is in place for examining and presenting results.

Three basically different approaches to symbolic music matching have been proposed: dynamic programming [MS90], state matching [WM92], and n-gram-based methods that employ standard information retrieval techniques [WMB99]. All have been used to implement practical melody retrieval systems. Dynamic programming techniques work by calculating, in an efficient manner, the "edit distance" between the query and each melody in the database. The lower the distance, the better the match. Rudimentary edit operations include adding a note, deleting a note and substituting one note for another, along with more subtle changes such as consolidating a series of notes at the same pitch into one note of the composite duration. State based matching also works by adding, deleting and substituting notes to provide an approximate match, but its implementation takes a different form. It uses a matrix of bits that records the state of partial matching so far, and achieves efficiency by encoding the matrix as an array of machine words. Given this data-structure only shifts and bitwise Boolean operators are needed to implement the matching progress. Unlike dynamic programming, state-based matching does not keep track of which edits were made, and its running time is proportional to the number of errors that are allowed. N-gram-based methods work by mapping both queries and melodies to textual words (n-letters long) and then using full-text retrieval to locate documents in the database that contain the “words” included in a given query.

The aim of this paper is to provide a comparative evaluation of these three methods for searching digital music libraries. We explore three orthogonal axes. The first measures “query length uniqueness”, and is the minimum length of query (in notes) needed to unambiguously determine a unique melody. The second is ranked return—how high up the sought-after melody appears in the ordered list of returned matches. The third is the computational efficiency of the searching method.

We begin by introducing the workbench we have developed to perform these experiments, and then describe the experiments themselves and the results obtained. We conclude with a summary of our findings.

2 A Workbench for Symbolic Music Information Retrieval

To support practical research in this area we have developed a workbench for symbolic music information retrieval. Designed to fit into a larger digital library software architecture, Greenstone [WRBB00,WBB01], work is divided into two phases: the assimilation phase and the runtime phase. The former is responsible for collecting files together and creating from them the necessary indexes and/or databases. The latter, guided by user input, supports experimentation and evaluates performance measures. While assimilation is typically performed once for a given collection, the runtime phase is executed many times to gather results from different experiments.

The assimilation phrase is controlled by a configuration file that determines what files are gathered together and how they are processed and indexed. It begins with an “import” process that is capable of reading the plethora of different file formats associated with music data and normalizing them by converting them into a canonical format. For this we use the XML version of Guido [HRG01]: it is general, expressive enough for our needs, and straightforward to parse.

The workbench implements three broad types of algorithm for symbolic music information retrieval: state-based matching, dynamic programming, and text-based information retrieval of n-grams. These require different levels of support at assimilation time, and the configuration settings that govern the assimilation phase dictate what indexes and databases are built.

In the runtime phase, users issue commands to interact with the workbench. They can provide sample inputs and match them against the database using different matching algorithms, examining and comparing the results. Each matching method has optional arguments that modify its behavior. For example, one can seek matches only at the start of melodies, rather than at any position within them. Instead of using exact pitch intervals one can match a pitch contour, which records for each pitch change whether it rises and falls, rather than the amount by which it rises or falls. The workbench implements many other matching options. The outcome of a search is held as a result set from which statistics are extracted, graphs plotted, tables generated, and so on.

Interactive use has its limitations, particularly when setting up and running large experiments. Consequently there is a facility for users to develop a “script” that defines a particular series of experiments. This script is then run by the workbench in batch mode, and the results are recorded in files for later examination by the user.

A third mode of operation is to allow a different process, rather than an online user or a pre-prepared script, to access the facilities of the workbench. The workbench can be accessed through a web-based user interface, using the CGI mechanism, to perform music-content retrieval and format the data returned in a suitable format. This allows its use directly by digital library software, for example, Greenstone. The advantage is that exactly the same implementation and options are used for live retrievals as have been evaluated in interactive and off-line experiments.

The workbench design is capable of supporting polyphonic matching. However, the experiments reported here focus on monophonic-to-monophonic matching.

The music information retrieval workbench will be released under the GNU public license. It will provide a uniform basis for evaluating melody matching algorithms. More importantly, other research groups will be able to add their retrieval algorithms to it, allowing a comprehensive comparison of their strengths and weaknesses against the prior state of the art without the need to continually re-implement earlier methods. An alternative strategy, which has been adopted in other communities (e.g. text compression [AB97] and machine learning [BKM98]), is to develop a standard corpus of material against which different algorithms are evaluated, and publish the results of these evaluations. Indeed a public domain workbench is complimentary to such an approach, however in the context of music, on-going efforts to form such a tested have been stymied by issues of copyright.

3 Experimentation

The purpose of our experiments is to shed light on how well commonly-used music information retrieval algorithms perform under a wide variety of conditions. This provides the basic information needed to design and configure a digital music library. Such information is necessary to make a sensible choice of any algorithms used to support query by music content in practice; it is also necessary to fine-tune particular parameter settings. Conditions differ from one digital library to the next, depending on factors such as the user community being served, the computer infrastructure that underpins the service, and the type and size of the collection. Our aim is to provide design data for digital music libraries. If, in addition, a library uses our workbench to respond to queries, the implementation is guaranteed to be the same as was used to produce the design data.

3.1 Dataset

For evaluation, we need to use standard corpora of melodies. Recall the legal difficulties, mentioned above, of creating and distributing corpora. In the absence of a globally used corpus we have used a dataset of folksongs that is available to us internally. The dataset combines songs from the Essen and Digital Tradition collections [BNMW^{+}99] to form a dataset of nearly 10,000 songs.

3.2 Summary of Experiments

Our experiments are based on earlier work by McNab [McN96], Downie [Dow99], and Rand *et al.* [RB01]. Where possible, we follow the same experimental procedures, expanding and extending the details appropriately.

In the first experiment we examine how many notes each method requires for the list of matches to contain just one tune—the sought-after one. This is repeated for a range of different circumstances. The second experiment studies the ranking performance for each method under more realistic user conditions, albeit simulated. The experiment quantifies retrieval quality in terms of the position in which the sought-after tune appears in the returned list, and thereby helps establish how useful each method is from a user's perspective. The third experiment considers computation cost, and introduces a hybrid approach. It is known that music-based n-gram systems are computationally very efficient and have high recall, but suffer from low precision [Dow99]. Motivated by this observation, the final experiment evaluates a system that first performs an n-gram query and then applies (in one variation) the dynamic programming approach to refine the resulting set of search results, and (in a second variation) the state-based matching algorithm.

3.3 Uniqueness

It is interesting to establish how many notes a user must sing before a music information retrieval algorithm returns one match—namely, the sought-after tune. We call this measure *uniqueness* since it helps gauge how selective or inclusive a technique is. For instance, when using exact matching one would expect the number of returned tunes to drop off more quickly than with approximate matching, but does this actually happen, and is the difference significant?

Many parameters can be varied in this experiment: with or without duration, interval or contour, match at the start or anywhere within the tune. Within these broad categories further choices must be made. For example, Downie [Dow99] considered four different ways to map musical notes to text characters and varied the size of n-grams from 4 to 6. We have experimented with many variations, but (due to space limitations) we present selective results that indicate the trends we observed, and note which were the most influential parameters and which had little effect.

Figures 1–3 show the uniqueness measure applied to the folksong dataset. We used exact and approximate matching, ignored durations and took them into account, and described pitch in absolute, interval, and contour terms. For

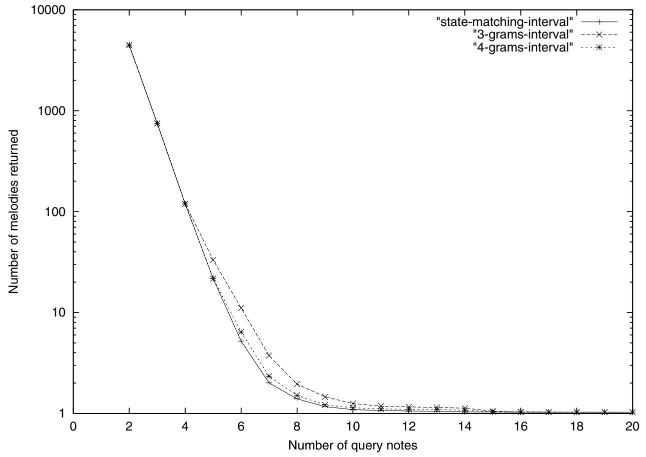


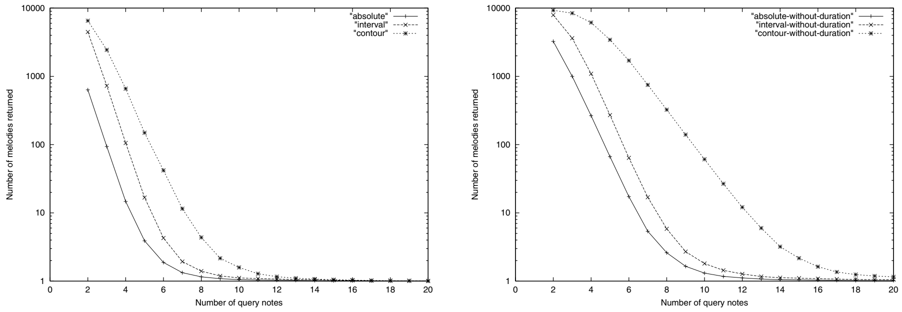
Fig. 1. Testing the uniqueness of matching algorithms versus indexing algorithms (exact interval matching anywhere in the melody, using duration).

300 melodies chosen at random from the database, an initial query of two notes was extended a note at a time up to a limit of twenty notes, and the resulting number of tunes returned was recorded. The average result for each query length was then calculated and plotted.

Figure 1 tests the idealized circumstance that the notes sung in the query—ignoring rests—are exactly the same as the corresponding notes in the sought-after tune in the database. This is as good as it gets! It represents the best results any matching algorithm can hope to achieve using pitch interval (without using rests or resorting to additional features of music notation such as dynamic markings). The figure shows state-based matching and two versions of n-gram (3-gram and 4-gram). Dynamic programming is omitted because it produces *exactly* the same result as state-based matching.

The n-gram methods yield slightly higher curves because in the version tested, detected n-grams do not need to occur consecutively. A stricter version of the search can be accomplished by issuing the query as a “phrase search,” in which case it too would produce the lower of the three curves. While dynamic programming and state-based matching are implicitly designed for approximate matching, this is not true for n-gram-based matching. Removing the requirement that n-grams be consecutive is one way to incorporate flexibility, so it is interesting to see how this variation performs. The second category of experiment (see Section 3.4) evaluates how relaxing this requirement affects the quality of the search results.

Figure 2(a), for state-based matching, compares the result of using absolute pitch (so the query needs to match the exact key the song in is), intervals, and contours. The contour mapping used recorded if a note’s pitch was above, below or the same as the previous note’s pitch (category C3 in Downie’s work [Dow99]).



(a) (b)

Fig. 2. Uniqueness experiment for state-based matching with exact matching anywhere in melody (a) with duration (b) without duration.

Not surprisingly, absolute pitch queries require fewer notes than interval queries, which in turn need fewer than contour queries. For example, a user who sings a query and wants only four songs returned (which can then be checked manually by playing them) must sing 5 notes for the absolute match, 6 notes for the interval match and 8 notes for the contour match.

Figure 2(b) repeats the experiment but ignores note duration. The progression is as before, but contour matching is considerably worse. To return just four songs a further two more notes must be sung for absolute and interval matching, but five for contour matching.

Repeating these two experiments with dynamic programming and consecutive n-gram matching yields exactly the same results. The trend for the more relaxed n-gram version is to return a greater number of melodies than the comparable versions of the other two algorithms, and for the disparity to be at its most significant when the query is 8–14 notes long.

So far we used the dynamic programming and state-based matching algorithms in an "exact matching" mode. This shows how well things go in an ideal situation, and allows comparison with n-gram algorithms that are inherently exact-match based. We now measure uniqueness for approximate matching.

Figure 3 shows what happens when 1, 3 and 5 mismatches are allowed between the state-based matching algorithm and the dataset. Naturally more melodies are returned than before for the same query, and the number increases as more mismatches are allowed. If only four songs are to be returned, users must now sing on average 8, 11, and 14 notes respectively. Using approximate contour matching the values increased further to 11, 15 and more than 20.

The dynamic programming algorithm yields similar trends. However, the two are not equivalent because of differences in how the cost function is calculated.

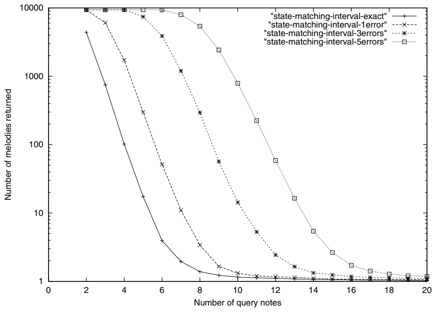


Fig. 3. Exact versus approximate interval matching for the state-based technique with duration, matching anywhere in melody.

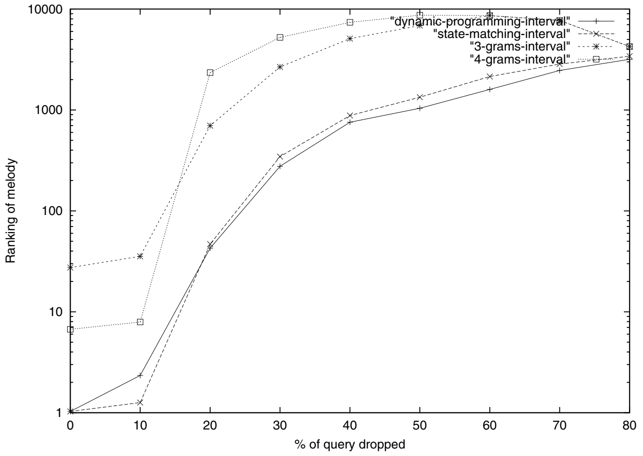
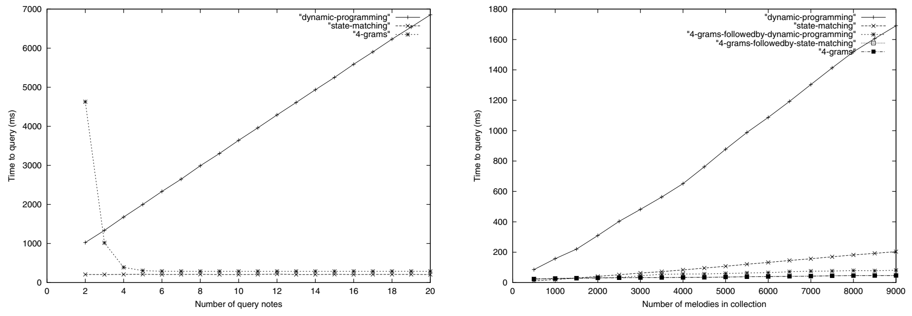


Fig. 4. Simulation of dropping notes from queries with approximate matching of intervals and duration used.

3.4 Ranking

In the uniqueness experiments, the sample queries were replicated excerpts of the sought-after tune in the database. This does not reflect the reality of a digital music library, where the user's query may not match any tune in the database for many reasons—different arrangements of the same song, imprecise recollection, the original does not exist in notated form, and so on. This section describes an experiment that takes into account such errors. It establishes how useful a matching algorithm is from a user's perspective by studying tune ranking, that is, the position in which the sought-after melody appears in the list of returned melodies.



(a) (b)

Fig. 5. Computational cost for matching anywhere in query with duration and ignoring rests (a) increasingly long queries with exact interval matching (b) increasingly large collections with contour matching.

Figure 4 shows the ranked position of the sought-after melody when some notes are omitted from the query, which is originally ten notes long. The x-axis shows the percentage of notes omitted, and values are averaged over 100 queries.

For dynamic programming, state-based matching and 3-grams the sought-after melody appears in the top 10 when zero or one note is missing from the query. Beyond that point the plots rise steeply indicating a marked worsening of the melody's ranked position. The same trend appears in the 4-gram method, only its ranked scores are notably poorer.

Towards the end of the graph (percentage error 70%–80%) both the 3- and 4-grams improve slightly. However, the improvement is of little practical import because it moves the sought-after tune up to rank 3,000—still very far from the start of the list. To see why the effect occurs, consider the size of the n-gram at this point. In the case of 4-grams, by the time the experiment has dropped 7 or more notes, the query is shorter than the length of the n-gram. In order to support queries of two or three notes—something a user could reasonably expect—we modified the n-gram algorithm to also calculate 3- and 2-grams, and it is this part of the index that performs the matching in the experiment's final stages.

Repeating the experiment for contour matching produces a similar graph to Figure 4. The main difference is a softening of the gradient when moving from one to two missing notes. This is because in all cases the ranked position when one note is missing is higher than the equivalent interval ranked position, but the when two notes are missing the ranked positions are comparable.

3.5 Computational Cost

Now we examine the computational cost of the three methods. Figure 5(a) shows the stark difference between dynamic programming and state-based matching.

The 4-gram method's cost is high at first, but quickly reduces before stabilizing at a consistent value that is a little larger than that for state matching. The initial expense is caused, once again, by the fact that when the length of the n-gram (4) exceeds that of the query, we drop down to 3-grams and 2-grams.

An approximate version of the dynamic programming algorithm gives exactly the same results, since runtime is not proportional to the error value. An approximate version of state-based matching gives values that are slightly greater than for exact matching, depending on the degree of error. This is because the runtime complexity of this algorithm is directly proportional to the error value.

The great efficiency, but disappointing effectiveness, of n-grams leads one to consider whether they can be used as a pre-filtering step for the more expensive, but more effective, dynamic programming and state-based matching algorithms. The effect of this is shown in Figure 5(b).

In this experiment the collection size was varied from 500 songs to 9000 songs in 500 song increments. For each collection size approximate matching versions of the dynamic programming and state-based matching algorithms (allowing up to two mistakes) were run and the time taken compared with versions that used 4-gram matching to prefilter the dataset.

The difference observed between dynamic programming and state-based is marked, with dynamic programming taking on average 8 times longer to perform a match. Although it should be remembered that the former is unrestricted in the number of mismatches that can occur, whereas the version of the state-based matching tested allowed only two mistakes.

The cost of matching 4-grams stays consistently low. Although it is hard to make out in Figure 5(b) due to the high density of lines, by the time the dataset contains 1,500 melodies, its cost is cheaper than state matching. For the hybrid methods, once the collection size has crossed the same threshold 1,500 it too represents a faster search that state matching alone. Both versions of the hybrid algorithm fall between the lines plotted for 4-gram state-based matching solutions: 4-gram followed by dynamic programming is roughly twice as time-consuming as the 4-gram method alone; 4-gram followed by state-based matching is so close to the 4-gram base line it cannot be visually distinguished.

4 Conclusion

We conclude this paper by relating the outcomes of our experimentation to forming a digital music library.

The uniqueness experiments—Figures 2(a)–3—help gauge how many notes make a useful query for a digital music library. Turning the issue around, having determined the typical number of notes sung by a user in a query, what are the implications of selecting a particular matching algorithm with certain parameter settings? The ranking experiment—Figure 4—helps gauge how much sifting through the list of returned songs may be required by the user to locate the sought-after melody. Together these experiments help place limits on what parameters are acceptable for a given matching algorithm in a digital library.

Say the designer of a digital music library expects users to sing queries with around 6–8 notes, and—because users rarely go beyond the first page of search results [JCMB00]—wants the sought-after melody to be ranked in the top 10. The graphs of uniqueness and ranking show that this rules out performing any contour matching without duration. It also rules out state-based matching with three or more errors,[1] and 3-grams. This leaves interval matching with or without duration, and contour matching with duration, as strong candidates for use, and 4-grams with interval and duration as a tolerable option.

Issues of computational efficiency are revealed by the third set of experiments. Greater speed without compromising accuracy is a strong factor driving the implementation of a digital music library. Figure 5(b) shows that there is a definite advantage in a digital library system using a hybrid approach, particularly in the case of pre-filtering the dynamic programming algorithm. The initial high cost of n-gram matching for queries with fewer notes than the basic n-gram size is of minor concern and can be handled by warning the user, if they issue such a query, that so few notes are likely to return a high number of songs and asking if they wish to continue.

This discussion applies to the folksong dataset. These recommendations can probably be extrapolated to other collections, but some caution must be exercised. The dataset used is monophonic, based on notated form, and is from one genre—folk music. What happens if the genre is different? What happen if the dataset is sourced from MIDI files, a readily available form of music but one that is much noisier. For instance, duration information is less reliable because the song is typically entered using a synthesizer keyboard, and passages of the music that repeat are played out rather than appearing once and being notated as repeating. Further experimentation is required to understand how such changes alter the requirements of a digital music library, and what better way to manage this than through a workbench for music information retrieval!

References

[AB97] R. Arnold and T. Bell. A corpus for the evaluation of lossless compression algorithms. In *Designs, Codes and Cryptography*, pages 201–210, 1997.

[BB01] D. Bainbridge and T. Bell. The challenge of optical music recognition. *Computers and the Humanities*, 2001.

[BD00] D. Bird and J.S. Downie, editors. *Proceedings of the 1st. Int. Symposium on Music Information Retrieval: ISMIR 2000*, Plymouth, Massachusetts, USA, 2000. Available through *www.music-ir.org*.

[BKM98] C. Blake, E. Keogh, and C.J. Merz. *UCI Repository of Machine Learning Databases.* http://www.ics.uci.edu/~mlearn/mlrepository.html, University of California, Department of Information and Computer Science, Irvine, CA, Irvine, CA, USA, 1998.

[BNMW^{+}99] D. Bainbridge, C. Nevill-Manning, I. Witten, L. Smith, and R. McNab. Towards a digital library of popular music. In *The 4th ACM conference on Digital Libraries*, pages 161–169, 1999.

[1] A similar cutoff threshold for dynamic programming can also be determined. However, the necessary graph is not shown in the paper and the value is specific to the cost functions used to calculate edit distance.

[DB01] J.S. Downie and D. Bainbridge, editors. *Proc. of the 2nd Int. Symposium on Music Information Retrieval*, Indiana University, Bloomington, IN, USA, 2001. Available through *www.music-ir.org*.

[Dow99] J.S. Downie. *Evaluating a Simple Approach to Musical Information Retrieval: Conceiving Melodic N-Grams as Text*. PhD. thesis, University of Western Ontario, Canada, 1999.

[GR69] B. Gold and L. Rabiner. Parallel processing techniques for estimating pitch periods of speech in the time domain. *J. Acoust. Soc. Am.*, 46(2):442–448, 1969.

[HRG01] H. Hoos, K. Renz, and M. Gorg. GUIDO/MIR: An experimental musical information retrieval system based on guido music notation. In J. Stephen Downie and David Bainbridge, editors, *Proc. of the 2nd Int. Symposium on Music Information Retrieval: ISMIR 2001*, pages 41–50, 2001.

[JCMB00] S. Jones, S.J. Cunningham, R.J. McNab, and S. Boddie. A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3(2):152–169, 2000.

[McN96] R. McNab. *Interactive applications of music transcription*. MSc thesis, Department of Computer Science, University of Waikato, NZ, 1996.

[MS90] M. Mongeau and D. Sankoff. Comparison of musical sequences. *Computers and the Humanities*, pages 161–175, 1990.

[Par75] D. Parsons. *The Directory of Tunes and Musical Themes*. Spencer Brown, Cambridge, 1975.

[RB01] W. Rand and W. Birmingham. Statistical analysis in music information retrieval. In J. Stephen Downie and David Bainbridge, editors, *Proc. of the 2nd Int. Symposium on Music Information Retrieval*, pages 25–26, Indiana University, Bloomington, IN, USA, 2001.

[WBB01] I. Witten, D. Bainbridge, and S. Boddie. Greenstone: open source dl software. *Communications of the ACM*, 44(5):44, 2001.

[WM92] S. Wu and U. Manber. Fast text searching allowing errors. *Communications of the ACM*, 35(10):83–91, 1992.

[WMB99] I. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: compressing and indexing documents and images*. Morgan Kaufmann, San Francisco, CA, 1999.

[WRBB00] I. Witten, McNab R., S. Boddie, and D. Bainbridge. Greenstone: a comprehensive open-source digital library software system. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 113–121, San Antonio, Texas, June 2000.

The NUS Digital Media Gallery – A Dynamic Architecture for Audio, Image, Clipart, and Video Repository Accessible via the Campus Learning Management System and the Digital Library

Phoon Moi Ying[1] and Junius Soh Hock Heng[2]

[1]National University of Singapore, Central Library, 12 Kent Ridge Crescent,
Singapore 119275
moiying@nus.edu.sg

[2]National University of Singapore, Centre for Instructional Technology, Kent Ridge Drive 4,
Level 3, Computer Centre, Singapore 119275
citsohhh@nus.edu.sg

Abstract. This paper describes the development of a digital media gallery (DMG) that allows users to search, preview digital contents, and link them to the course website within the Integrated Virtual Learning Environment (IVLE). The IVLE is a web-based learning management system designed and developed by the Centre for Instructional Technology (CIT) at the National University of Singapore (NUS) to manage and support teaching and learning. The gallery is a joint collaboration between CIT and the NUS Library, applying a user oriented perspective to developing, organizing and using digital resources for teaching and research. The digital resources are broadly categorized into Audio, Video, Image and clipart. Within each grouping, the contents are classified and meta tagged according to the IMS specification. For video recordings, an interface developed by VIDTOOLS, a commercial company, was used to facilitate efficient navigation of video content via keyframe indexing, with a time slider to support quick skimming through the recording.

1 Introduction

The purpose in developing a Digital Media Gallery (DMG) is to bring together digital works created by the various teaching faculties within the campus to a central repository. The DMG provides the storage space for the preservation of the works and a classification scheme to facilitate access and retrieval. A metadata scheme of tagging according to the IMS specification, with an alternative to XML representation, is adopted to facilitate searching and interoperability between systems. An interface is provided for teaching staff to select digital resources from the DMG for integration into the campus learning management system called IVLE (Integrated Virtual learning Environment). A similar interface will also be developed to provide access to the DMG resources via the Digital Library. It is a collaborative attempt among faculties, systems engineers and librarians to develop, organize and use digital resources for research and instruction from a user oriented perspective.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 141-152, 2002.

2 Objectives of Developing a DMG

Technology has created new opportunities for librarians, teaching staff, technologists and instructional developers to work closely to create new learning environments to support instructional and research needs. The DMG is a collaborative attempt to achieve four main objectives.

- The first objective is to provide a central repository for digital resources for the campus community, to facilitate resource sharing and to provide storage space for the preservation of works.
- The second objective is to provide a tagging system to manage the digital resources to facilitate search and retrieval. The contents are IMS tagged with an alternative XML representation, as XML representation is the next generation of markup language for internet media as recommended by W3C
- The third objective is to provide an interface for teaching staff to select from a wider variety of media and integrate the digital resources into the course multimedia bank within the learning management system called IVLE. A similar interface will be developed in the future to allow a one stop search facility across all library resources and the DMG
- The fourth objective is to centralize the processes of procurement, storage and management of digital resources across the faculties. This will improve the level of service to support teaching and learning. The library will streamline the workflow for archiving, manage the digital media, and provide end users with important information about the media content. CIT as co-partner will anticipate the adoption of new technologies in this changing field and set up the IT equipment and infrastructure to support the delivery of content in the DMG.

3 Building the Digital Archives

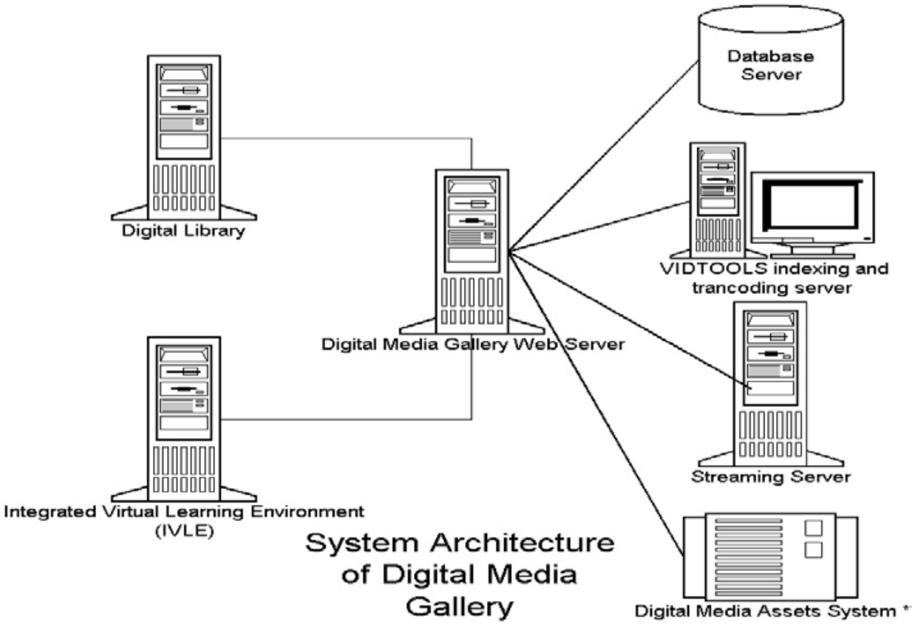
There are currently four categories of resources in the DMG, with room for future expansion to include other formats. The four categories are audio, video, images and clipart files. A set of guidelines on the audio and video materials to be included in the DMG was drawn up. All existing digitised audio and video recordings in the on-demand streaming service were migrated into the DMG. Also included were in-house video recordings of speeches by distinguished visitors and interviews by or hosted by faculty staff. Commercially purchased recordings identified for course use with licenses paid for were also included.

Holding road shows on the DMG in teaching departments is our strategy to identify collections held by the teaching departments. For departments which have established their mini repositories, CIT will convince them of the benefits of resource sharing and offer parallel linking from the DMG and their website. For departments without any formal organization of resources, assistance will be offered to advise them on copyright issues and to provide technical support to upload the contents into the DMG.

4 System Architecture

The Digital Media Gallery (DMG) consists of the following components:

- *Web server*. The web server contains the homepages for DMG and the administration page for managing DMG. It is also the repository for images and clipart. The administration web page provides a means for submission and meta-tagging of media into DMG. The Library, being the administrator of the system, is responsible for ensuring that the resources are tagged according to the IMS Specification and adhere to copyright
- *Streaming Server*. The streaming server stores the audio and video content of DMG. Content is delivered to clients via streaming techniques using Windows Media Technology and Real, with plans in the pipeline to experiment with Quicktime
- *Database Server*. The database server contains all the media parameters such as title, description, keywords, IMS metadata, etc.
- *Indexing and transcoding server*. This server takes in MPEG files for keyframe extraction for indexing as well as transcoding to a streaming format such as Windows Media and Real to be transferred to the streaming server.
- *Digital Media Assets System*. This system houses all the archival copies of the audio, video and images that are found in DMG. The audio and video content are digitized at high quality for archiving purposes as well as for transcoding to a streaming form in DMG. As for images, they are scanned at high dpi for archiving, and sent to DMG at a lower resolution.

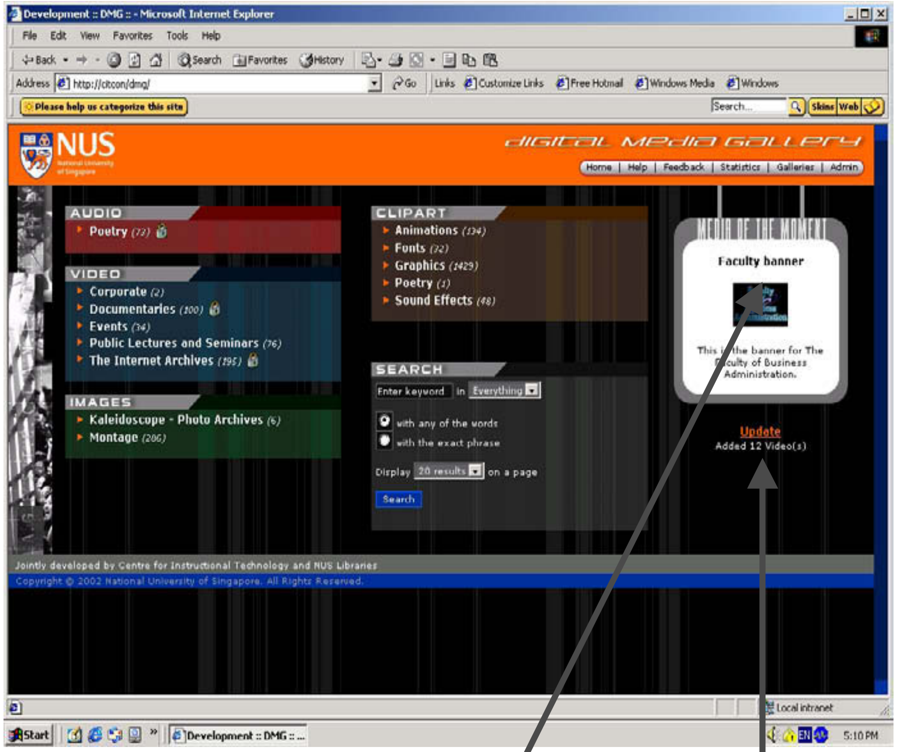

Database Server
Digital Library
VIDTOOLS indexing and trancoding server
Digital Media Gallery Web Server
Streaming Server
Integrated Virtual Learning Environment (IVLE)
Digital Media Assets System *

System Architecture of Digital Media Gallery

5 Digital Contents and Layout of the Digital Media Gallery

The digital contents are broadly grouped into Audio, Images, Clipart and Video. The contents consist of files created by the campus community and resources acquired by the Library. The licenses of purchased digital resources are cleared for use before inclusion in the portal.

There are three levels of categorization. The first level is by format, grouped broadly into Audio, Images, Clipart and Video. The second level groups each format by genre/form. The third level of classification is metadata tagging. (Fig. 1).


Development :: DMG :: - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://citcon/dmg/
Go Links Customize Links Free Hotmail Windows Media Windows
Please help us categorize this site
Search... Skins Web
NUS
National University of Singapore
digital media gallery
Home | Help | Feedback | Statistics | Galleries | Admin
Local intranet
Start
Development :: DMG :: ...
EN
5:10 PM

AUDIO
- Poetry (73)

VIDEO
- Corporate (2)
- Documentaries (100)
- Events (34)
- Public Lectures and Seminars (76)
- The Internet Archives (195)

IMAGES
- Kaleidoscope - Photo Archives (6)
- Montage (206)

CLIPART
- Animations (134)
- Fonts (32)
- Graphics (1429)
- Poetry (1)
- Sound Effects (48)

SEARCH

Enter keyword in Everything

● with any of the words

○ with the exact phrase

Display 20 results on a page

Search

MEDIA OF THE MOMENT

Faculty banner

This is the banner for The Faculty of Business Administration.

Update

Added 12 Video(s)

Jointly developed by Centre for Instructional Technology and NUS Libraries

Copyright © 2002 National University of Singapore. All Rights Reserved.

Fig. 1. Digital resources grouped by genre/form

Media of the Moment randomly display media from repository

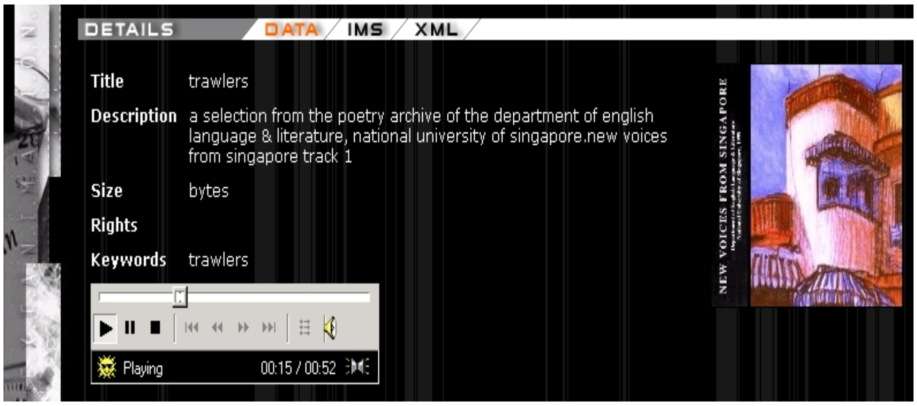
Showing the latest collection

5.1 Audio Files

Audio files range from poetry recitations to recordings of interviews. The contents are delivered to the users via streaming technology. The policy on what audio materials to incorporate into DMG and what to archive has yet to be formulated, as efforts have been concentrated on developing and establishing the video component of the gallery first. (Fig. 2)

5.2 Images

The image collection consists of photographs taken by library photographers and staff of CIT, as well as photograph collections deposited with the library by individuals, teaching departments and external organizations. The bulk of these photographs are without negatives. The photographs were scanned at 300 dpi for archiving and 100 dpi for DMG use. The library started identifying and documenting the photographs from two separate collections. One set of photographs on the *historical past* of the University contains photographs taken at university related events as well as personalities. The other set on *Vanishing Scenes of Singapore* is of buildings earmarked for demolition. The identification and documentation of the photographs were most time consuming. (Fig. 3)



Title trawlers

Description a selection from the poetry archive of the department of english language & literature, national university of singapore.new voices from singapore track 1

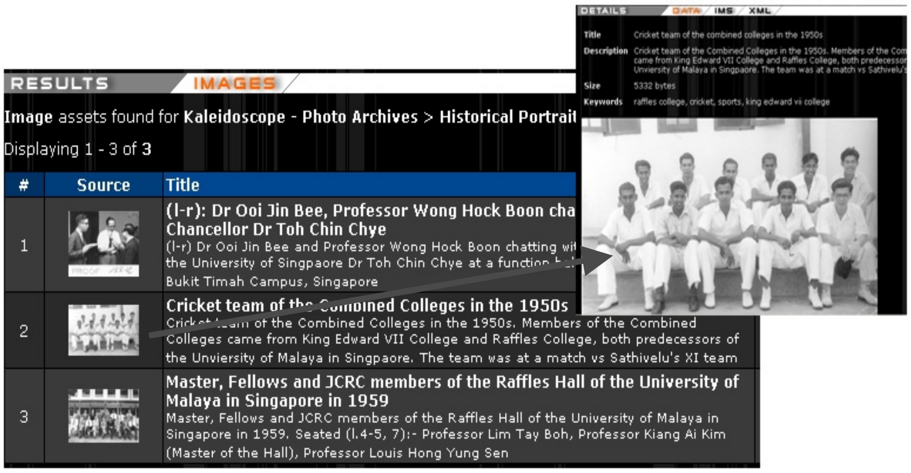
Size bytes

Rights

Keywords trawlers

Playing 00:15 / 00:52

Fig. 2. An example of an audio file display and playback from the poetry recital recording, taken from the poetry archives of the Department of English Language and Literature. The image on the right shows the CD cover



Title Cricket team of the combined colleges in the 1950s

Description Cricket team of the Combined Colleges in the 1950s. Members of the Com came from King Edward VII College and Raffles College, both predecessor Unviersity of Malaya in Singpaore. The team was at a match vs Sathivelu's

Size 5332 bytes

Keywords raffles college, cricket, sports, king edward vii college

Image assets found for **Kaleidoscope - Photo Archives > Historical Portrait**

Displaying 1 - 3 of **3**

#	Source	Title
1		**(l-r): Dr Ooi Jin Bee, Professor Wong Hock Boon cha Chancellor Dr Toh Chin Chye** (l-r) Dr Ooi Jin Bee and Professor Wong Hock Boon chatting wi the University of Singpaore Dr Toh Chin Chye at a function Bukit Timah Campus, Singapore
2		**Cricket team of the Combined Colleges in the 1950s** Cricket team of the Combined Colleges in the 1950s. Members of the Combined Colleges came from King Edward VII College and Raffles College, both predecessors of the Unviersity of Malaya in Singpaore. The team was at a match vs Sathivelu's XI team
3		**Master, Fellows and JCRC members of the Raffles Hall of the University of Malaya in Singapore in 1959** Master, Fellows and JCRC members of the Raffles Hall of the University of Malaya in Singapore in 1959. Seated (l.4-5, 7):- Professor Lim Tay Boh, Professor Kiang Ai Kim (Master of the Hall), Professor Louis Hong Yung Sen

Fig. 3. Samples from the Historical Past of the University photograph collection

5.3 Clipart

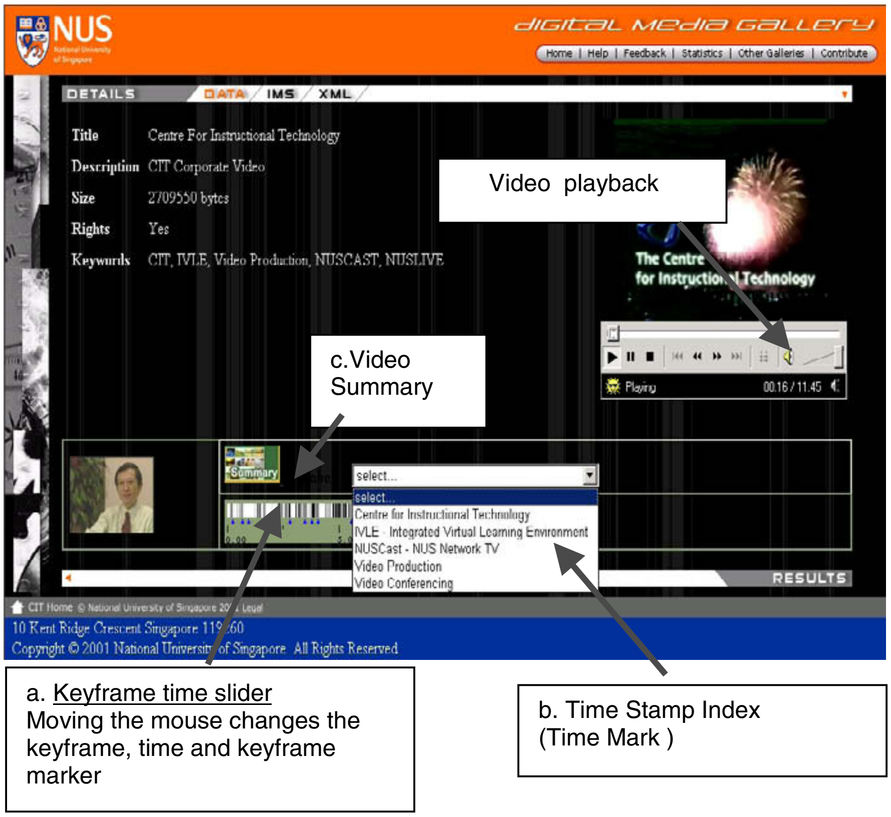
These resources are mounted for use in powerpoint presentations and in the design of web pages. The digitized files include Animation, Fonts, Graphics and Sound Effects.

5.4 Video Files

Currently, there are four broad categories of video recordings : corporate videos of faculties and departments; documentaries and training programs; recordings of public lectures and seminars organized by the University; and licence free video recordings downloaded from the internet or works from NUS staff. All the analog video recordings are digitized at high quality for archiving and are also preserved on tape. The video contents are then transcoded into streaming format viewer for the gallery.

Visual Interfaces

The Centre for Instructional Technology collaborated with VIDTOOLS[1] to create three interfaces for navigating video recordings.


NUS
National University of Singapore
digital media gallery
Home | Help | Feedback | Statistics | Other Galleries | Contribute
DETAILS
DATA
IMS
XML
The Centre for Instructional Technology
Playing
00:16 / 11:45
Summary
select...
RESULTS
CIT Home © National University of Singapore 20_ Legal

Title Centre For Instructional Technology
Description CIT Corporate Video
Size 2709550 bytes
Rights Yes
Keywords CIT, IVLE, Video Production, NUSCAST, NUSLIVE

select...
Centre for Instructional Technology
IVLE - Integrated Virtual Learning Environment
NUSCast - NUS Network TV
Video Production
Video Conferencing

10 Kent Ridge Crescent Singapore 119260
Copyright © 2001 National University of Singapore All Rights Reserved

Video playback

c. Video Summary

a. Keyframe time slider
Moving the mouse changes the keyframe, time and keyframe marker

b. Time Stamp Index (Time Mark)

Fig. 4. The 3 visual interfaces to video recordings: a,b,c

[1] VIDTOOLS is the commercial company that developed the software

The three visual interfaces are :

- The *Keyframe time slider interface* gives a quick overall impression of the video's content, as moving the mouse through the slider bar changes the keyframes (or still images extracted from the video). This interface allows very quick skimming through the video. Clicking anywhere on the timeline opens the video and starts playback at the corresponding time.
- The *Time stamp index* provides a means to segment video content based on the time indices. The video can be indexed into chapters 1,2, 3, etc.
- The *Video Summary* provides another form of overview for the video, based on video shots demarcated by scene or camera shot changes. The captured keyframes are sized according to the importance measure of the segment (e.g. duration), with larger keyframes representing higher-importance segments. An overview of the video's content is displayed on a single page as a visual summary. Once the user identifies an interesting segment, clicking on its keyframe starts video playback from the beginning of that segment. (Fig. 5)

Video Keyframe Indexing

Besides searching for video content in DMG using video descriptions and titles, VIDTOOLS provided the means to search within a video segment based on the time mark (or time stamp index) and keyframe indices. Thus users are able to search for short clips within a video segment. (Fig. 6)

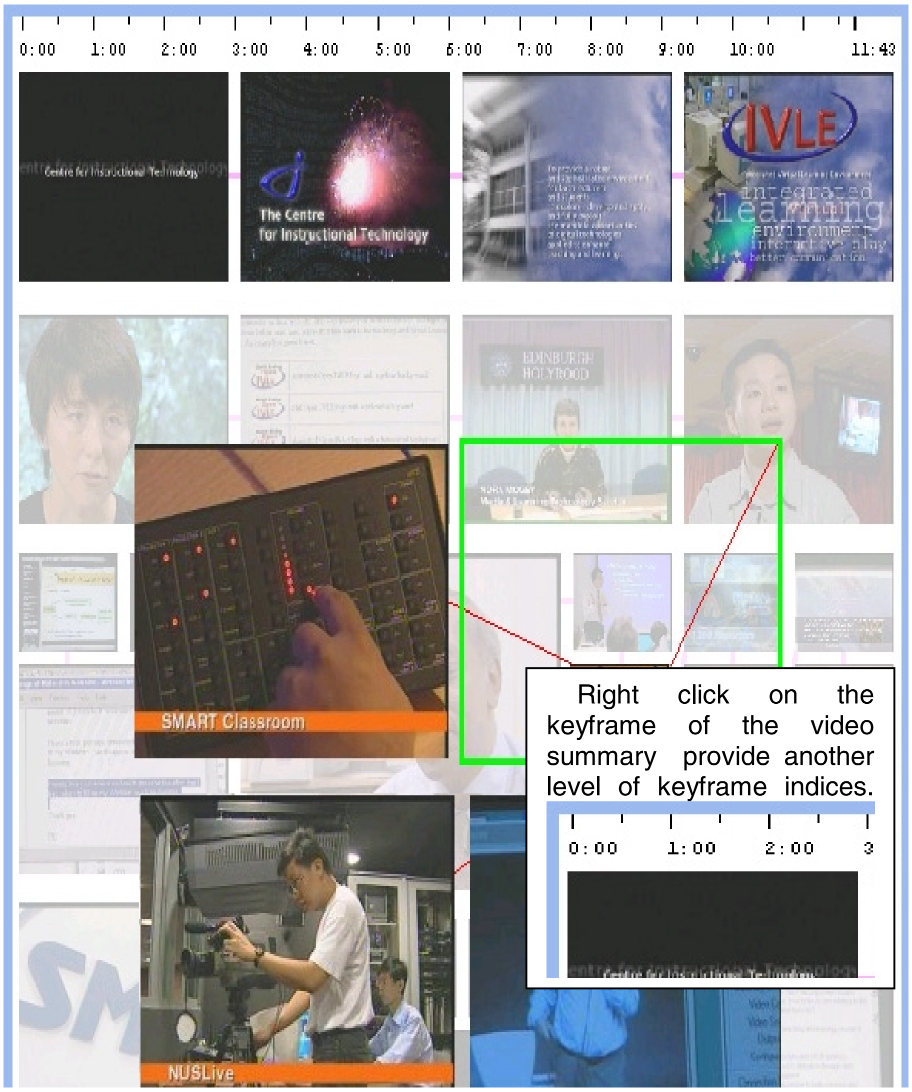
5 Indexing Content with Metadata Tags

5.1 Metadata

The digital resources are considered learning objects for integration into the campus learning management system.

The SingCORE schema of meta tagging is adopted to ensure interoperatiblity, convenience, flexibility and efficiency in design, delivery and administration. SingCORE is Singapore's learning resource identification specification. It is a streamlined subset of elements from the IMS Learning Resources Meta-data Specificiation version 1.2.2, customized for the Singapore context. As of July 2002, the IMS metadata specification was accepted by the IEEE as a standard. SingaCORE has eight categories and 42 elements, and its key vocabularies are defined for the education environment in Singapore.

The descriptive terms are easy to find and use, providing users with a more efficient structure with elements to describe the learning resource. The contents are described so that important characteristics of the resource (including ownership, terms and conditions of use, restrictions on access and copyright status) are given. A standard controlled vocabulary will be documented in the process of tagging so that standard terms are consistently applied with reference to Library of Congress subject headings and the recommended taxonomy for the Singapore Institutions of Higher Learning vocabulary. (Fig.7)


0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:43
Centre for Instructional Technology
The Centre for Instructional Technology
IVLE
integrated learning environment interactive play better communication
EDINBURGH HOLYROOD
SMART Classroom
NUSLive
0:00
1:00
2:00
3

Right click on the keyframe of the video summary provide another level of keyframe indices.

Fig. 5. Video Summary display

As the keyword search facility allows searching through the elements, this makes duplication of resources less likely. The metadata template gives creators and end users a common template to describe the resources. The process of finding and using a resource becomes more efficient, since the structure predefines the elements to describe (or catalogue) the contents, along with requirements about how the elements are to be used . The library will be responsible for the selection of resources and for checking the metadata tagging to ensure consistency and uniformity.

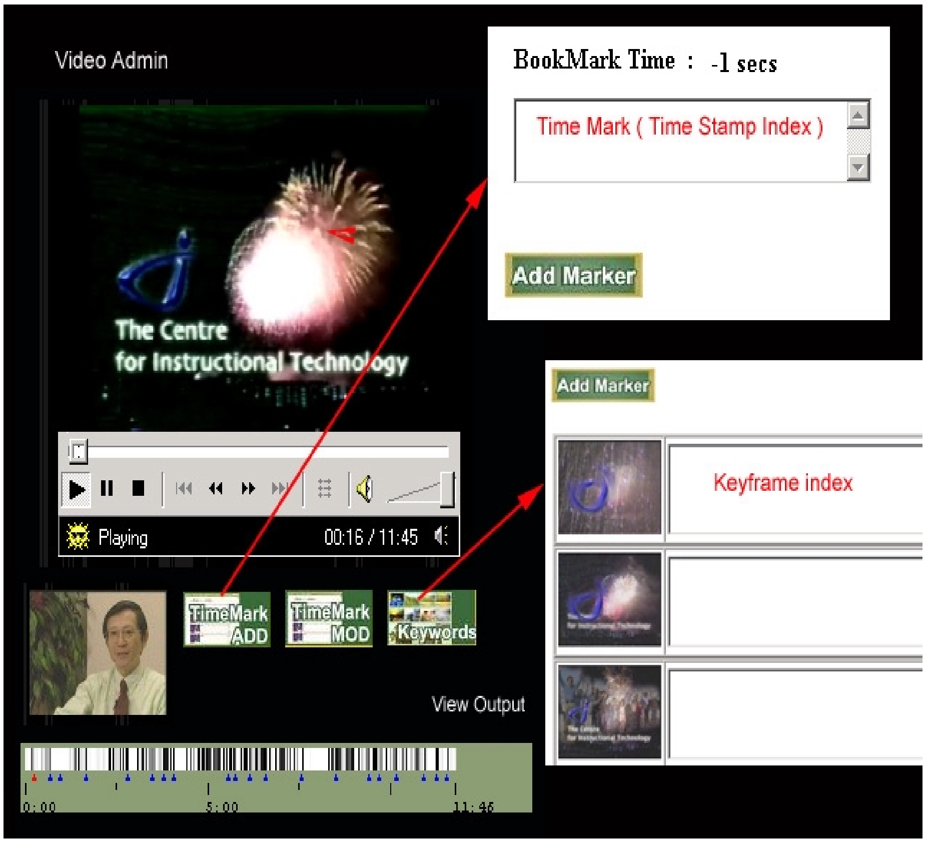


Fig. 6. Video keyframe indexing

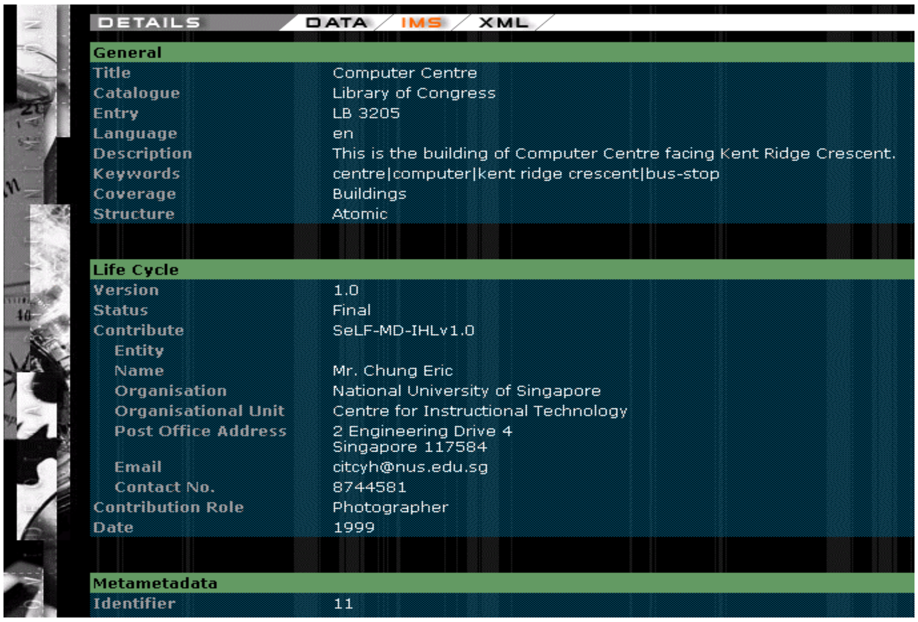
5.2 XML

To read the metadata, XML is used to bind IMS information together. Potentially, these bindings are used to exchange information with similar systems. Being IMS ready provides the Digital Media Gallery with an avenue to effectively grow in terms of resources. (Fig. 8)

6 Accessing the Digital Media Gallery (DMG)

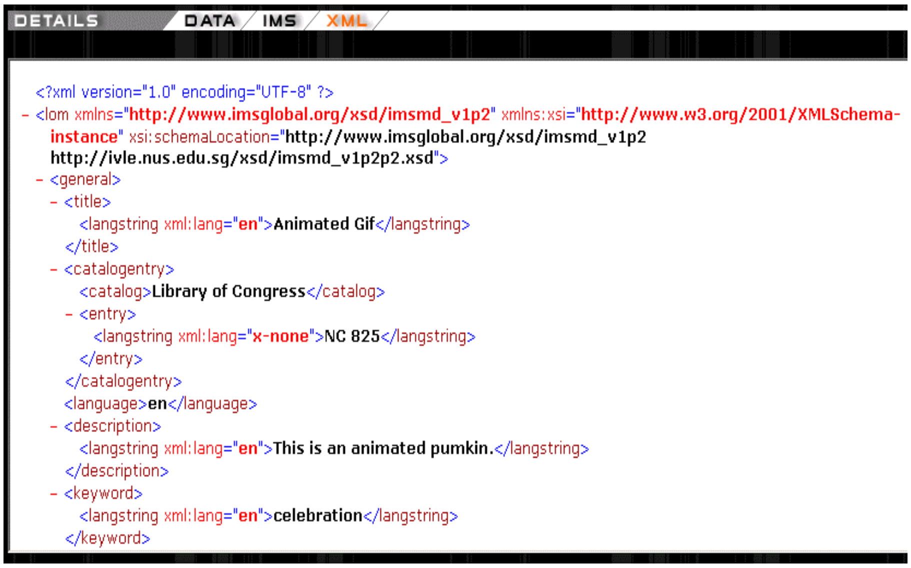
The contents of the Digital Media Gallery are accessible via the following means.

- At the *Digital Media Gallery* homepage, users can browse through the media categories or use the search facility. He can search across all formats or limit to a particular format, use one word or an exact phrase, and limit results from 10 to 50 hits on one page. (Fig. 1)



General	
Title	Computer Centre
Catalogue	Library of Congress
Entry	LB 3205
Language	en
Description	This is the building of Computer Centre facing Kent Ridge Crescent.
Keywords	centre|computer|kent ridge crescent|bus-stop
Coverage	Buildings
Structure	Atomic
Life Cycle	
Version	1.0
Status	Final
Contribute	SeLF-MD-IHLv1.0
Entity	
Name	Mr. Chung Eric
Organisation	National University of Singapore
Organisational Unit	Centre for Instructional Technology
Post Office Address	2 Engineering Drive 4 Singapore 117584
Email	citcyh@nus.edu.sg
Contact No.	8744581
Contribution Role	Photographer
Date	1999
Metametadata	
Identifier	11

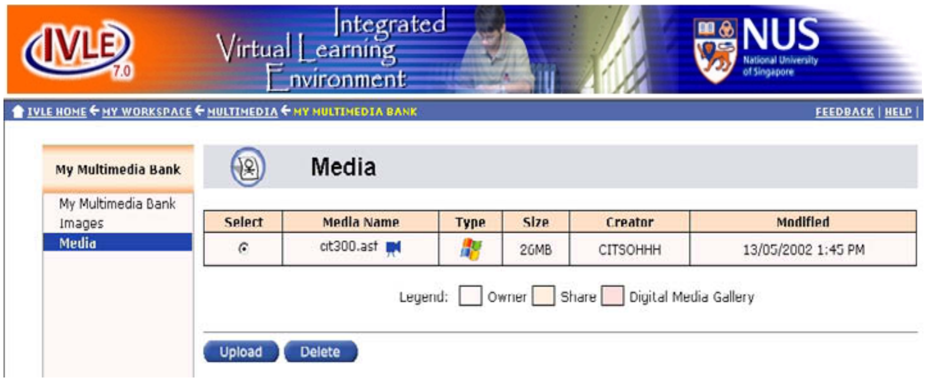
Fig. 7. IMS metadata elements



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <lom xmlns="http://www.imsglobal.org/xsd/imsmd_v1p2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
    instance" xsi:schemaLocation="http://www.imsglobal.org/xsd/imsmd_v1p2
    http://ivle.nus.edu.sg/xsd/imsmd_v1p2p2.xsd">
  - <general>
    - <title>
        <langstring xml:lang="en">Animated Gif</langstring>
      </title>
    - <catalogentry>
        <catalog>Library of Congress</catalog>
      - <entry>
          <langstring xml:lang="x-none">NC 825</langstring>
        </entry>
      </catalogentry>
      <language>en</language>
    - <description>
        <langstring xml:lang="en">This is an animated pumkin.</langstring>
      </description>
    - <keyword>
        <langstring xml:lang="en">celebration</langstring>
      </keyword>
```

Fig. 8. XML indexing

- *Integrated Virtual Learning Environment (IVLE)* . A user can search the DMG database from the IVLE and upload from the multimedia bank into the course website. Students enrolled in the course will have access to multimedia files anywhere and anytime from a desktop. Access by users is logged and statistics can be generated monthly or on an ad hoc basis. IVLE contains a Multimedia Bank for lecturers to upload their own media (images, audio, video) to enhance teaching and learning. With the DMG, lecturers will be able to access a wider variety of media shared across the campus. They are able to browse the content of DMG and select the necessary media links to be inserted into their own multimedia bank. (Fig. 9)
- *Digital Library.* An interface to search the DMG via the Digital Library will be developed in the near future to provide a one-stop search facility that works across all library resources. These include databases subscribed to by the library, catalogues , electronic resources and the DMG.



Select	Media Name	Type	Size	Creator	Modified
●	cit300.ast		26MB	CITSOHHH	13/05/2002 1:45 PM

Legend: □ Owner □ Share □ Digital Media Gallery

Fig. 9. IVLE (Integrated Virtual Learning Environment)

8 Conclusion

The success of the Digital Media Gallery (DMG) will depend on the teaching faculties agreeing to deposit their digital works with the repository and also on the Library managing the resources well. The ease of integrating resources from the DMG to the IVLE (Integrated Virtual Learning Environment) is an attractive feature of the DMG. The interfaces developed for video recordings (in terms of visual content and time indexing) help users to navigate and extract information from the recordings. As the DMG grows, there will be opportunities to develop more interfaces to access the digital resources in the gallery.

References

1. Singapore Standard SS 496 : Part 1: 2001 Specification for e learning Framework. Part 1- An overview.

2. Singapore Standard SS 496 : Part 2 : 2001 Specification for e learning Framework. Part 2 - Learning Resources Identification.
3. Keyframe-based user interfaces for digital video. IEEE Computer 34(9), 61-67 (2001).
4 http://www.imsglobal.org
5 http://courseware.nus.edu.sg/Standards

A Schema Language for MPEG-7

Ano Chotmanee, Vilas Wuwongse, and Chutiporn Anutariya

Computer Science & Information Management Program
Asian Institute of Technology
Klong Luang, Pathumtani 12120, Thailand
{a00602,vw,ca}@cs.ait.ac.th

Abstract. MPEG-7 provides a set of tools for the definition of a standard description for multimedia information. It uses XML Schema as its own Description Definition Language (DDL) to provide an ability to extend standard description to user application domains. However, due to its lack of expressive power, XML Schema cannot represent the implicit information of multimedia contents and hence cannot yield derived descriptions. In order to enhance the expressive power of multimedia annotations—described in terms of an XML document—and to maintain the advantages of XML Schema, there appears to be a need of a new schema language which can handle those abilities. XML Declarative Description (XDD) is a representation framework which can encode any XML document and application (e.g., XML Schema, RDF, RDF Schema, and DAML+OIL). It enhances XML expressive power by employing Declarative Description (DD) theory and also provides a mechanism for reasoning about new information from existing data. Moreover, it can specify constraints as well as ontological axioms. Thus, it is employed to construct a schema language for MPEG-7.

1 Introduction

In recent years, the amount of multimedia information has been increasing both in analog and digital formats and has begun to play an important role in our lives. The popularity of Internet makes the exchange of multimedia information in digital archives easy and many companies provide streaming broadcasting services for their customers on Internet, such as news, music, music video and movie trailer. In order to search, filter, access, retrieve, and manage multimedia contents, metadata are necessary. MPEG-7 (Multimedia Content Description Interface) [10]—an answer to this requirement—is a standard under development by MPEG (Moving Picture Experts Group) which will provide tools to describe multimedia contents. MPEG-7 has chosen XML Schema (XMLS) [3, 6, 11] as its Description Definition Language (DDL) because of XMLS's ability to express syntactic, structural, cardinality and data-type constraints required by MPEG-7. However, XMLS lacks an ability to represent implicit information of multimedia contents and an inferential mechanism needed to obtain derived information from other explicit descriptions. Therefore, a new data model must overcome this inability. Reference [8] has proposed to employ RDF [9], RDF Schema (RDFS) [5] and DAML+OIL [7] to deal with limitation. However, RDF and RDFS have limited specification of multiple range constraints on single proper-

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 153-164, 2002.

ties and lack an inferential mechanism to support reasoning. Even if these limitations can be resolved by using DAML+OIL markup language ontology, DAML+OIL itself has limited capabilities to express arbitrary constraints and axioms.

In order to maintain the advantages of XMLS as well as to enhance the expressive power of MPEG-7's Descriptors (Ds) and Description Schemes (DSs), a new schema language is proposed. By employing XML Declarative Description (XDD) [12] theory, the new schema language can preserve the semantics of MPEG-7 descriptions as well as provide reasoning mechanism. In order to increase interoperability among the existing markup languages—RDF, RDFS, and DAML+OIL—the new schema will employ XML [4] as its syntax, extend ordinary XML elements with variables and model relationships among XML elements by *XML clauses*.

Section 2 explains XDD theory, Section 3 introduces MPEG-7 standard, Section 4 proposes the new schema language for MPEG-7, and Section 5 concludes and discusses future work.

2 XML Declarative Description (XDD)

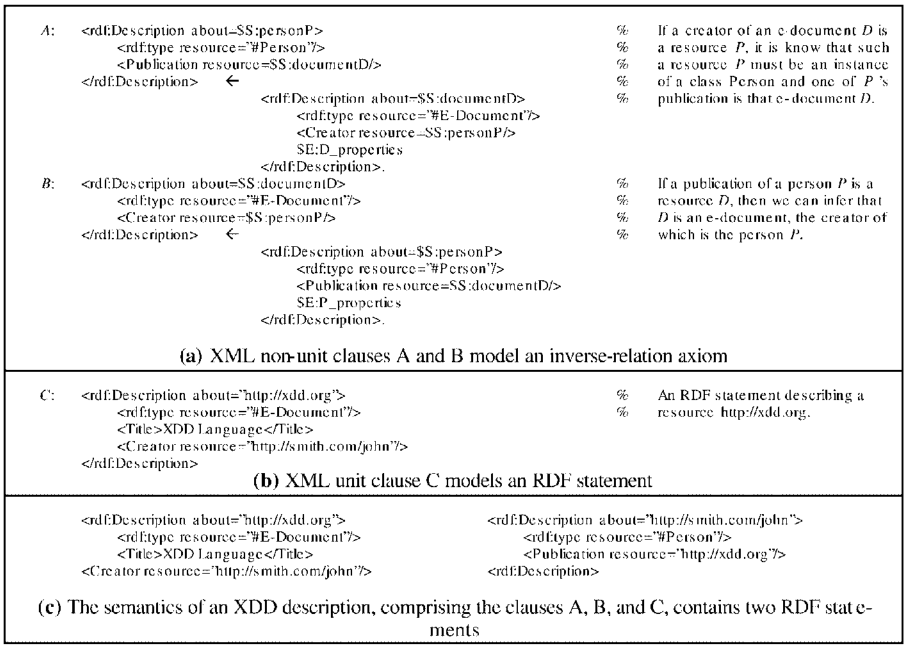
XML Declarative Description (*XDD*) [12] is an XML-based information representation, which extends ordinary, well-formed XML elements by incorporation of variables for an enhancement of expressive power and representation of implicit information into so called *XML expressions*. Ordinary XML elements—XML expressions without variable—are called *ground XML expressions*. Every component of an XML expression can contain variables, e.g., its expression or a sequence of sub-expressions (*E-variables*), tag names or attribute names (*N-variables*), strings or literal contents (*S-variables*), pairs of attributes and values (*P-variables*) and some partial structures (*I-variables*). Every variable is prefixed by ‘$*T*:’, where *T* denotes its type; for example, $S:value and $E:expression are *S*- and *E*-variables, which can be specialized into a string or a sequence of XML expressions, respectively.

An *XDD description* is a set of *XML clauses* of the form:

$$H \leftarrow B_1, \dots , B_m, \beta_1, \dots, \beta_n,$$

where m, $n \geq 0$, H and the B_i are XML expressions, and each of the β_i is a predefined *XML constraint*—useful for defining a restriction on XML expressions or their components. The XML expression H is called the *head*, the set $\{B_1, \dots, B_m, \beta_1, \dots, \beta_n\}$ the *body* of the clause. When the body is empty, such a clause is referred to as an *XML unit clause*, otherwise a *non-unit clause*, i.e., it has both head and body. A unit clause ($H\leftarrow$.) is often denoted simply by H and referred to as a *fact*. Moreover, an XML element or document can be mapped directly onto a *ground XML unit clause*. The elements of the body of a clause can represent constraints. Therefore, XML clauses (both unit and non-unit ones) can express facts, taxonomies, implicit and conditional relationships, constraints, axioms as well as ontology – set of taxonomies together with their axioms. Given an XDD description D, its meaning is the set of all XML elements which are directly described by and are derivable from the unit and non-unit clauses in D, respectively.

Fig. 1 gives an example of an XML description and its semantics.



```
A:   <rdf:Description about=$S:personP>                                  %   If a creator of an e-document D is
         <rdf:type resource="#Person"/>                                  %   a resource P, it is know that such
         <Publication resource=$S:documentD/>                            %   a resource P must be an instance
     </rdf:Description>     ←                                            %   of a class Person and one of P 's
                        <rdf:Description about=$S:documentD>             %   publication is that e-document D.
                            <rdf:type resource="#E-Document"/>
                            <Creator resource=$S:personP/>
                            $E:D_properties
                        </rdf:Description>.
B:   <rdf:Description about=$S:documentD>                                %   If a publication of a person P is a
         <rdf:type resource="#E-Document"/>                              %   resource D, then we can infer that
         <Creator resource=$S:personP/>                                  %   D is an e-document, the creator of
     </rdf:Description>     ←                                            %   which is the person P.
                        <rdf:Description about=$S:personP>
                            <rdf:type resource="#Person"/>
                            <Publication resource=$S:documentD/>
                            $E:P_properties
                        </rdf:Description>.
```

(a) XML non-unit clauses A and B model an inverse-relation axiom

```
C:   <rdf:Description about="http://xdd.org">                            %   An RDF statement describing a
         <rdf:type resource="#E-Document"/>                              %   resource http://xdd.org.
         <Title>XDD Language</Title>
         <Creator resource="http://smith.com/john"/>
     </rdf:Description>
```

(b) XML unit clause C models an RDF statement

```
     <rdf:Description about="http://xdd.org">                  <rdf:Description about="http://smith.com/john">
         <rdf:type resource="#E-Document"/>                         <rdf:type resource="#Person"/>
         <Title>XDD Language</Title>                                <Publication resource="http://xdd.org"/>
     <Creator resource="http://smith.com/john"/>               <rdf:Description>
```

(c) The semantics of an XDD description, comprising the clauses A, B, and C, contains two RDF statements

Fig. 1. An example of an XML description

3 Multimedia Content Description Interface (MPEG-7)

3.1 Overview of MPEG-7

MPEG started a new working group called MPEG-7 in October 1996; its aims were to provide standardized tools for the description of multimedia contents. MPEG-7's metadata, attached to multimedia contents, do not depend on the manner in which described content is coded or stored, i.e., it is possible to create a MPEG-7 description

Descriptors (Ds): represent features of multimedia contents and define the syntax and the semantics of each feature representation.
Description Schemes (DSs): specify the structure and semantics of the relationships between their components. These components may be both Descriptors and Description Schemes.
Description Definition Language (DDL): allows the creation of new Description Schemes and, possibly, Descriptors. It also allows the extension and modification of existing Description Schemes.
System tools: support multiplexing of descriptions, synchronization of descriptors with content, transmission mechanisms, coded representations (both textual and binary formats) for efficient storage and transmission, management and protection of intellectual property in MPEG-7 descriptions, etc.

Fig. 2. Main element of MPEG-7's standard

of an analogue movie or of a picture printed on paper in the same approach as contents in digital formats. The main elements of MPEG-7's standard are shown in Fig. 2 [10].

MPEG-7 standard comprises several parts under the general title Information Technology – Multimedia Content Description Interface:

1) *MPEG-7 System*,
2) *MPEG-7 Description Definition Language (DDL)*,
3) *MPEG-7 Visual (MPEG-7 Visual) description tools*—the Ds and DSs dealing with only Visual descriptions,
4) *MPEG-7 Audio (MPEG-7 Audio) description tools*—the Ds and DSs dealing with only Audio descriptions,
5) *MPEG-7 Multimedia Description Schemes (MPEG-7 MDS)*—the Ds and DSs dealing with generic features and multimedia descriptions,
6) *MPEG-7 Reference Software*—a software implementation of relevant parts of the MPEG-7 standard and
7) *MPEG-7 Conformance*—guidelines and procedures for testing conformance of MPEG-7 implementations.

MPEG-7 has decided to adopt the XML Schema as MPEG-7 DDL language. However, because XML Schema language has not been designed specifically for audiovisual content, certain extensions are needed in order to satisfy all MPEG-7 DDL requirements. Hence, the DDL consists of the components [10]:

1) XML Schema Structural components,
2) XML Schema Datatype components, and
3) MPEG-7 Extensions to XML Schema.

The DDL language forms a core part of the MPEG-7 standard and defines the syntactic rules to express, combine, extend and refine Ds and DSs. It also provides the basic Ds and DSs, which users can extend to their own application domain. The next section will introduce the important description tools into the MPEG-7 standard.

3.2 Semantic Description Tools in MPEG-7

The most important tools in MPEG-7 are semantic description tools. They provide standard tools to describe the semantics of multimedia contents in MPEG-7 MDS. Those tools embrace Abstraction Model, Semantic Entities, Semantic Attributes, and Semantic Relations. The details of all DSs and their functionalities in an abstraction model through semantic relation description tools can be found in [1], and additional, readily understood information in [2]. Although MPEG-7 allows textual description in semantic description tools, their real power depends on the ability to describe semantic entities and their relationships in a manner which supports inference and reasoning. However, all Ds and DSs are described by XMLS which lacks the capability to represent implicit information as well as a reasoning mechanism.

4 Schema Language for MPEG-7

As has been mentioned in Section 3, there seems to be a need for changing the representation of Ds and DSs into a new data model with a reasoning ability as well as an inferential mechanism. RDF and RDFS may be a suitable candidate of such a new representation. However, they lack an inferential mechanism, extensive capability to represent implicit information as well as a flexible method for stating constraints on resources. For instance, let there be given a picture of fruits in Fig. 3, and descriptions of *leftOf* relation such as *grapes leftOf apple*, *apple leftOf strawberry*, and *strawberry leftOf banana*. What will happen if users query "list all fruits that are on the right hand side of grapes"? Based on the above descriptions, the system does not return an answer, because there are no descriptions or keywords corresponding to such users' queries. However, the *rightOf* relation, the inverse relation of *leftOf*, can be derived from *leftOf* relation, even though it does not express it explicitly. Moreover, RDF and RDFS lack a mechanism to express comparison constraints such as *greater than (>)*, *less than (<)*, *greater than or equal (≥)* and *less than or equal (≤)*, or mathematic operations, such as *add (+)*, *subtract (-)*, *divide (÷)*, and *multiply (×)*.

Hence a new schema language for MPEG-7 is proposed which employs XDD theory—the approach used to model each part of semantic description tools—and gives examples, which illustrate how one can apply a new schema language with semantic descriptions of multimedia contents.

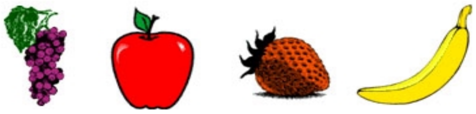


Fig. 3. A picture of fruits

4.1 Modeling MPEG-7 Schema with XDD Descriptions

Due to the inabilities of XMLS to express implicit information of multimedia content and the need to increase the expressive power of MPEG-7's Ds and DSs, XDD theory is employed as a new schema language for MPEG-7. In order to maintain interoperability and increase the expressive power of Ds and DSs, the proposed framework will reuse the terminology from existing ontology languages such as RDF, RDFS and DAML+OIL as well as model their constraints and ontological axioms. Fig. 4 presents an overview of the framework.

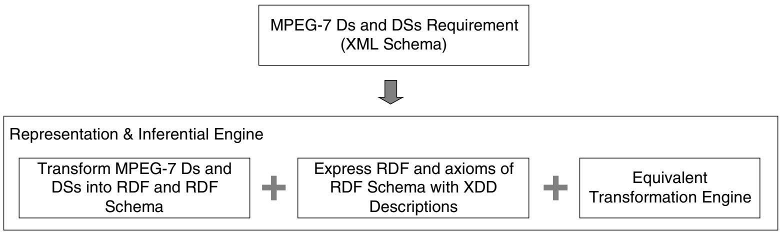

MPEG-7 Ds and DSs Requirement (XML Schema)
Representation & Inferential Engine
Transform MPEG-7 Ds and DSs into RDF and RDF Schema
+
Express RDF and axioms of RDF Schema with XDD Descriptions
+
Equivalent Transformation Engine

Fig. 4. Overview of framework

For the sake of simplicity, the example will omit certain details of Ds and DSs as well as employ RDF(S) syntax as their descriptions; however, DAML+OIL ontology can be employed instead, but all information will be expressed in terms of XDD descriptions. Table 1 summarizes the modeling MPEG-7 elements in XDD descriptions. Details of the modeling are described in the following 3 sub-sections under the headings: domain ontology, ontological axioms and application rules modeling.

Table 1. Modeling MPEG-7 Semantic elements

MPEG-7 Semantic Elements	Expressed as	Description
Domain Ontology Abstraction Model Semantic Entities Semantic Attributes Semantic Relations	XDD descriptions using RDF(S) or DAML+OIL (XML unit and non-unit clauses)	Modeling concepts, properties, and their relations of terms used in application domain.
Ontological axioms	XDD descriptions (XML non-unit clauses)	Modeling the axioms of terms used in application domain.
Application rules	XDD descriptions (XML non-unit clauses)	Modeling rules, axioms, constraints and queries in application domain.

4.2 Modeling Domain Ontology

All semantic Ds and DSs of semantic entities are represented by XMLS. There is a need to transform them into RDFS which has more expressiveness. The transformation translates child elements of the XMLS *complexType* into property elements attached to the RDFS class. Fig. 5 illustrates an example of a SemanticPlace DS expressed as an RDFS graph. The corresponding descriptions in RDF syntax are shown in Fig. 6.

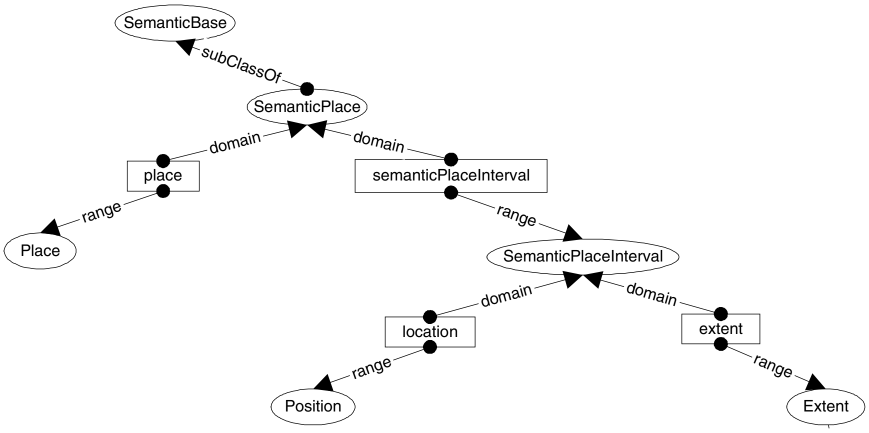


Fig. 5. SemanticPlace DS expressed in RDFS graph

```
<rdfs:Class rdf:ID="SemanticPlace">
        <rdfs:label>SemanticPlace DS</rdfs:label>
        <rdfs:subClassOf rdf:resource=”#SemanticBase”/>
</rdfs:Class>

<rdf:Property rdf:ID="place">
        <rdfs:domain rdf:resource="#SemanticPlace"/>
        <rdfs:range rdf:resource="#Place"/>
</rdf:Property>

<rdf:Property rdf:ID="semanticPlaceInterval">
        <rdfs:domain rdf:resource="#SemanticPlace"/>
        <rdfs:range rdf:resource="#SemanticPlaceInterval"/>
</rdf:Property>

<rdfs:Class rdf:ID="SemanticPlaceInterval">
        <rdfs:label>Class SemanticPlaceInterval</rdfs:label>
</rdfs:Class>

<rdf:Property rdf:ID="location">
        <rdfs:domain rdf:resource="#SemanticPlaceInterval"/>
        <rdfs:range rdf:resource="#Position"/>
</rdf:Property>

<rdf:Property rdf:ID="extent">
        <rdfs:domain rdf:resource="#SemanticPlaceInterval"/>
        <rdfs:range rdf:resource="#Extent"/>
</rdf:Property>
```

Fig. 6. RDFS descriptions of SemanticPlace DS

In order to preserve the difference of XMLS *attributes* and *child elements* in RDFS property elements, the postfix *attr* is attached to property names, such as attribute *datum* in a GeographicPosition DS will become a *datum_attr* property of class GeographicPosition. The rest of the Ds and DSs in the semantic elements are processed similarly.

Semantic relation elements are distinct from other elements. They have the same structures of DSs and differ only in their names. The importance of a semantic relation is its ability to describe the complex meaning of multimedia content. The power of semantic relations depends on the ability to capture meaning or express implicit information of multimedia contents. For this purpose, they need a schema which provides a flexible method for expressing rules, constraints and axioms as well as supports an inferential mechanism. Specification [1] has four relation categories:

- *Semantic Entity Relation Tools*: describe relations among semantic entities,
- *Semantic Relation – Semantic Entity Relation Tools*: describe relations among semantic relation and semantic states,
- *Segment – Semantic Entity Relation Tools*: describe relations among segments and semantic entities,
- *Analytic Model – Semantic Entity Relation Tools*: describe relations among analytic models and semantic entities.

Each category has predefined relation names with functionalities. Moreover, almost all relations have inverse relations, i.e., they increase the expressive power of semantic relation tools. However, they must be expressed explicitly in the descriptions. The disadvantages of express semantic relations and their inverses explicitly are redundancy of information and, if they express at different locations, they are not easily noticed; hence an inconsistency will occur when one of them is removed. Instead of expressing them explicitly, one could add axioms of their inversion. By employing XDD descriptions, not only axioms of a relation's inverse property can be expressed, but also other axioms such as a transitive property which makes relation description more powerful.

4.3 Modeling Ontological Axioms

RDFS uses two important constructs—*subClassOf* and *subPropertyOf*—to specify hierarchical relationship among sets of classes and properties, respectively. It also provides *domain* and *range* constructs to specify constraints on a property's value and on types of objects to which a property can be applied. The meanings of *subClassOf* and *subPropertyOf* are transitive and include some notions of implication. Since there are many predefined relation names in a specification, the following example (Fig. 7) illustrates the modeling axioms of the relation name *leftOf* as well as its inversion axioms.

The meaning of the remaining relations and their inversions can be described by XML non-unit clauses in a similar manner.

```
CX1:  <ObjectObjectRelation rdf:about=$S:rel_AC>                                 %  Modeling the transitive axioms of
          <name>leftOf</name> <source rdf:resource=$S:resourceA/>              %  leftOf relation: if resource A lies on
          <target rdf:resource=$S:resourceC/>                                  %  the left hand side of resource B and
      </ObjectObjectRelation>                                                  %  resource B also lies on the left hand
      ←   <ObjectObjectRelation rdf:about=$S:rel_AB>                           %  side of resource C, then resource A
              <name>leftOf</name> <source rdf:resource=$S:resourceA/>          %  lies on left hand side of resource C.
              <target rdf:resource=$S:resourceB/>                              %
          </ObjectObjectRelation>,
          <ObjectObjectRelation rdf:about=$S:rel_BC>
              <name>leftOf</name> <source rdf:resource=$S:resourceB/>
              <target rdf:resource=$S:resourceC/>
          </ObjectObjectRelation>,
          CONCAT($S:rel_AB, $S:rel_BC, $S:rel_AC).
```

```
CX2:  <ObjectObjectRelation rdf:about=$S:rel_BA>                                 %  Modeling the inverse axioms of
          <name>rightOf</name>      <source rdf:resource=$S:resourceA/>        %  leftOf relation, rightOf : if resource
          <target rdf:resource=$S:resourceB/>                                  %  B lies on left hand side of resource
      </ObjectObjectRelation>                                                  %  A, then resource A lies on right hand
      ←   <ObjectObjectRelation rdf:about=$S:rel_BA>                           %  side of resource B.
              <name>leftOf</name> <source rdf:resource=$S:resourceB/>
              <target rdf:resource=$S:resourceA/>
          </ObjectObjectRelation>.
```

Fig. 7. XML non-unit clauses express axioms and inversion axioms of *leftOf* relation

4.4 Modeling Application Rules

Suppose there is a video segment which contains some scenes involving animals. How can the user produce a new video segment which only contains animal scenes. In this example, the video segment VS1 represents a video sequence comprising five scenes. Video segment VS1 is decomposed into five video segments, VS2, VS3, VS4, VS5 and VS6 as depicted by Fig. 8 and this decomposition has neither gaps nor overlaps.

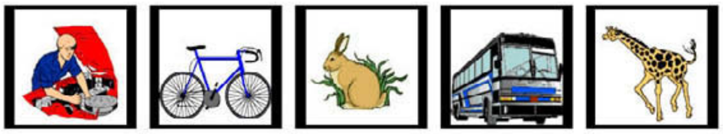


Fig. 8. Video segment named "travel.mpg"

The descriptions of Fig. 9 give merely physical properties of each video segment such as its parent, the beginning point of the segment counted from the starting point of its parent and its length (duration). The other descriptions are required by the semantics of each segment. For the sake of simplicity, each segment contains only one semantic. Fig. 10 shows semantic description of each segment.

C_{G1}:	<VideoSegment rdf:about="VS1"> <TemporalSegmentLocator rdf:about="TSL1"> <mediaURI rdf:resource="http://travel.mpg"/> <MediaTime rdf:about="MT1"> <mediaTimePoint>T00:00:00</mediaTimePoint> <mediaDuration>PT25M</mediaDuration> </MediaTime> </TemporalSegmentLocator> <TemporalDecomposition rdf:about="TD1"> <gap>false</gap> <overlap>false</overlap> </TemporalDecomposition> </VideoSegment> ←.	% % % % % %	Clause C_{G1} expresses information of the video segment named "travel.mpg". It is 25 minutes long, and has 5 video segments, which do not overlap, and have no gaps in between.
C_{G2}:	<VideoSegment rdf:about="VS2"> <TemporalSegmentLocator rdf:about="TSL2"> <mediaURI rdf:resource="http://travel.mpg"/> <MediaTime rdf:about="MT2"> <mediaTimePoint>T00:00:00</mediaTimePoint> <mediaDuration>PT5M</mediaDuration> </MediaTime> </TemporalSegmentLocator> </VideoSegment> ←.	% % % %	Clause C_{G2} gives information about the video sub-segment VS2. It starts at position 0 minute and lasts 5 minutes.
C_{G3}:	<VideoSegment rdf:about="VS3"> <TemporalSegmentLocator rdf:about="TSL3"> <mediaURI rdf:resource="http://travel.mpg"/> <MediaTime rdf:about="MT3"> <mediaTimePoint>T00:05:00</mediaTimePoint> <mediaDuration>PT5M</mediaDuration> </MediaTime> </TemporalSegmentLocator> </VideoSegment> ←.	% % % %	Clause C_{G3} gives information about the video sub-segment VS3. It starts at position 5 minutes from the beginning and lasts 5 minutes.
C_{G4}:	<VideoSegment rdf:about="VS4"> <TemporalSegmentLocator rdf:about="TSL4"> <mediaURI rdf:resource="http://travel.mpg"/> <MediaTime rdf:about="MT4"> <mediaTimePoint>T00:10:00</mediaTimePoint> <mediaDuration>PT5M</mediaDuration> </MediaTime> </TemporalSegmentLocator> </VideoSegment> ←.	% % % %	Clause C_{G4} gives information about the video sub-segment VS4. It starts at position 10 minutes from the beginning and lasts 5 minutes.
C_{G5}:	<VideoSegment rdf:about="VS5"> <TemporalSegmentLocator rdf:about="TSL5"> <mediaURI rdf:resource="http://travel.mpg"/> <MediaTime rdf:about="MT5"> <mediaTimePoint>T00:15:00</mediaTimePoint> <mediaDuration>PT5M</mediaDuration> </MediaTime> </TemporalSegmentLocator> </VideoSegment> ←.	% % % %	Clause C_{G5} gives information about the video sub-segment VS5. It starts at position 15 minutes from the beginning and lasts 5 minutes.
C_{G6}:	<VideoSegment rdf:about="VS6"> <TemporalSegmentLocator rdf:about="TSL6"> <mediaURI rdf:resource="http://travel.mpg"/> <MediaTime rdf:about="MT6"> <mediaTimePoint>T00:20:00</mediaTimePoint> <mediaDuration>PT5M</mediaDuration> </MediaTime> </TemporalSegmentLocator> </VideoSegment> ←.	% % % %	Clause C_{G6} gives information about the video sub-segment VS6. It starts at position 20 minutes from the beginning and lasts 5 minutes.

Fig. 9. Description of individual video segment

However, the descriptions in Fig. 10 do not express which segment concerns Animal, whence additional knowledge about animals is needed—clauses C_{G12} – C_{G14} in Fig. 11. In real applications, not only facts can be entered into the system, but one can define constraints which derive new information from facts.

```
CG7:   <SegmentSemanticBaseRelation rdf:about="SSBR1">                          %   Clause CG7 gives semantics of the
           <name>mediaPerceptionOf</name> <source rdf:resource="#VS2"/>        %   video segment VS2, which is a
           <target rdf:resource="#Man"/>                                       %   scene of man.
       </SegmentSemanticBaseRelation>   ←.
------------------------------------------------------------------------------------------------------------
CG8    <SegmentSemanticBaseRelation rdf:about="SSBR2">                          %   Clause CG8 gives semantics of the
           <name>mediaPerceptionOf</name> <source rdf:resource="#VS3"/>        %   video segment VS3, which is a
           <target rdf:resource="#Bicycle"/>                                   %   scene of bicycle.
       </SegmentSemanticBaseRelation>   ←.
------------------------------------------------------------------------------------------------------------
CG9:   <SegmentSemanticBaseRelation rdf:about="SSBR3">                          %   Clause CG9 gives semantics of the
           <name>mediaPerceptionOf</name> <source rdf:resource="#VS4"/>        %   video segment VS4, which is a
           <target rdf:resource="#Rabbit"/>                                    %   scene of rabbit.
       </SegmentSemanticBaseRelation>   ←.
------------------------------------------------------------------------------------------------------------
CG10:  <SegmentSemanticBaseRelation rdf:about="SSBR4">                          %   Clause CG10 gives semantics of the
           <name>mediaPerceptionOf</name> <source rdf:resource="#VS5"/>        %   video segment VS5, which is a
       <target rdf:resource="#Bus"/>                                           %   scene of bus.
       </SegmentSemanticBaseRelation>   ←.
------------------------------------------------------------------------------------------------------------
CG11:  <SegmentSemanticBaseRelation rdf:about="SSBR5">                          %   Clause CG11 gives semantics of the
           <name>mediaPerceptionOf</name> <source rdf:resource="#VS6"/>        %   video segment VS6, which is a
           <target rdf:resource="#Giraffe"/>                                   %   scene of giraffe.
       </SegmentSemanticBaseRelation>   ←.
```

Fig. 10. Semantic of video segment VS2, VS3, VS4, VS5 and VS6

```
CG12:  <rdfs:Class rdf:ID="Animal">                                               %   Clauses CG12 –CG14 give infor-
           <rdfs:label>Animal Kingdom</rdfs:label>                                %   mation about animal.
       </rdfs:Class>   ←.

CG13:  <rdfs:Class rdf:ID="Rabbit">
           <rdfs:subClassOf rdf:resource="#Animal"/>
       </rdfs:Class>   ←.

CG14:  <rdfs:Class rdf:ID="Giraffe">
           <rdfs:subClassOf rdf:resource="#Animal"/>
       </rdfs:Class>   ←.
------------------------------------------------------------------------------------------------------------
CR:    <BinaryTemporalSegmentRelation rdf:about=$S:segment_A>                     %   Clause CR defines constraints
           <name>before</name>     <source rdf:resource=$S:segment_A/>            %   of relation before, which will use
           <target rdf:resource=$S:segment_B/>                                    %   to combine each segment of
       </BinaryTemporalSegmentRelation>                                           %   video together.
       ←   <NonSeqSegment rdf:about=$S:segment_A>                                 %   Constraint ADD computes the
               <TemporalSegmentLocator rdf:about=$S:A_Loc>                        %   end point of each segment and
                   <mediaURI rdf:resource=$S:same_segment/>                       %   use constraint LT to check
                   <MediaTime rdf:about=$S:A_MT>                                  %   which segment comes before
                       <mediaTimePoint>$S:A_TP</mediaTimePoint>                   %   the other.
                       <mediaDuration>$S:A_MD</mediaDuration>
                   </MediaTime>
               </TemporalSegmentLocator>
           </NonSeqSegment>,
           <NonSeqSegment rdf:about=$S:segment_B>
               <TemporalSegmentLocator rdf:about=$S:B_Loc>
                   <mediaURI rdf:resource=$S:same_segment/>
                   <MediaTime rdf:about=$S:B_MT>
                       <mediaTimePoint>$S:B_TP</mediaTimePoint>
                       <mediaDuration>$S:B_MD</mediaDuration>
                   </MediaTime>
               </TemporalSegmentLocator>
           </NonSeqSegment>,
           ADD($S:A_TP, $S:A_MD, $S:A_Result),
           LT($S:A_Result, $S:B_TP).
------------------------------------------------------------------------------------------------------------
CQ:    <NonSeqSegment rdf:about=$S:vs_X>                                          %   Clause CQ selects video seg-
           $E:VS_Property                                                         %   ment which contains scenes of
       </NonSeqSegment>                                                           %   animals.
       ←   <VideoSegment rdf:about=$S:vs_X>
               $E:VS_Property
           </VideoSegment>,
           <SegmentSemanticBaseRelation rdf:about=$S:ssbr>
               <name>mediaPerceptionOf</name>
               <source rdf:resource=$S:vs_X/><target rdf:resource="#Animal"/>
           </SegmentSemanticBaseRelation>.
```

Fig. 11. Rules and constraints of application are expressed in XDD

Query clause C_Q of Fig. 11 merely produces a number of single video segments (frames), and dose not combine them. In order to combine video segment, constraint C_R is defined as shown in Fig. 11.

The constraint C_R composes new multimedia information by defining the *before* relation between two sub-segments which come from the same video segment—restricted by *$S:same_segment*. It compares the end time of video segment *A*—computed by the constraint *ADD(v1, v2, result)* which will return true when *v1*, *v2*, and *result* are numeric data and *result* = *v1* + *v2*—to the start time of video segment *B*. If the end time of segment *A* is less than the start time of segment *B*, then segment *A* comes before segment *B*. Based on the information of clauses C_{G1} - C_{G14} and constraint C_R, clause C_Q can compose new multimedia information.

5 Conclusion

By employing XDD description as a schema language for MPEG-7, the expressive power of Ds and DSs to represent implicit information of multimedia content can be enhanced. It has a flexible method to express rules, axioms, and constraints which strengthen annotation description. Moreover, it reduces explicit annotation of semantic relations and their inversion, makes the descriptions consistent and modifications easy. The reasoning ability enhances the capability of descriptions to capture the complex semantics of multimedia content. Moreover, it also maintains interoperability among existing ontologies and enables their exchange of information.

MPEG-7 also contains normative structural relations such as *meets*, *overlaps*, *starts* and *finishes* which are useful when dealing with multimedia synchronization and spatial information. XDD could be used to represent and reason with these complex relations, thus yielding a schema language for efficient representation, retrieval and composition of temporal and spatial multimedia information.

References

1. Beek, P. V., Benitez, A. B., Heuer, J., Martinez, J., Salembier, P., Shibata, Y., Smith, J. R., and Walker, T.: Text of 15938-5 FCD Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes, Singapore, March 2001.
2. Benitez, A. B., Rising, H., Jorgensen, C., Leonardi, R., Bugatti, A., Hasida, K., Mehrotra, R., Tekalp, A. M., Ekin, A., and Walker, T.: Semantics of Multimedia in MPEG-7, Proceeding of IEEE 2002 Conference on Image Processing (ICIP-2002), Rochester, New York, USA, September 2002.
3. Biron, P. V., and Malhotra, A.: XML Schema Part 2: Datatypes, W3C Recommendation, May 2001, [http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/].
4. Bray, T., Paoli, J., Sperberg-McQueen, C. M., and Maler, E.: Extensible Markup Language (XML) 1.0 (Second Edition), W3C Recommendation, April 2002, [http://www.w3.org/TR/2000/REC-xml-20001006].
5. Brickley, D. and Guha, R. V.: RDF Vocabulary Description Language 1.0: RDF Schema, W3C Working Draft, April 2002, [http://www.w3.org/TR/2002/WD-rdf-schema-20020430/].
6. Fallside, D. C.: XML Schema Part 0: Primer, W3C Recommendation, May 2001, [http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/].

7. Harmelen, F. V., Patel-Schneider, P. F., and Horrocks, I.: Reference description of the DAML+OIL (March 2001) ontology markup language, March 2001, [http://www.daml.org/2001/03/reference].
8. Hunter, J.: Adding Multimedia to the Semantic Web – Building an MPEG-7 Ontology, [http://www.semanticweb.org/SWWS/program/full/paper59.pdf].
9. Lassila, O., and Swick, R. R.: Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation, February 1999, [http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/].
10. Martinez, J. M.: Overview of the MPEG-7 Standard (version 5.0), Singapore, March 2001.
11. Thompson, H. S., Beech, D., Maloney, M., and Mendelsohn, N.: XML Schema Part 1: Structures, W3C Recommendation, May 2001, [http://www.w3.org/TR/2001/REC-xmlschema-1-20010502/].
12. Wuwongse, V., Anutariya, C., Akama, K., and Nantajeewarawat, E.: XML Declarative Description: A language for the Semantic Web, IEEE Intelligent Systems, Vol. 16, No. 3, May/June 2001, pp. 54-65.

Bitmap-Based Indexing for Multi-dimensional Multimedia XML Documents

Jong P. Yoon[1], Sung H. Kim[2], Go U. Kim[2], and Venu Chakilam[1]

[1]Center for Advanced Computer Studies
University of Louisiana, Lafayette, LA 70504-4330
{jyoon,vmc0583}@cacs.Louisiana.edu
[2]Division of Information Science
Sookmyung Women's University, Seoul, Korea 710
{ksh,gounkim}@sookmyung.ac.kr

Abstract. XML is a new standard for exchanging and representing information on the Internet. Documents can be hierarchically represented in XML-elements and also available for sophisticated content-based retrieval. For fast retrieval, XML documents may be indexed. Typical indexing techniques, however, are not satisfactory for multi-dimensional and irregularly hierarchical XML documents. In this paper, we propose a scalable bitmap indexing that can index not only document-path-content (or –word) information but also additional information such as the occurrence and reference/de-reference information of words and paths, or multimedia features in digital libraries. Querying XML document collections can be performed based on combinations of primitive operations such as slice, project, and dice. Bit-wise operations are outperformed in bitmap indexes. We also define the notion of distances in bitmap indexes suitable for sophisticated or proximity approximation retrievals. Experiments show that the bitmap-based indexing for multiple features of XML documents can be constructed efficiently, and the distance operations can be performed more efficiently with the BitCube than with other alternatives.

1 Introduction

The eXtensible Markup Language (XML) is a standard for representing and exchanging information on the Internet, e.g., .NET™ XML Web Servers, and UDDI. As such, documents in digital libraries can be represented in XML-elements, and therefore content-based retrieval is possible. However, because the size of XML documents is very large and the features associated with those documents are multi-dimensional, typical database indexing techniques are not satisfactory.

We consider an XML document database (D). Each document (d) is represented in XML. So, d contains XML-elements (p), where p has zero or more terms (w) bound to it. Typical indexing requires a frequency table that is a two-dimensional matrix indicating the number of occurrence of the terms used in documents. We developed a three-dimensional matrix of (d, p, w) [15]. By extending this indexing technique towards the multi-dimensional information of XML documents, we propose a scalable BitCube indexing that can include not only the three-dimensions (d, p, w) but also

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 165-176, 2002.

additional dimensions, *f, s, e, r, c*, etc., where *f* denotes frequency or occurrences of *w*, *s* denotes sensitivity of *p*, *e* denotes encryption of *p*, *r* denotes references of *d*, *c* denotes de-references of *d*, etc.

We also consider a query that may be in any combination (subset) of multi-dimensions (*d*, *p*, *w*, *f, s, e, r, c*). Proximity queries can be computed by distance functions. Distance functions for multi-dimensional XML documents will be developed in this paper. In many cases on the Internet, typical evaluation of proximity queries is often too slow and costly. A simple way to speed up query answering is to construct indexes, but approximation on multi-dimensions is again unsatisfactory. To solve this difficulty, in this paper, we propose a scalable bitmap indexing technique, and approximation operations that can retrieve such documents efficiently. Before going further, consider the following examples that illustrate why a sophisticated indexing technique is needed.

1.1 Motivating Examples

Example 1. Consider a query Q1 that is posed to find all documents that contain the word "XML" three times in "any" figure caption(s) of "any-level" subsections. This type of queries is processed slowly because they are posed to a large number of XML document collections. To speed up the process, one may consider an indexing technique. Typical indexing techniques (e.g., B+-tree or hash indexing) are unsatisfactory partly because XML documents are hierarchical and irregular in structure and partly because XML documents are multi-dimensional, i.e., multiple features in (multimedia) documents in digital libraries.

Example 2. Consider a query Q2 that is requested for those documents that contain similar element contents to document *d1*. Suppose that *d1* contains *w1* and *w2* in an element *e1*, and a document *d2* has *w1* in an element *e2* that in turn contains *w2* in sub-elements (e.g., the first two documents in Figure 1). They are approximately the same because those two document d1 and d2 are "similar" in structure, but they are "exactly the same" in content. Serving this type of approximation queries is not satisfactorily possible in typical indexes because XML elements are again hierarchically irregular in structure.

1.2 Related Work

Digital libraries contain multimedia documents. Such multimedia documents have been represented in XML [5]. Recently, MPEG7 [7] formally named "Multimedia Content Description Interface", as a multimedia content description standard that describes multimedia contents, so that users can search, browse, and retrieve those contents more efficiently and effectively than they could using today's leading text-based search engines.

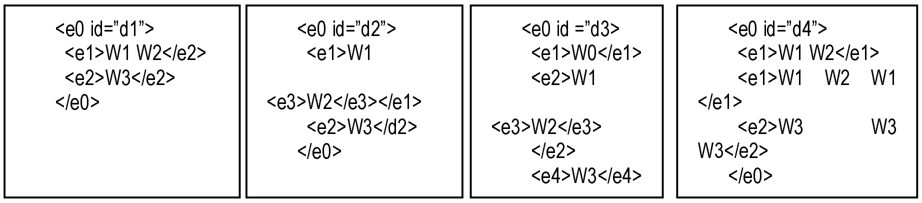
Typical indexing techniques in the database community are B+-tree indexing and hash-based indexing [8]. They have been employed in XML databases. For example, B+-tree has been employed in XQEngine (www.fatdog.com), hash-based indexing has been employed in XYZfinder (www.xyzfinder.com), and bitmap-based indexing has been employed in BitCube [15]. In principle, an index is constructed on one XML-element (or one attribute) or a combination of more than one XML-element. Recently some researchers invented new indexing techniques for XML documents: Xmills [6] and X-tree [1]. A new data structure, called X-tree, has been introduced for storing very high dimensional data [2]. XML documents are indexed using database techniques [4,9,12].

Full-text documents are in the form of (*d, w*), where *d* denotes document and *w* denotes word. They are therefore regarded as a sequence of words without situating them in appropriate attributes, and so whole keywords of the full-text documents can be indexed in a B+-tree or hash-based index. However, notice that XML documents are in the form of (*d, p, w*), where *d* denotes document, *p* denotes path of XML-elements, and *w* denotes word. They have contents that are encompassed hierarchically in multiple-level XML-elements. Typical database indexing techniques are not satisfactory for the case that *w* is encompassed in multiple-level *p* in document *d*. Such indexes are costly and result in poor performance in space and in time.

The collection of bitmaps in a bitmap index forms a 2-dimensional [3] and 3-dimensional bitmap index [15,5]. A bitmap index has been used to optimize queries [3,10,14]. Bit-wise operations developed in the earlier work were also generalized to the 3-dimensional bit matrix context [15]. In this paper, we will extend BitCube to a multi-dimensional bit matrix, which can manipulate multiple dimensions (*d*, *p*, *w*, *f*, *s*, *t*, *e*, *l*, *r*) of XML document collections. Bit-wise operations developed in the earlier work will also be generalized to the 3-dimensional bit matrix context.

2 Preliminaries

An XML document is defined as a sequence of ePaths and contents associated with them. An XML document database contains a set of XML documents for simplicity purpose. We use a bitmap index for an XML document database. Consider a 2-dimensional bitmap index. In a document-ePath bitmap index, a bit column represents an ePath, and a row represents an XML document. Of course, element contents, that is, values or words, need to be taken into account. In doing so, we need to consider multi-dimensional bitmap index, which will be discussed in detail in Section 4. As an example of a bitmap index, assume those XML documents in Figure 1.



```
<e0 id="d1">
  <e1>W1 W2</e2>
  <e2>W3</e2>
</e0>
```

```
    <e0 id="d2">
      <e1>W1

<e3>W2</e3></e1>
      <e2>W3</d2>
    </e0>
```

```
    <e0 id ="d3>
      <e1>W0</e1>
      <e2>W1

<e3>W2</e3>
      </e2>
      <e4>W3</e4>
```

```
    <e0 id="d4">
      <e1>W1 W2</e1>
      <e1>W1    W2    W1
</e1>
      <e2>W3            W3
W3</e2>
    </e0>
```

Fig. 1. Example of XML Documents

Figure 1 is a set of simple XML documents. First, we need to define ePaths as follows:

```
p0=e0.e1,  p1=e0.e2,  p2=e0.e1.e3,  p3=e0.e2.e3,
p4=e0.e4.
```

Notice that Wi is a (key) word that is chosen from simple content to be used for search.

The construction of a bitmap index is as follows: If a document has ePath, then the corresponding bit for the ePath is set to 1, and otherwise, set to 0. For each ePath, documents can be represented as shown in Figure 2.

Definition 2.1. (*Size of Bitmap*) $|b_i|$ denotes the size of a bitmap b_i, which is the number of 1's in a bitmap b_i, and $||b_i||$ denotes the cardinality of a bitmap b_i, which is the number of 1's or 0's.

	p0	p1	p2	p3	p4
d1	1	1	0	0	0
d2	1	1	1	0	0
d3	1	1	0	1	1
d4	1	1	0	0	0

Fig. 2. A Bitmap Index for Figure 1

Before extending the notion of similarities, we first consider the indexing technique which can represents multi-dimensional XML documents.

3 Three-Dimensional Indexing

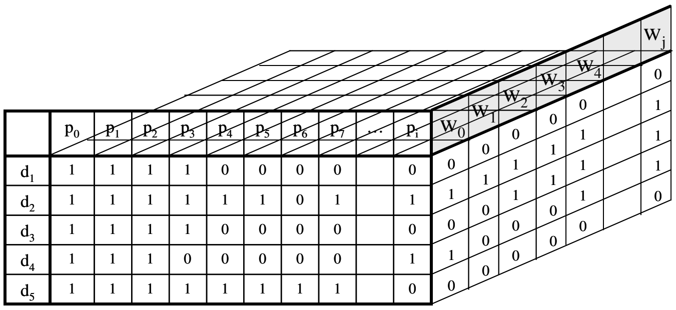
In this section, we review a 3-dimensional bitmap index, called "BitCube," and apply to digital libraries. A BitCube represents a set of documents together with both (1) a set of ePaths (or XML-elements) and (2) a set of words (or element contents) for each ePath. For a BitCube, we propose two types of slice (slice of ePath and slice of Content) and project (of documents).

3.1 BitCube

A BitCube for XML documents is defined as BitCube = (*d, p, w, b*), where *d* denotes XML document, *p* denotes ePath, *w* denotes word or content for ePath, and *b* denotes 0 or 1, the value for a bit in BitCube (if ePath contains a word, the bit is set to 1, and 0 otherwise).

For example, consider a collection of XML documents: D={d_1, d_2, d_3, d_4, d_5}. Each documents d_1={(p_0, w_1), (p_1, w_2), (p_1, w_3), (p_1, w_5), (p_2, w_3), (p_2, w_8) }, .., d_3={(p_0,w_{11}), (p_1, w_2), (p_1, w_7), (p_2, w_3), (p_2, w_9) …, (p_i, w_{i2}), (p_i, w_{i3}), (p_i, w_{i4}), …, (p_i,

w_{ij})}, and so on. Notice that *P* denotes all ePaths and *W* denotes all (key) words used in *D*. The BitCube for this collection looks like in Figure 3.



	p_0	p_1	p_2	p_3	p_4	p_5	p_6	p_7	…	p_i
d_1	1	1	1	1	0	0	0	0		0
d_2	1	1	1	1	1	1	0	1		1
d_3	1	1	1	1	0	0	0	0		0
d_4	1	1	1	0	0	0	0	0		1
d_5	1	1	1	1	1	1	1	1		0

w_0	w_1	w_2	w_3	w_4		w_j
0	0	0	0	0		0
1	1	1	1	1		1
0	0	1	1	1		1
1	0	0	0	1		1
0	0	0	0	0		0

Fig. 3. A BitCube: Example

3.2 BitCube Operations

The basic three operations are described in this section: (1) ePath-Slice, (2) word-Slice, and (3) document-Project. The outcome of these operations, if applied against a BitCube, is a 2-dimensional bitmap index. In addition to these query operations, this section describes some operations that are needed to manipulate a BitCube: (4) unfolding, and (5) focusing-Element.

ePath_Slice. Each bit for a particular ePath is sliced. This operation takes a Path as input and returns a set of documents with words associated with it.

```
P_Slice(ePath) = {(doc, word) | ePath is used in doc,
and word is associated with the ePath}.
```

The outcome of this slicing is a 2-dimensional bitmap index that represents a set of documents with a set of words. Typical web searches may not possible for ePath.

Word-Slice. Each bit for a particular word can be sliced. This operation takes a (search key) word as input and returns a set of documents.

```
W_Slice(word) = {(doc, ePath) | word is associated with
the ePath which is in turn used in doc}.
```

The outcome is a 2-dimensional bitmap index that represents a set of documents with a set of ePath with which the word is associated. Typical web searches are based on this word-Slice operation if they search XML documents.

Document Project. Each row of a BitCube can be projected. This operation takes a document as input and returns a set of ePaths with words associated with those ePaths.

Project (doc) = {(ePath, word) | entire content and ePath pairs appeared in doc}.

The outcome is a 2-dimensional bitmap index that represents a set of ePaths with their content (or words). A typical method for this project operation is a web browsing.

Dice: A Combination of Basic Operations. The operations "Slice" and "Project" can be applied multiple times. Multiple operations can be treated as an operation that specifies a range. If we mix those three operations, each of which specifies a range, the outcome is again a BitCube that is smaller or equal to the original BitCube.

Dice([d_s..d_e],[p_s..p_e],[w_s..w_e]) = {(doc, ePath, word) | doc∈[d_s..d_e], ePath∈[p_s..p_e], and word∈[w_s..w_e]}, where x∈[x_s..x_e] implies that $x_s \leq x \leq x_e$.

Unfolding. This operation is applied to a BitCube to flatten a hierarchical structure of XML documents. Given an element, the Unfolding operation removes all sub-elements.

Unfolding (*, p_i, *) = {(doc, ePath, word) | all words for all leaf elements of ePaths in docs but ePath subsumed by p_i}.

As an example of document d2 in Figure 1, the operation Unfolding (p1) returns <e1>W1<e3>W2</e3></e1> is unfolded into <e1>W1 W2</e1> which is the same as document d1.

4 Scalable Bitmap Indexing

It is likely that XML document collections in digital libraries contain not only ePath and word but also the information about reference and de-reference, and the information about their frequency, encryption, and security. Meaning that the XML documents are multi-dimensional. In this section, we extend BitCube to accommodate additional information available in XML document collections.

Consider XML documents in multi-dimensions (*d*, *p*, *w*, *f*, *s*, *e*, *r*, *c*). XML document can be indexed in BitCube only on the limited dimensions (*d*, *p*, *w*). A BitCube is not sufficient enough to represent multi-dimensional XML documents. For example, the documents d1 and d4 in Figure 1 are the same in BitCube in Figure 2. However, if we think the content of the documents, they are not exactly the same because the word W3 appears once in d1, but three times in d4. Topologically speaking, the structures of the two documents are not the same because the element <e1> appears one in d1 but twice in d4. Such differences motivate us to propose to include multiple aspects of XML information in our indexing technique.

In this section, we propose how the 3-dimensional information of BitCube can be converted to multi-dimensional structures, called "InfoCube." For one BitCube, there may be one or more InfoCubes, depending upon the number of dimensions. We first define bitmap mask vectors which can represent additional information *f*, *s*, *e*, *r, c* and show how each such bitmap mask vector can be applied to BitCube indexing and therefore generate an InfoCube.

InfoCube is 3-dimensional matrix, similar to BitCube, but it contains any values. Each value in InfoCube therefore implies the frequency of words, or the number of references, etc, which are additional to the ones available in a BitCube.

4.1 Bitmap Mask Vectors

A bitmap mask vector is a one-dimensional matrix. Each entry in the matrix is corresponding to a non-zero value in a BitCube. The number of occurrences (or frequency) of a dimension or topological information will be encoded into a bitmap mask vector. There are two types of bitmap mask vectors:

- *BitCube Mask Vector*. For each additional dimension, *f*, *s*, *e*, *r, c*, two types of bitmap mask vectors may be produced: one for word dimension, and another for ePath dimension. For example, for frequency dimension, two bitmap masks can be considered: the word frequency bitmap mask for Figure 2 is [1 1 1 1 1 1 1 1 1 1 3 2 3], while ePath frequency bitmap mask is [1 1 1 1 1 1 1 1 1 2 1]. For encryption, one may consider encryption on elements or encryption on contents only [13].
- *Tree Mask Vectors*. Topological information can be also encoded into a *tree mask vector*. For the same BitCube, there may be different topological tree structures of XML documents. A tree mask vector contains non-negative numbers that indicate the position where an ePath is branched. If the position is 0, then the ePath is attached to the root node of a tree. If the position is *n*, then ePath is attached from the node of that position. For example, consider the two ePath: p0 = e0.e1 and p1 = e0.e1.e2. Assume that <e0> is the root node. The ePath p1 may or may not be a subpart of p0. Therefore <e2> can be a sub-element of the existing element <e1> in one case. But, in another case, <e2> can be a new element from the root node <e0> thru the brand new element <e1>. The tree mask vector for the former case is [0 1], while the latter is [0 0].

4.2 Conversion Functions

A bitmap mask vector can be applied to a BitCube to convert to an InfoCube. This section describes two conversion functions available for XML document libraries.

- *InfoCube Generator*. Each value of a bitcube mask vector can be applied to a non-zero value of a BitCube to generate an InfoCube. Consider Figure 4 for example. Two mask vectors are applied to a BitCube. By applying a mask vector, an InfoCube is generated in (a). Notice that we use only 2-dimensional representation for simplicity purpose. This InfoCube contains the occurrence information about ePaths. Similarly, an InfoCube with de-reference information is generated in (b).

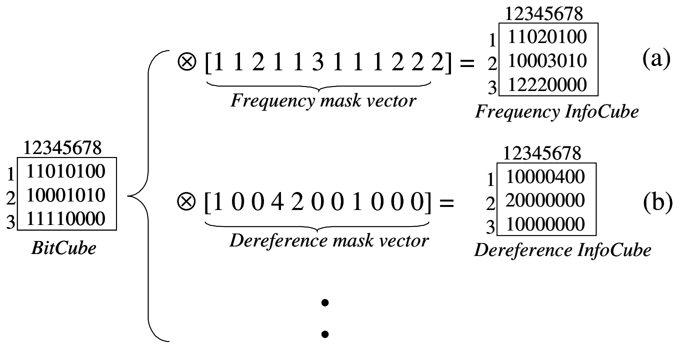


Fig. 4. Conversion of BitCube to InfoCube

- *Tree Generator*. Each value of a tree mask vector can be applied to a non-zero value of a BitCube or an InfoCube to generate tree structures. A tree for XML document schema is generated from BitCube, while trees of XML documents themselves are generated from InfoCube. For example, Figure 5 illustrates that a BitCube is converted into a schema tree by applying a tree mask vector [00222007], meaning that P_0 and P_1 are from the root, while P_2, P_3, and P_4 are from P_1 (the second bit). All rows in the BitCube are first ORed. Then the ORed matrix is applied with the tree vector to generate a tree. This tree represents schema information for all XML document instances. The very same tree mask vector is applied to the frequency InfoCube that was generated in Figure 4 (a). On the other hand, an InfoCube can also be converted to topological tree structures for XML documents.

4.3 InfoCube

There are numerous InfoCubes. XML documents in digital libraries are various in type, multi-dimensional in feature, irregular in structure. Consider XML documents (d, p, w, f, s, e, r, c). Then, there are InfoCubes available as follows: (1) InfoCube on content (or word) frequency, (2) InfoCube on ePath frequency, (3) InfoCube on content security, (4) InfoCube on content encryption, (5) InfoCube on ePath/content encryption, (6) InfoCube on content reference, and (7) InfoCube on content de-reference. Using these conversion functions, meaningful queries that cannot be served in typical indexing can be manipulated.

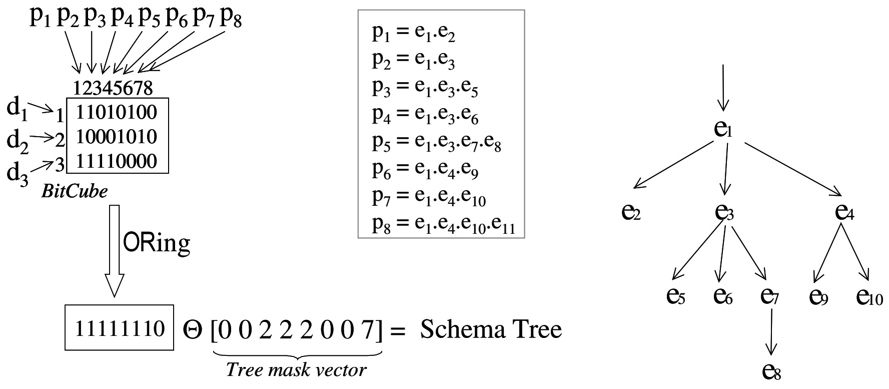


Fig. 5. Conversion of BitCube to a Schema Tree

5 Two Tiered Similarities for XML Documents

Traditional similarity matching techniques [2,11] are not satisfactory for proximity query processing partly because XML documents are organized irregularly and hierarchically, and partly because similarities on bitmap-based indexing are not fully developed yet.

As discussed in Section 4, although an XML document collection is indexed in 3-dimensional bitmap indexing, by applying various mask vectors we can represent and generate useful and meaningful InfoCubes. Therefore, similarity match can take place on not only BitCube but also InfoCubes. We propose two tiered similarity matches: Similarity on BitCube and Similarity on InfoCube. The BitCube mach uses the normalized Hamming distance to speed up the matching process, while the InfoCube match uses Euclidean distance.

5.1 Tier-1 Distance

We normalize Hamming distance as follows:

Definition 5.1 (*Normalized Hamming Distance on BitCube*) The distance between two documents can be defined: $\mathrm{dist}(d_i, d_j) = |\mathrm{xOR}(d_i, d_j)| / \mathrm{MAX}\ (\|d_i\|, \| d_j\|)$, where xOR is an exclusive OR operator.

Definition 5.2 (*Similarity*) The similarity of two XML documents (or bitmap rows), d_i and d_j, $\mathrm{sim}(d_i, d_j) = 1 - \mathrm{dist}(d_i, d_j)$. Two documents, d_i and d_j are ξ-similar if $\mathrm{sim}(d_i, d_j) \geq \xi$, where $0 \leq \xi \leq 1$, and ξ is given.

For example, in Figure 2, the similarity of d_1 and d_2 is 1-1/3 = 2/3, while the similarity of d_1 and d_3 is 1-2/4=1/2. That is, d_1 is closer to d_2 than d_3 in terms of ePath. This similarity check can be applied to a bit vectors (ePath-wise in this case). Again, if a bitmap takes element contents or words into account, the similarity in terms of words will be obtained.

5.2 Tier-2 Distance

By taking the frequency of ePath into account, we can check the similarity in topological structure between two documents. Even for the same ePath, there may be more than one topological structure as shown in Section 4. Before going further, in any case of similarity match, one may concern only part of documents. For example, one may want to similar XML articles not all sections but only the Business section. We first define similarity on a target element. Notice that a target element is obtained from the Focusing_Element operation.

Definition 5.3 (*Distance on Focusing Element*) The distance of two XML documents based on particular focusing elements p_k is defined as follows: focusing_dist (d_i, d_j, p_k) = dist (Focusing_Element (d_i , p_k), Focusing_Element (d_j , p_k)).

For example, in Figure 2, although dist (d_1, d_2) = 2/3 and dist (d_1, d_3) = 1/2, if we target only p0, then focusing_dist (d_1, d_2, p_0) = focusing_dist (d_1, d_3, p_0) = 0. That is, d_1, d_2 and d_3 are the same on that target element. Up to this point, we consider distances only on BitCube indexing. The distance will be more significant if InfoCube is considered in addition to BitCube. See the following definition.

Definition 5.4 (*Normalized Euclidean Distance on InfoCube*) Consider two rows of InfoCube. $d_i = [\ b_{i0}, b_{i1}, b_{i2}, \ldots, b_{in}\ ]$, and $d_j = [\ b_{j0}, b_{j1}, b_{j2}, \ldots, b_{jn}\ ]$, where b_i denotes column in the bitmap indexing. The distance of the two XML documents based on an InfoCube is defined as follows:

$$\text{dist\_on\_InfoCube } (d_i, d_j) = \sqrt{\frac{\left(\frac{b_{i0}-b_{j0}}{m_0}\right)^2 + \left(\frac{b_{i1}-b_{j1}}{m_1}\right)^2 + \ldots + \left(\frac{b_{in}-b_{jn}}{m_n}\right)^2}{N}}\text{, where}$$

the two normalization factors m_0 and N are: m_0 = MAX (b_{i0} , b_{j0}) and N = the number of bits in InfoCube (which is n+1 in this case). Two documents, d_i and d_j are ξ-similar if 1 - dist_on_InfoCube (d_i, d_j) ≥ ξ, where 0 ≤ ξ ≤1, and ξ is given.

Notice that the maximum number (m_0) of each bit and the number (N) of bits are used for normalization. For example, in Figure 4,

$$\text{dist\_on\_InfoCube } (d_1, d_2) = \sqrt{\frac{0+1/4+0+4/4+4/9+1+1+0}{8}} = 0.75 \text{ and}$$

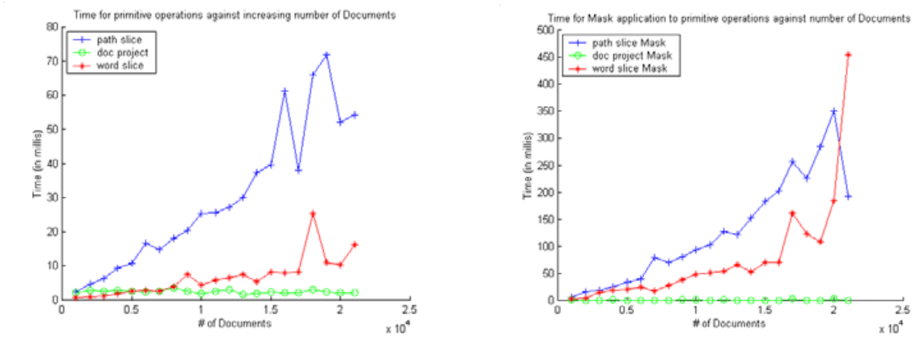
$$\text{dist\_on\_InfoCube } (d_1, d_3) = \sqrt{\frac{0+1/4+4/4+0+0+1+0}{8}} = 0.53.$$

It means that the document d1 is closer to d3 than d2 with respect to the frequency of ePath. Of course, this is more significant if we use the frequency of contents. Then, this distance on InfoCube with respect to content can be used to search similar documents in content.

6 Experimental Results

In order to evaluate the construction and manipulation of BitCubes, InfoCubes, and similarities, we generated over one million documents with the help of a tool called XMLGenerator (http://www.alphaworks.ibm.com/tech/xmlgenerator). We measured the execution time (in milliseconds) of all the InfoCube operations with increasing number of documents. The InfoCube's performance is evaluated for various primitive operations The experiments are conducted for several words, ePaths and documents chosen randomly from the document set. It is made sure that same words/ePaths/documents are queried from the document set across the different tools that are compared against each other. The experimental environment is Windows 2000 with 256M Byte Memory.

We measured the execution time for the three operations: word-Slice, path-Slice, and document-Project with the increasing document set size for the three tools both in BitCubes and InfoCubes. The experimental results in terms of average execution times are plotted in the graphs of Figure 6. Notice that (b) in Figure 6 takes a query like Q1 shown in Section 1.1. It can be easily read from the both graphs that path-slice takes longer than the other operations, doc-project and word-slice. Word-slice increases sharply in (b) due to the number of words increases dramatically in the synthetic XML documents. In any case, processing time in InfoCubes takes five times longer than in BitCubes. It means that sophisticated queries by nature require more processing time. It can be easily read from the graph that the performance of BitCube is better than that of XYZfind and XQEngine.



(a) In BitCube (b) In InfoCube

Fig. 6. Primitive Operation Processing Time

7 Conclusion

The main contributions of this paper are (1) the scalable bitmap indexing to represent XML document collections and speed up query processing, (2) two-tiered distance functions on both BitCubes and InfoCubes. Regular queries can be processed in

BitCubes, while sophisticated queries are in InfoCubes. This scalable bitmap indexing technique further contributes to manipulate multiple aspects and dimensions of information available in digital libraries and to speed up the performance of information retrieval. We plan to develop a clustering technique for multi-dimensional XML documents using the information available in both BitCube and InfoCube.

References

1. Berchtold, S., Keim, D.A., and Kriegel, H.P., The X-tree: An Index Structure for High-Dimensional Data, in Proc. Intl. Conf. On Very Large Data Bases, Bombay, India, pp. 28-39, 1996.
2. Bozkaya, T., and Ozsoyoglu, M., Indexing Large Metric Spaces for Similarity Search Queries, ACM Transactions on Database Systems, pp. 361-404, 1999.
3. Chan, C., and Ioannidis, Y., Bitmap Index Design and Evaluation, in Proc. of Int'l ACM SIGMOD Conference, pp. 355-366, 1998.
4. Deutsch, A., Fernandez, M., and Suciu, D., Storing Semistructured Data with STORED, ACM SIGMOD Int'l Conf. on Management of Data, pp. 431-442,1999.
5. Kim, B., Chakilam, V., and Yoon, J., Spatial Relationship Modeling and Indexing for XML Multimedia Data Retrieval, 3rd Intl ACM SIGMM 2001 Workshop on Multimedia Information Retrieval, Ottawa, Canada, October 5, 2001.
6. Liefke, H., and Suciu. D., XMill: an Efficient Compressor for XML Data, Proc. of Int'l ACM SIGMOD Int'l Conf. on Management of Data, pp. 153-164, 2000.
7. Introduction to MPEG-7 (v3.0), ISO/IEC JTC1/SC29/WG11/N4032, Singapore, March 2001.
8. Ramakrishnan, R., and Gehrke, J., *Database Management Systems*, McGraw Hill, 2000.
9. Rizzolo F. and Mendelzon, A., Indexing XML Data with ToXin, Technical Report, University of Toronto, 2001.
10. Rinfret, D., O'Neil, P., and O'Neil, E., Bit-Sliced Index Arithmetic, ACM SIGMOD Int'l Conf. on Management of Data, pp. 47-57, 2001.
11. Shasha, D., and Wang, T, New Techniques for Best-Match Retrieval, ACM Tractions on Information Systems, pp. 140-158, 1990.
12. Tian, F., DeWitt, D., Chen, J., and Zhang, C., The Design and Performance Evaluation of Alternative XML Storage Strategies, SIGMOD Record, Vol. 31, pp.5-10, 2002.
13. van den Akker, T., Snell, Q. Clement, M., The Yguard Access Control Model: Set-based Access Control, Proc. Of the 6th ACM Symposium on Access Control Models and Technologies, pp. 75-84, 2001.
14. Wu, M., Query Optimization for Selections using Bitmaps, in Proc. Int'l ACM SIGMOD Conference, pp. 227-238, 1999.
15. Yoon, J., Raghavan, V., Chakilam, v., and Kerschberg, L., BitCube: A Three-Dimensional Bitmap Indexing for XML Document, Journal of Intelligent Information Systems, vol. 17, pp. 241-254, 2001.

What People Do When They Look for Music: Implications for Design of a Music Digital Library

Sally Jo Cunningham

Department of Computer Science, University of Waikato,
Private Bag 3105, Hamilton, New Zealand
sallyjo@cs.waikato.ac.nz

Abstract. This paper describes work in progress on an ethnographic study of the searching/browsing techniques employed by people in the researcher's local community. The insights provided by this type of study can inform the development of searching/browsing support for music digital libraries.

1 Introduction

There is a dearth of a priori research on information behavior as regards to music - that is, how a given group of people prefer to locate music and the strategies that they employ for searching or browsing. Without a rich understanding of user needs and information behaviors, the MIR community runs the risk of developing systems ill-suited to their eventual users. The focus here is on eliciting the searching and browsing strategies that people 'natively' employ in searching for music, to provide empirical grounding for design decisions in creating music digital libraries.

2 Methodology

The primary data gathering techniques used are semi-structured interviews and observation of music shoppers. The shopping observations are opportunities to study searching/browsing strategies in a 'natural' setting, while the interviews provide contextual information for the behaviors observed. Participants are asked about their music shopping strategies, their preferred sources for music and information about music, the ways they typically 'use' music, and the social contexts for locating music.

3 Preliminary Results

Observations and interviews confirm that people conduct known-item searches - they look for a specific music document based on known features of the document.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 177-178, 2002.

Features guiding the search were bibliographic (primarily artist/group name, album name, and song title), indicating the importance of including quality bibliographic data in a music digital library. While there is anecdotal evidence that music-seekers may sing remembered bits of songs, no such behavior was observed, and interview subjects professed a reluctance to approach either librarians or music shop workers with queries (sung or spoken). Taken together, these observations suggest that while sung queries may not be common, a 'query by humming' interface might be welcomed, as users could avoid the discomfiture of singing in the hearing of another person.

Browsing is more exploratory and less directed than searching. Music shoppers browse the contents of a shop or public library mainly by genre. The genres do not have to be tightly defined—shoppers are often willing to spend significant amounts of time flipping through a category of interest. This observation suggests that a music digital library should also support browsing by genre, but that care should be taken to avoid the construction of overly narrow clusters of music documents.

The shoppers appreciate stores whose genre categorizations correspond to their personal view of which artists or compilations are 'like' each other. People may also develop their own, idiosyncratic categories, such as 'gym music' or 'study music', or group music by its emotional impact ('depressing', etc.). Since the construction of these novel categories involves mentally grouping familiar songs, the individual has at hand examples of songs falling within the novel genre; locating new songs in an idiosyncratic genre could be supported by search tools that allow the user to ask for 'more songs like these'. A next step is to clearly identify the musical facets most useful for characterizing genres—timbre, instrumentation, rhythm, etc—and to develop interfaces for specifying musical query-by-example searches.

Shoppers keep up to date on the music available for sale through *monitoring*: they peruse the store displays to stay current on the locally obtainable offerings. A digital library that supports such monitoring would provide "what's new in this collection" summaries that could be subdivided by genre as well as date.

The images on CD covers are used by shoppers in a variety of ways: to speed searches through a large stack of CDs; to scan distant stacks for interesting images; to provide guidance on the genre of music by an unfamiliar artist; and to quickly indicate an interesting CD to wandering companions (by holding up the CD). In a music digital library, images could be used as thumbnails accompanying search hits or browsing displays, to support fast scanning for items or genres of interest.

This study suggests that findings from studies of music information behavior, as displayed in settings such as music stores and public libraries, can be used to inform the design of useful and usable music digital libraries.

Distributing Relevance Feedback in Content Based Image Retrieval Systems

Raghu Menon and Raj Acharya

Department of Electrical Engineering, SUNY at Buffalo, Buffalo, New York, USA

Abstract. Most Content Based Image Retrieval (CBIR) systems use low level features such as texture, color and shape to formulate the query. Relevance feedback (RF) from the environment has also been used as a means of training CBIR systems and improving the performance of feature based query schemes. Query schemes based on feature extraction methods generally make recognition errors of different types, and hence a scheme that would exploit this "error independence" among these schemes could be used to improve the performance of a combined system using these features. We propose a scheme for combining results from a number of low level feature based image classifiers based on the relative relevance of the features and distributing RF from the environment to each of the low level classifiers to improve their performance.

1 Distributed Relevance Feedback

The combination of results from multiple feature based image classification schemes, and the refinement of these results using a distributed relevance feedback method form the basis of our approach to improving retrieval accuracy

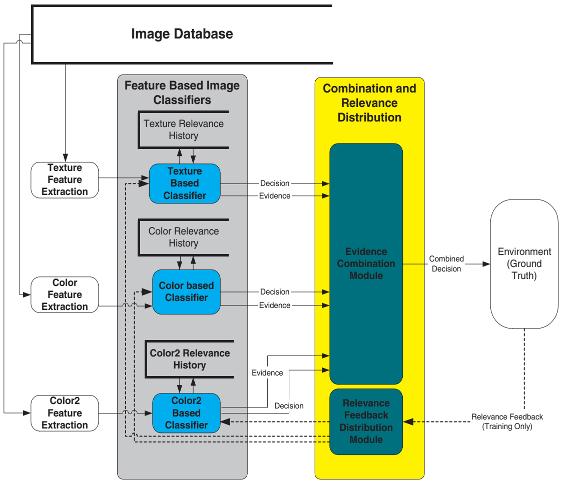


Image Database
Feature Based Image Classifiers
Texture Relevance History
Texture Feature Extraction
Texture Based Classifier
Color Relevance History
Color Feature Extraction
Color based Classifier
Color2 Relevance History
Color2 Feature Extraction
Color2 Based Classifier
Decision
Evidence
Decision
Evidence
Evidence
Decision
Combination and Relevance Distribution
Evidence Combination Module
Relevance Feedback Distribution Module
Combined Decision
Environment (Ground Truth)
Relevance Feedback (Training Only)

Fig. 1. Diagram of proposed Distributed Relevance Feedback Based CBIR System

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 179–181, 2002.

Table 1. Recall Results on image database

Individual classifier on Vistex DLP				**Individual classifier on Berkeley DLP**			
Classifier	*Kohonen*	*LVQ*	*RF(Reinforcement)*	*Classifier*	*Kohonen*	*LVQ*	*RF(Reinforcement)*
Red	72.8	88.47	86.42	Red	65.1	73.2	71.31
Blue	67.28	81.27	80.45	Blue	57.6	73.1	72.3
Green	76.54	88.47	86.54	Green	63.1	73.15	71.4
HSV moments	77.16	89.09	88.4	HSV moments	68.1	74.3	73.4
Wavelet moments	70.3	84.97	82.3	Wavelet moments	63.1	74.51	73.21
ASF	71.4	87.3	84.3	ASF	64.2	75.41	74.23
Evidence Combination(EC) and RF on VisTex				**Evidence Combination (EC)and RF on Berkeley DLP**			
EC (except blue)	90.53	91.6	91.7	EC (except blue)	82.4	82.	82.2
EC (including blue)	91.3	92.0	92.1	EC (including blue)	83.2	83.4	83.6
Distributed RF	93.2	93.2	93.2	Distributed RF	90.51	90.56	90.42

and relevance in a CBIR system. An overview of the system is shown in figure 1. Each of these feature classifiers can be considered "error-independent" of the other classifiers. i.e. the retrieval errors due to one classifier are independent of the errors made by other classifiers. This error-independence is exploited to improve retrieval performance of a combined system. Low level features are used to classify the images into classes using a self-organizing neural network. Evidences are computed for each classifier. Dempster-Shafer (DS) based evidence combination is used to provide a combined decision to the environment. In training mode, a set of labeled data from the image database provide the ground truth information for the environment to present binary RF back to the system. In retrieval mode, all the images belonging to the chosen class are displayed in order of evidence measure as the retrieved results. RF from the environment is distributed to each classifier. Each classifier modifies its decision basis through a weight re-adjustment based on the RF. We will discuss briefly the process of distributing RF. At time instant t, let $O(t)$ be the output presented to the environment. The environment, after receiving the output, computes a reinforcement signal $r_t \in [-1, 1]$ as an evaluation of the previous output of the system at time t. Let $h_k(t) \in [-1, 1]$ be a measure of the agreement between classifier k and the evidence combination module in its decision. We employ additional reinforcement for each classifier k through a reinforcement estimator term $re_k(t)$. To develop this term, we diminish the contribution of past reinforcements through a normal distribution. The reinforcement estimator is given by $re_k(t) = [\sum_{j=1}^{T} r_k(t-j)\frac{e^{(-j/2\sigma)}}{\sigma^2}]/T$ where T denotes the history of reinforcements and σ is the standard deviation of a normal distribution used to diminish importance of past reinforcements with time. Then the reinforcement to classifier k is $r_k(t) = (r(t).h_k(t) + re_k(t))/2$. The computed reinforcement is used by each individual classifier to adjust its performance.

2 Results and Conclusions

The proposed algorithms were tested on the Berkeley Digital Library Project data set and the VisTex dataset from MIT. Precision and recall were used as metrics for evaluating performance. The experiments were performed for each image feature type using unsupervised(Kohonen feature maps), supervised (Learning Vector Quantization), and Relevance feedback (Reinforcement Learning) based image classifiers. The supervised classifier performed better than the unsupervised and reinforcement learning based classifiers. The improvement in performance was lower in the Berkeley DLP project database as there were a larger number of categories. Thus, the limitations of individual classifiers was apparent in the results. The second set of experiments consisted of evaluating the performance of evidence combination and distributed reinforcement learning on the CBIR system. The recall results were as presented in table 1. The columns represent the starting points of the evidence combination process. i.e. in the first case, the converged weights of the unsupervised classifier formed the initial weights for the RF learning process. The following inferences can be drawn from the presented results. Combination of evidence and distributed RF give consistently better results than any of the individual classifiers. Even though the blue classifier was the poorest performing individual classifier, on evidence combination, the collective performance was improved on adding the blue classifier. Thus, in an information poor environment, the method provides a means to exploit all available information. Distributed learning performed better than all other methods. Further, the retrieval performance for the distributed RF was more or less the same regardless of the starting point of the process. Thus, the convergence performance of the method is borne out in experimentation. Results show that the proposed method can significantly improve the performance in CBIR systems.

Multistrategy Learning of Rules for Automated Classification of Cultural Heritage Material

G. Semeraro, F. Esposito, S. Ferilli, N. Fanizzi,
T.M.A. Basile, and N. Di Mauro

Dipartimento di Informatica
Università di Bari
via E. Orabona, 4 - 70125 Bari - Italia
{semeraro, esposito, ferilli, fanizzi, basile, nicodimauro}@di.uniba.it

Abstract. This work presents the application of a new, enhanced version of the incremental learning system INTHELEX (INcremental THEory Learner from EXamples), the learning component in the architecture of the EU project COLLATE, dealing with the annotation of cultural heritage documents. Due to the complex shape of the handled material, the addition of multistrategy capabilities was needed to improve the effectiveness and efficiency of the learning process. Some results demonstrating the benefits that the addition of each strategy can bring are also reported.

1 Introduction: The COLLATE Project

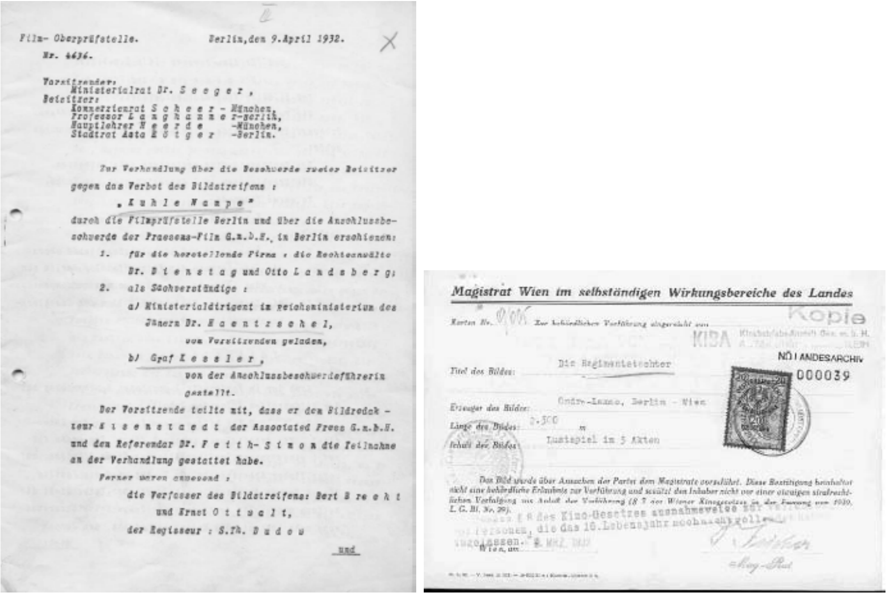
Many important historic and cultural sources, which constitute a major part of our cultural heritage, are fragile and distributed in various archives. Such a situation is an obstacle to full access, knowledge and usage. Also, many informal and non-institutional contacts between archives constitute specific professional communities, which today still lack effective and efficient technological support for cooperative and collaborative knowledge working. The COLLATE project[1] aims at developing a WWW-based *collaboratory* [7] for archives, researchers and end-users working with digitized historic/cultural material.

Though the developed tools and interfaces are generic, the chosen experimental domain concerns historic film documentation. Multi-format documents on European films of the early 20th century, provided by three major national film archives, include a large corpus of rare historic film censorship documents from the 20s and 30s, as well as newspaper articles, photos, stills, posters and film fragments. An in-depth analysis and comparison of such documents can give evidence about different film versions and cuts, allow restoration of lost/damaged films, and identify actors and film fragments of unknown origin.

All material is analyzed, indexed, annotated and interlinked by film experts. The COLLATE system will provide suitable task-based interfaces and knowledge

[1] IST-1999-20882 project COLLATE - *Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material* (URL: http://www.collate.de).

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 182–193, 2002.



Film- Oberprüfstelle. Berlin, den 9.April 1932.
Nr. 4636.

Vorsitzender:
Ministerialrat Dr. Seeger,
Beisitzer:
Kommerzienrat Scheer - München,
Professor Langhammer-Berlin,
Hauptlehrer Neerde -München,
Stadtrat Asta Rötger -Berlin.

Zur Verhandlung über die Beschwerde zweier Beisitzer gegen das Verbot des Bildstreifens :

„Kuhle Wampe"

durch die Filmprüfstelle Berlin und über die Anschlussbeschwerde der Praesens-Film G.m.b.H. in Berlin erschienen:

1. für die herstellende Firma : die Rechtsanwälte Dr. Dienstag und Otto Landsberg;
2. als Sachverständige :
   a) Ministerialdirigent im Reichsministerium des Innern Dr. Haentzschel, vom Vorsitzenden geladen,
   b) Graf Kessler, von der Anschlussbeschwerdeführerin gestellt.

Der Vorsitzende teilte mit, dass er dem Bildredakteur Eisenstaedt der Associated Press G.m.b.H. und dem Referendar Dr. Feith-Simon die Teilnahme an der Verhandlung gestattet habe.

Ferner waren anwesend :

die Verfasser des Bildstreifens: Bert Brecht und Ernst Ottwalt,
der Regisseur : S.Th. Dudow

und

(a)

Magistrat Wien im selbständigen Wirkungsbereiche des Landes

Karten Nr. 9/06 Zur behördlichen Vorführung eingereicht von

Kopie

KIBA Kinobetriebs-Anstalt Ges. m. b. H.

NÖ LANDESARCHIV
000039

Titel des Bildes: Die Regimentstochter

Erzeuger des Bildes: Ondra-Lamac, Berlin - Wien

Länge des Bildes: 2.300 m

Inhalt des Bildes: Lustspiel in 5 Akten

Das Bild wurde über Ansuchen der Partei dem Magistrate vorgeführt. Diese Bestätigung beinhaltet nicht eine behördliche Erlaubnis zur Vorführung und schützt den Inhaber nicht vor einer etwaigen strafrechtlichen Verfolgung aus Anlaß der Vorführung (§ 7 des Wiener Kinogesetzes in der Fassung von 1930, L. G. Bl. Nr. 29).

Gemäss § 8 des Kino-Gesetzes ausnahmsweise zur Vorführung vor Personen, die das 16. Lebensjahr noch nicht vollendet haben, zugelassen.

Wien, am 8. MRZ. 193[illegible]

Mag.-Rat

(b)

Fig. 1. Sample COLLATE documents

management tools to support both individual work and collaboration. Continuously integrating newly derived user knowledge into its digital data and metadata repositories, the system can offer an improved content-based retrieval functionality. Enabling users to create and share valuable knowledge about the cultural, political and social contexts allows in turn other end-users to better retrieve and interpret the historic material.

Supported by successful experience in the application of symbolic learning techniques to the classification and understanding of paper documents [4,6,11], our aim is applying INTHELEX to these documents. The objective is learning to automatically identify and label document classes and significant components, to be used for indexing/retrieval purposes and to be submitted to the COLLATE users for annotation. Combining results from the manual and automatic indexing procedures, elaborate content-based retrieval mechanisms can be applied [2]. The challenge comes from the low layout quality and standard of such materials, which introduce a considerable amount of noise in their description (see Figure 1). Regarding the layout quality, it is often affected by manual annotations, stamps that overlap sensible components, ink specks, etc.. For the layout standard, many documents are typewritten sheets, that consist of all equally spaced lines in Gothic type. Such a situation should account for a profitable use of automated reasoning capabilities in INTHELEX, such as abduction and abstraction. While the former can make the system more flexible in the absence of particular layout components due to the typist's style, the latter can help in focusing on layout patterns that are meaningful to the identification of interesting ones, neglecting less interesting details. Preliminary experiments showed that INTHELEX is able to distinguish at least 3 classes of COLLATE censorship documents, and to single out a number of logical components inside them. For instance, it learns rules that can separate the censorship authority, applicant and decision in documents like the one in Figure 1-b.

The following section presents the system INTHELEX along with its multistrategy capabilities. Section 3 describes the experiment and discusses the results. Lastly, Section 4 draws some conclusions and outlines future work directions.

2 INTHELEX: The Learning Component

Incremental learning is necessary when either incomplete information is available at the time of initial theory generation, or the nature of the concepts evolves dynamically. Both cases are very frequent in real-world situations, hence the need for incremental models to complete and support the classical batch ones, that perform learning in one step and thus require the whole set of observations to be available from the beginning.

INTHELEX (INcremental THEory Learner from EXamples) is a learning system for the induction of *hierarchical* logic theories from examples [5]: it learns theories expressed in a first-order logic representation from positive and negative examples; it can learn simultaneously *multiple concepts*, possibly related to each other (recursion is not allowed); it retains all the processed examples, so to guarantee validity of the theories learned from all of them; it is a *closed loop* learning system (i.e. a system in which feedback on performance is used to activate the theory revision phase [1]); it is *fully incremental* (in addition to the possibility of refining a previously generated version of the theory, learning can also start from an empty theory); it is based on the *Object Identity assumption* (terms, even variables, denoted by different names within a formula, must refer to different objects)[2].

INTHELEX incorporates two refinement operators, one for generalizing hypotheses that reject positive examples, and the other for specializing hypotheses that explain negative examples. It exploits a (possibly empty) previous version of the theory, a graph describing the dependence relationships among concepts, and a historical memory of all the past examples that led to the current theory. Whenever a new example is taken into account, it is stored in such a repository and the current theory is checked against it.

If it is positive and not covered, generalization must be performed. One of the definitions of the concept the example refers to is chosen by the system for generalization. If a generalization can be found that is consistent with all the past negative examples, then it replaces the chosen definition in the theory, or else another definition is chosen to be generalized. If no definition can be generalized

[2] This often corresponds to human intuition, while allowing the search space to fulfill nice properties affecting efficiency and effectiveness of the learning process [10].

in a consistent way, the system checks if the exact shape of the example itself can be regarded as a definition that is consistent with the past negative examples. If so, it is added to the theory, or else the example itself is added as an exception.

If the example is negative and covered, specialization is needed. Among the theory definitions involved in the example coverage, INTHELEX tries to specialize one at the lowest possible level in the dependency graph by adding to it positive information, which characterize all the past positive examples and can discriminate them from the current negative one. In case of failure on all of the considered definitions, the system tries to add negative information that is able to discriminate the negative example from all the past positive ones, to the definition related to the concept the example is an instance of. If this fails too, the negative example is added to the theory as an exception. New incoming observations are always checked against the exceptions before applying the rules that define the concept they refer to.

Another peculiarity in INTHELEX is the integration of other forms of automated reasoning, that may help in the solution of the theory revision problem by pre-processing the incoming information [6]. Namely, deduction is exploited to fill observations with information that is not explicitly stated, but is implicit in their description. There is thus the possibility of better representing the examples and, consequently, the inferred theories. Conversely, abduction aims at completing possibly partial information in the examples (adding more details), whereas abstraction removes superfluous details from the description of both the examples and the theory. Thus, even if with opposite perspectives, both aim at reducing the computational effort required to learn a correct theory with respect to incoming examples.

2.1 Deduction

INTHELEX requires the observations to be expressed only in terms of the predicates that make up the description language for the given learning problem. To ensure uniformity of the example descriptions, such predicates have no definition. Nevertheless, since the system is able to handle a hierarchy of concepts, combinations of these predicates might identify higher level concepts that are worth adding to the descriptions in order to raise their semantic level. Thus, INTHELEX implements a saturation operator that exploits deduction to recognize such concepts and explicitly add them to the description of the examples.

The system can be provided with Background Knowledge containing (complete or partial) definitions in the same format as the theory rules. The background knowledge is supposed to be correct, and hence is not modifiable. This way, any time a new example is considered, a preliminary saturation phase can be performed, that adds the higher level concepts whose presence can be deduced from such rules by subsumption and/or resolution. Differently from abstraction, described below, all the specific information used by saturation is left in the example description. Hence, it is preserved in the learning process until other evidence reveals it is not significant for the concept definition, which is a more cautious behaviour. This is fundamental if some concepts to be learnt are related, since their definition could not be stable yet, and hence one cannot afford to drop the source from which deductions were made in order to be able to recover from deductions made because of wrong rules.

2.2 Abduction

Induction and abduction are both important strategies to perform hypothetical reasoning (i.e., inferences from incomplete information). Induction yields the inference, from a certain number of significant observations, of regularities and laws valid for the whole population. Abduction was defined by Peirce as hypothesizing some facts that, together with a given theory, could explain a given observation.

According to the framework proposed in [8], an *abductive logic theory* is made up of a normal logic program [9], a set of *abducibles* and a set of *integrity constraints* (each corresponding to a combination properties/relations that is not allowed to occur). Abducibles are the predicates about which assumptions (*abductions*) can be made: They carry all the incompleteness of the domain (if it were possible to complete these predicates, then the theory would be correctly described). Integrity constraints provide indirect information about them and, since several explanations may hold for this problem setting, are also exploited to encode preference criteria for selecting the best ones.

The proof procedure implemented in INTHELEX starts from a goal and a set of initial assumptions, and results in a set of consistent hypotheses (abduced literals) by intertwining *abductive* and *consistency derivations*. Intuitively, an abductive derivation is the standard Logic Programming derivation suitably extended in order to consider abducibles. As soon as an incompleteness is detected in an observation, the corresponding information, if abducible, is added to the observation itself, provided that any related integrity constraint is satisfied. This is checked by starting a consistency derivation. Each integrity constraint related to the abduced fact is considered satisfied if at least one of its components does not hold. In the consistency derivation, when an abducible is encountered, an abductive derivation for its complement is started in order to prove its falsity.

This procedure can be exploited in order to complete the document descriptions contained in the examples, so that the system is less sensitive to missing information (e.g., missing layout features).

2.3 Abstraction

Abstraction is a pervasive activity in human perception and reasoning. When we are interested in the role it plays in Machine Learning, inductive inference must be taken into account as well. The exploitation of abstraction concerns the shift of representation languages from the language in which the theory is described to a higher level one.

According to the framework proposed in [12], concept representation deals with entities belonging to three different levels. Concrete objects reside in the *world*, but any observer's access to it is mediated by his *perception* of it. To be available over time, these stimuli must be memorized in an organized *structure*, i.e. an *extensional* representation of the perceived world. Finally, to reason about the perceived world and communicate with other agents, a *language* is needed, that describes it *intensionally*. Modifications to the structure and language are just a consequence of differences in the perception of the world (due, e.g., to the medium used and the focus-of-attention). Thus, abstraction takes place at the world-perception level, and then propagates to higher levels, by means of a set of operators. An abstraction theory contains information for performing the shift specified by the abstraction operators.

In INTHELEX, it is assumed that the abstraction theory is already given (i.e. it has not to be learned by the system), and that the system automatically applies it to the learning problem at hand before processing the examples. The implemented abstraction operators allow the system to replace a number of components by a compound object, to decrease the grain-size of a set of values, to ignore whole objects or just part of their features, and to ignore the number of occurrences of some kinds of object.

3 Experiments

Some experiments were run to test the improvement coming from the addition of abduction and abstraction to the process of learning definitions, for some classes of censorship documents provided by FilmArchiv Austria (FAA) and Deutsches FilmInstitut (DIF). Specifically, we used registration cards coming from the former ('`faa_registration_card`', see Figure 1-b) and censorship decisions coming from the latter ('`dif_censorship_decision`', see Figure 1-a). The dataset consisted of 34 documents for the class `faa_registration_card`[3], 19 documents for the class `dif_censorship_decision`[4] and 61 reject documents, obtained from newspaper articles and DIF registration cards. Note that the symbolic method adopted allows the trainer to specifically select prototypical examples to be included in the learning set. This explains why theories with good predictiveness can be obtained even from few observations.

The first order descriptions of such documents, needed to run the learning system, were automatically generated by the system WISDOM++ [3]. Starting from scanned images, such a system is able to identify the layout blocks that make up a paper document, along with their type and relative positions. Each document was then described in terms of its composing layout blocks, along with their size

[3] A certification that the film has been approved for exhibition in the present version by the censoring authority. The "registration cards" were given to the distribution company, who had to pay for this. They enclosed the cards with the prints. The police checked the cinemas from time to time, and the owner or projectionist had then to show the registration card.

[4] Decision whether a film could or could not, and in which version, be distributed and shown throughout a country. The "censorship decision" is often a protocol of the examination meeting and is issued by the censorship office or headquarters.

Table 1. `faa_registration_card` Classification

	Clauses	Length	lgg	Runtime	Accuracy	Pos	Neg	t-test
'Manual' Discretization	1	10,5	9,5	11,9	0,98	0,94	1,0	1,0
Numeric abstraction	1	13,1	8	19,29	0,99	0,97	1,0	1,0
Speckle abstraction	1	12,9	8,7	18,24	0,98	0,94	1,0	1,0
Abduction	1,1 (2)	13,1 (40,7)	8,7 (9,7)	123,69 (40,9)	0,99 (0,99)	0,97 (0,97)	1,0 (1,0)	1,0 (1,0)
Abduction+Abstraction	1 (2)	12,5 (42)	8 (9)	90,13 (36,42)	1,0 (0,99)	1,0 (0,97)	1,0 (1,0)	1,0 (1,0)

Table 2. `dif_censorship_decision` Classification

	Clauses	Length	lgg	Runtime	Accuracy	Pos	Neg	t-test
'Manual' Discretization	1,6	52,3	8	71,92	0,94	0,74	0,99	0,98
Numeric abstraction	1	23,3	8,4	49,6	0,97	0,84	1,0	0,99
Speckle abstraction	1	23,3	8,6	47,64	0,97	0,84	1,0	0,99
Abduction	1,1 (1,2)	28,3 (33,4)	8,7 (8,7)	5667,8 (50,37)	0,95 (0,95)	0,74 (0,74)	1,0 (1,0)	0,98 (0,97)
Abduction+Abstraction	1,1 (1,2)	25,8 (33,4)	9,2 (8,7)	5761,8 (50,37)	0,96 (0,94)	0,79 (0,74)	1,0 (0,99)	0,99 (0,99)

(height and width), position (horizontal and vertical), type (text, line, picture and mixed) and relative position (horizontal/vertical alignment, adjacency). The description length of the documents for class `faa_registration_card` ranges between 40 and 379 literals (144 on average); for class `dif_censorship_decision`, it ranges between 54 and 263 (215 on average).

Each document was considered as a positive example for the class it belongs to, and as a negative example for the other class (to be learned from); reject documents were considered as negative examples for both classes. Definitions for each class were learned, starting from the empty theory and with all the negative examples at the beginning (in order to simulate a batch approach), and their predictive accuracy was tested according to a 10-fold cross validation methodology, ensuring that all the learning and test sets contained the same proportion of positive and negative examples.

3.1 Experimental Baseline

Various experiments were performed on both classes, whose results (all averaged on the 10 runs) are reported in Table 1 (as regards `faa_registration_card`) and in Table 2 (concerning `dif_censorship_decision`). For each case, the following data are reported - *Clauses* : number of clauses (i.e., alternative definitions for the concept) in the learned theory; *Length* : number of literals composing the clauses; *lgg* : number of generalizations needed to obtain the theory; *Runtime* : computational time required to learn the theory (expressed in seconds); predictive accuracy rate computed on the test set (*Accuracy* overall, *Pos* on positive examples only, *Neg* on negative examples only); *t-test* : expected predictive accuracy according to a t-test with confidence level $\alpha = 0,05$.

In the following paragraphs, we will discuss the outcomes in more detail, and try to justify the conclusions we draw.

A preliminary problem was the fact that INTHELEX is currently unable to handle numeric descriptors, whereas WISDOM++ expresses the blocks' dimensions and positions as numeric values (number of pixels). Hence, a discretization was needed to assign each specific value to a symbolic label representing an interval. When abstraction had not yet been added to INTHELEX, we had to perform such a transformation through a purposely implemented routine. Now, we are able to delegate this task to the system itself, by exploiting one of its abstraction operators (the one acting on the grain size). Comparing the first two rows of Tables 1 and 2, which report on this performance in the two cases, we note that there is no loss in the 'Numeric abstraction' case (second row) compared to 'Manual discretization' (first row). Actually, there is a slight improvement except for computational time for the first case, probably due to abstraction changing the ordering of the literals in the observations. Thus, in all the subsequent experiments, INTHELEX ran the numeric discretization phase by abstraction.

3.2 Motivation for Multistrategy Learning in COLLATE

Since the available documents were often affected by the presence of speckles, identified by WISDOM++ as layout components and hence appearing in the description of the document, we decided to use abstraction to eliminate them. The underlying rationale is that such a 'cleaning' should hopefully help the system in at least two respects. First, by focusing on significant layout components, that are more discriminant, this should lead to more characterizing definitions of the concepts. Second, having shorter example descriptions can have a positive effect on the learning time. Specifically, the abstraction theory considered as speckles all the blocks without a clear type (mixed) that are short and/or narrow. Another issue to be faced in the COLLATE project is the low quality of some documents, due to their age and to the absence, in many cases, of a standard layout. While the documents in our dataset were chosen to have an acceptable quality and a sufficiently standard layout, it is foreseeable that worse documents will be available in the future. To simulate the behaviour of INTHELEX in such a possible scenario, we corrupted part of the documents in the dataset, and tried to apply abduction in order to overcome the problems raised by missing components. Specifically, incomplete documents were generated by randomly dropping 10% of the description from 30% of the available documents, and then letting INTHELEX use abduction. All the basic predicates in the description language concerning block dimensions, types and positions were considered as abducibles, while integrity constraints were set to express the mutual exclusion among layout block sizes, types and positions. INTHELEX was allowed to exploit abduction to hypothesize facts, concerning the above descriptors, only in case of failure in finding a correct generalization, before adding a new clause to the theory.

Abduction makes sense in this environment since the absence of a layout block in a document could be due to the writer not fulfilling the style requirements, and not to the insignificance of that block for a correct definition. In other words, a block should not be dropped from the definition just because a few examples miss it; conversely, integrity constraints are in place to ensure that superfluous blocks found in the first few examples do not introduce unnecessary blocks that can be always abduced in the future. The last three rows of Tables 1 and 2 report the experiments with speckles abstraction only, abduction on the corrupted dataset only, and both, respectively (in the last case, abstraction precedes abduction). Rows involving abduction also report, in parentheses for comparison purposes, the performance obtained on the corrupted dataset without exploiting abduction.

3.3 Discussion of Experimental Results

Let us first focus on the experiments concerning the `faa_registration_card` class (see Table 1). Abstraction of speckles cannot, of course, improve the number of clauses (that is already 1). Nevertheless, the shorter example descriptions have a beneficial effect on the learning time, in spite of the greater number of generalizations performed (probably due to the absence of speckles in negative examples, that formerly helped to avoid their coverage). The greater difficulty in avoiding coverage of negative examples also results in a more specific clause, as showed by the slight decrease in predictive accuracy on positive examples.

As for abduction, the experiments prove that it is able to balance the corruption in the examples, even if at the cost of a slightly worse number of clauses and lggs, and of a significant increase in the runtime (due to the necessity of checking the consistency of each hypothesized fact). Indeed, predictive accuracy is the same as when only inductive operators are employed, even if a great portion of the descriptions is now missing. The benefit becomes more evident, especially as regards the number of clauses and their length, if we compare the performance to what would be obtained on the corrupted dataset without exploiting abduction, as reported in parentheses.

Finally, the results of the joint effort of abduction and abstraction are shown in the last row. The theories learned in this case are, indeed, the best of all cases: they outperform all the previous ones as regards predictive accuracy (100%), and are better than any single strategy as to the number of clauses and lgg's performed. Also the runtime, even though worse than that of abstraction only, is far better than that of abduction only. The average number of literals in each clause is almost stable, with a slight decrease when abstraction is used, signifying that the system was always able to grasp the core concept and that abstraction actually eliminated only superfluous information.

Considering the experiments on the other class, `dif_censorship_decision` (see Table 2), we immediately note that runtime is always higher than for the previous class, while the predictive accuracy is always lower (even if still very high). The former suggests that we are facing an intrinsically harder learning problem (indeed, documents in this class have a larger size and a higher number of identified layout components – typically, each row in the document is considered a separate block). The latter may be due to the availability of fewer examples, leading to a less refined theory that is not predictive enough, as supported by the number of literals in the definitions. It seems that more characteristics than in the other class must be preserved to find a solution. Indeed, the portion of the original example length that is dropped ranges from 85,68% to 89,17% (against the 90,91–92,71% for the other class). The above comments concerning speckle abstraction generally still hold, including the improvement that it can bring when added to abduction.

On the other hand, this class seems to be particularly idiosyncratic with respect to abduction. In this respect, it should be noted that runtime in the last two rows is heavily affected by the search for useful abductions in one fold, as proved by the fact that, if we restrict the search to the remaining 9 folds, the mean decreases to 54,87 and 38,87, respectively. The poor performance of abduction in this case seems to be confirmed by comparison with the statistics in parentheses, particularly as regards the clause length, whereas the other class showed a significant improvement. This could be due to the fact that corrupting the descriptions makes an already difficult problem even more difficult. Another possible explanation is the fact that corrupting the description of documents in this class results in the elimination of many literals. This can be compensated up to some extent by the generalizations, that are able to find other common (but probably less characterizing and meaningful) features among the given documents. Such accidental similarities could cause overfitting on the training data, that abduction is not able to compensate during the test phase. This is confirmed by the lower predictive accuracy being concentrated in the coverage of positive examples, whereas negative examples are always rejected.

3.4 Insight into Multistrategy Operators' Performance

The difference in effectiveness of abduction in the two classes led us to closely examine the phenomenon, in order to better understand it and its possible causes. Table 3 summarizes, for both classes, the average number of examples on which literals have been abduced (first row) and the average number of literals abduced for each of them (second row), both with and without abstraction of speckles. Our expectations were confirmed in that, for the class `dif_censorship_decision`, nearly no abduction was made, which explains why no improvement was obtained with respect to the other cases (including learning without abduction on corrupted examples). More specifically, we found that the only abductions were carried out for the 'computation-intensive' fold, which again suggests that this class seldom requires abduction, and in those cases it is a hard task anyway. On the contrary, as regards class `faa_registration_card`, abduction improves performance since it succeeds on hypothesizing more literals (2 out of about 13 that make up the definition) and on more examples.

Similarly, we have collected in Table 4 information on the effects of eliminating speckles from document descriptions through abstraction. For both types of documents, the number of examples in which abstraction took place is reported

Table 3. Abduction performance

	`faa_registration_card`		`dif_censorship_decision`	
	w/ abstraction	w/o abstraction	w/ abstraction	w/o abstraction
Examples	1,9	1,6	0,1	0,1
Literals	1,53	1,5	0,2	0,2

Table 4. Speckles abstraction performance

	`faa_registration_card`	`dif_censorship_decision`	Reject
Examples	14	9	3
Literals	114	37	140

(first row), along with the average number of dropped literals (second row). It turns out that only a part of the whole training set suffered from speckles, and in those cases a relevant portion of the descriptions was removed. It is noticeable that more speckles were found in `faa_registration_card` documents than in `dif_censorship_decision` ones, even if the average description length of the latter was larger than that of the former, which indicates a better layout quality in the latter. This also means that abstraction cannot lower the document descriptions' complexity for the latter class, which results in harder work for the other operators.

4 Conclusions and Future Work

Multistrategy approaches to machine learning can help to improve efficiency, and are necessary in a number of real-world situations. The incremental learning system INTHELEX works on first-order logic representations. Its multistrategy learning capabilities have been further enhanced to improve effectiveness and efficiency of the learning process, by augmenting pure induction and abduction with abstraction and deduction.

This paper presented and discussed some experimental results demonstrating the benefits that the addition of each strategy can bring to the task of document classification. Even if the performance obtained exclusively by the inductive operator was very good, multistrategy operators contributed to make it even better from both the effectiveness and the efficiency viewpoints. INTHELEX is included in the architecture of the EU project COLLATE, to learn rules for automated classification of cultural heritage documents dating back to the 20s and 30s.

Future work will concern more extensive experimentation, aimed at finding tighter ways of cooperation among the learning strategies. It will also be interesting to apply the same techniques to learn rules for interpreting the semantic role played by meaningful layout components in the documents. An analysis of the complexity of the presented techniques is also planned. Moreover, the addition of numeric capabilities can be considered fundamental for effective learning in some contexts, and hence deserves further study.

Acknowledgments. This work was partially funded by the EU project IST-1999-20882 COLLATE "Collaboratory for Annotation, Indexing and Retrieval of Digitized Historical Archive Material".

References

[1] J. M. Becker. Inductive learning of decision rules with exceptions: Methodology and experimentation. B.s. diss., Dept. of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA, 1985. UIUCDCS-F-85-945.

[2] H. Brocks, U. Thiel, A. Stein, and A. Dirsch-Weigand. Customizable retrieval functions based on user tasks in the cultural heritage domain. In P. Constantopoulos and I.T. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries*, number 2163 in LNCS. Springer, 2001.

[3] F. Esposito, D. Malerba, and F.A. Lisi. Machine learning for intelligent processing of printed documents. *Journal of Intelligent Information Systems*, 14(2/3):175–198, 2000.

[4] F. Esposito, D. Malerba, G. Semeraro, N. Fanizzi, and S. Ferilli. Adding machine learning and knowledge intensive techniques to a digital library service. *International Journal on Digital Libraries*, 2(1):3–19, 1998.

[5] F. Esposito, G. Semeraro, N. Fanizzi, and S. Ferilli. Multistrategy Theory Revision: Induction and abduction in INTHELEX. *Machine Learning Journal*, 38(1/2):133–156, 2000.

[6] S. Ferilli. *A Framework for Incremental Synthesis of Logic Theories: An Application to Document Processing*. Ph.D. thesis, Dipartimento di Informatica, Università di Bari, Bari, Italy, November 2000.

[7] R.T. Kouzes, J.D. Myers, and W.A. Wulf. Collaboratories: Doing science on the internet. *IEEE Computer*, 29(8), 1996.

[8] E. Lamma, P. Mello, F. Riguzzi, F. Esposito, S. Ferilli, and G. Semeraro. Cooperation of abduction and induction in logic programming. In A. C. Kakas and P. Flach, editors, *Abductive and Inductive Reasoning: Essays on their Relation and Integration*. Kluwer, 2000.

[9] J. W. Lloyd. *Foundations of Logic Programming*. Springer-Verlag, Berlin, second edition, 1987.

[10] G. Semeraro, F. Esposito, D. Malerba, N. Fanizzi, and S. Ferilli. A logic framework for the incremental inductive synthesis of datalog theories. In N. E. Fuchs, editor, *Logic Program Synthesis and Transformation*, number 1463 in LNCS. Springer, 1998.

[11] G. Semeraro, S. Ferilli, N. Fanizzi, and F. Esposito. Document classification and interpretation through the inference of logic-based models. In P. Constantopoulos and I.T. Sølvberg, editors, *Research and Advanced Technology for Digital Libraries*, number 2163 in LNCS. Springer, 2001.

[12] J.-D. Zucker. Semantic abstraction for concept representation and learning. In R. S. Michalski and L. Saitta, editors, *Proceedings of the 4th International Workshop on Multistrategy Learning*, Desenzano del Garda, Italy, 1998.

VideoCube: A Novel Tool for Video Mining and Classification[*]

Jia-Yu Pan and Christos Faloutsos

Computer Science Department, Carnegie Mellon University
Pittsburgh PA 15213, USA

Abstract. We propose a new tool to classify a video clip into one of n given classes (e.g., ”news”, ”commercials”, etc). The first novelty of our approach is a method to *automatically* derive a “vocabulary” from each class of video clips, using the powerful method of “Independent Component Analysis” (ICA). Second, the method is *unified* in that it works with both video and audio information, and gives vocabulary describing not only the still images, but also motion and the audio part. Furthermore, this vocabulary is *natural* in that it is closely related to human perceptual processing. More specifically, every class of video clips gives a list of “basis functions”, which can compress its members very well.
Once we represent video clips in “vocabularies”, we can do classification and pattern discovery. For the classification of a video clip, we propose using compression: we test which of the “vocabularies” can compress the video clip best, and we assign it to the corresponding class.
For data mining, we inspect the basis functions of each video genre class and identify genre characteristics such as fast motions/transitions, more harmonic audio, etc. In experiments on real data of 62 news and 43 commercial clips, our method achieved overall accuracy of ≈81%.

1 Introduction

Video classification is useful for organizing and segmenting video clips of digital video libraries [23]. It also facilitates browsing and content-based retrieval.

The process of classifying video clips into several predefined genre classes (such as news and commercials) usually involves two steps: first, building a model of each genre class from training video clips of that class; second, classifying video clips of unknown genre by comparing them to class models. The design decisions in the first step includes: *How do we represent a video clip? What kind of features are used?* At the second step, we have to decide: *how do we determine which class a video clip belongs to? What kind of similarity function is used to determine the closeness between a video clip and a class?*

In this paper, we propose a novel video classification method with the following characteristics:

[*] This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IRI-9817496, IIS-9988876, IIS-0083148, IIS-0113089, IIS-0209107 and by the Defense Advanced Research Projects Agency under Contract No. N66001-00-1-8936.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 194–205, 2002.

- Feature extraction is done *automatically*, as opposed to handpicking the features.
- The method is *unified* in that it deals with both visual and auditory information, and captures both spatial and temporal characteristics.
- The extracted features are "*natural*", in the sense that they are closely related to human perceptual processing.

This paper is organized as follows. In section 2, we give a brief survey on the previous work on video classification. In section 3, we introduce Independent Component Analysis (ICA) and its relationship to human perceptual processing. In section 4, we describe our proposed method for video classification. Section 5 describes our experiments and discusses the results. We conclude in section 6.

2 Survey of Previous Work

There are studies based on pixel-domain visual information (color histograms) [14], and transform (compressed) domain visual information [6]. Other kinds of meta-data, such as motion vectors [19] and faces [4], have also been used. On the other hand, time-domain and frequency-domain auditory features have also been used in classifying video genres [15,18]. Recent studies [5] combined visual and auditory features for video classification. However, these features are usually hand-picked, whose applicability relies on the experience of the researchers, and their understanding of the deploying domain. Another issue of feature selection is whether the features are global or local [20]. With increased complexity, the latter provides class models that are more accurate, with higher representative power [24].

Several approaches have been examined for the representation of genre classes, namely, statistical (Gaussian) modeling [18] and hidden Markov model [4]. There are also works representing class models using general classifiers, such as decision trees or neural networks.

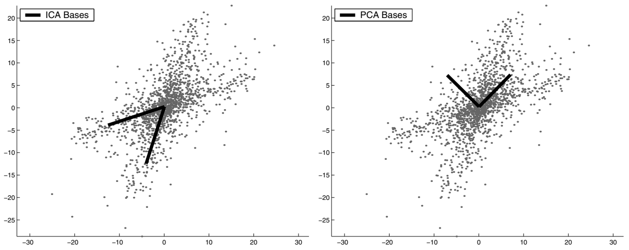
Below are three studies on video classification:

- Roach'01 [18] used audio features on 5 classes: sports, cartoon, news, commercials and music. They achieved ≈76% accuracy on classification.
- Truong'00 [21] used visual statics, dynamics, and editing features on 5 classes: sports, cartoon, news, commercials and music videos. They reported an accuracy of ≈80%.
- Liu'98 [15] used audio features on 3 classes: commercials, report (news and weather) and sports. Their classification accuracy is 93%.

Because of the different sets of genre classes and the different collections of video clips these three studies used, it is difficult to compare the performance of their different approaches and features.

3 Independent Component Analysis

Independent Component Analysis (ICA) [9,10], like principal component analysis (PCA), has been proven a useful tool for finding structure in data. Both techniques represent the given multivariate data set by a linear coordinate system. Unlike PCA, which gives orthogonal coordinates (or bases) and is based on the second-order statistics (covariance) of the data, ICA is more generalized and gives non-orthogonal bases determined by the second- and higher-order statistics of the data set [12]. Figure 1 demonstrates the difference between ICA and PCA. In this case, ICA captures the non-orthogonal underlying components of the data set which PCA misses.



(a) ICA bases (b) PCA bases

Fig. 1. ICA bases and PCA bases. PCA fails to capture the real components of the data set, while ICA does.

Specifically, let an observed data point be $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ a zero-mean m-dimensional random vector. Then under the assumption of ICA,

$$\mathbf{x} = \mathbf{A}\mathbf{s},$$

where $\mathbf{A}$ is a m-by-n matrix (usually $n < m$), and $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$ is a n-dimensional vector of source variables. s_i's are called the *independent components*. $\mathbf{A}$ is called the *mixing matrix* and its columns are the *bases* that are used to construct the observed $\mathbf{x}$. Given a set of data points $\mathbf{x}$'s, ICA finds the matrix $\mathbf{A}$ and the vector $\mathbf{s}$ which satisfies the above equation, subject to some optimization constraints on $\mathbf{s}$ such as maximizing non-gaussianity (or kurtosis, or independence of s_i's) [10].

3.1 ICA and Human Perceptual Processing

The mechanism of how the human perceptual system represents things we see and hear has long been intriguing. Barlow [1] proposed that neurons perform redundancy reduction and make up a factorial code for the input data, i.e. a representation with independent components, which is supported by recent studies on efficient natural image encoding, either from the direction of sparse coding [17] (maximizing redundancy reduction), or from the direction of independent components [3]. These experimental results show that human perceptual processing is based on *independent features which encode the input signals efficiently*. The independent components, either of visual or auditory signals, are generally filters resembling wavelet (Gabor) filters, which are oriented, localized in space (or time), band-pass in the frequency domain [3,13], and resembles the receptive fields of neurons in mammals' cortex.

Analysis has also been extended to the spatial-temporal domain [22], where the independent components of natural image sequences (video clips) and color images [7] are examined. The results are qualitatively similar to those of the static and gray-scale images, and are again closely related to the results from human perceptual studies.

Due to the fact that human perceptual processing is based on independent components of signals which ICA is able to compute, ICA has been used in applications such as face recognition [2], object modeling [24] and speech recognition [11], and achieved better or comparable performance than conventional approaches based on hand-picked features.

4 Proposed Method

Our proposed method includes a unified, automatic feature extraction method and the classification process based on these features. We will first explain how the features are extracted, how effective they are in representing the video genres, and how they are used in classification.

4.1 Features

We want to extract visual and auditory features which capture both spatial and temporal characteristics of a video genre. We first describe our extraction method for visual features. Extracting auditory features is similar.

Visual features are derived from pixel information. We say two pixels are **spatially adjacent** if they are adjacent to each other on the same video frame, and two pixels are **temporally adjacent** if they are at the same position of adjacent frames. For example, pixels (2,1) and (2,2) on frame 1 are spatially adjacent, and pixel (2,1) on frame 1 and pixel (2,1) on frame 2 are temporally adjacent. To consider the spatial and temporal information at once, spatially and temporally adjacent pixels of a video clip are grouped into cubes as basic processing units.

Definition 1. ***(VideoCube)*** *The pixel located at position (x,y) on frame t is denoted as p(x,y,t). A* ***n-by-n-by-n cube*** *located at p(x,y,t) consists of all pixels p(i,j,k) where i=x,...,(x+n-1), and j=y,...,(y+n-1), and k=t,...,(t+n-1). We called such cubes* ***VideoCubes****, which incorporate both spatial and temporal pixel information of the video frames.*

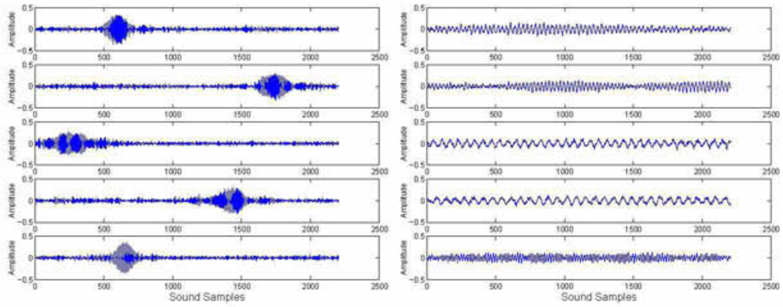
Definition 2. *(VideoBasis)* ***VideoBases*** *of a genre class are the spatial-temporal features extracted by performing ICA on a set of n-by-n-by-n VideoCubes sampled randomly from the training video clips. They are effectively the basis functions obtained from ICA. VideoBases of a genre class can be considered as the* ***vocabulary*** *commonly used by the clips of that genre to describe the visual content. Clips of the news genre share a set of VideoBases which is different from the set shared by commercial clips.*

Figure 5 shows several VideoBases of news stories and commercials. These basis functions are similar to the moving Gabor filters [22]. VideoBases of commercials have greater chopping effects along the time axis, which is because that more activities happen during commercials.

In a similar fashion, the auditory vocabulary of a video genre is extracted from the audio tracks of video clips.

Definition 3. *(AudioBasis)* ***AudioBases*** *of a genre class are the auditory features extracted by performing ICA on a set of audio segments of duration d seconds. Each segment is randomly sampled from the training clips. In the following, when both AudioBasis and VideoBasis are mentioned together, they are called* ***AV-Bases****.*

Figure 2 shows some of the AudioBases of news and commercial clips. The AudioBases of the two genres reveal their different characteristics, suggesting a way of classifying clips of the two genres.



(a) AudioBases (news) (b) AudioBases (commercials)

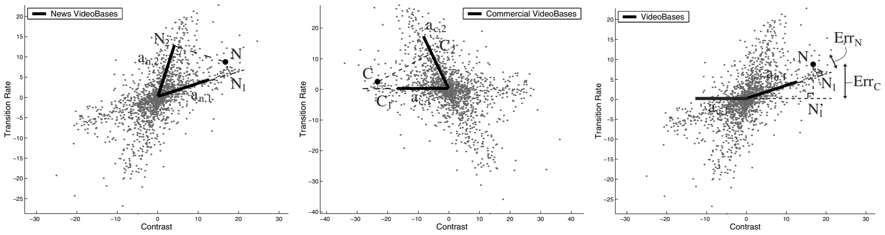
Fig. 2. AudioBases. AudioBases reveal the characteristic differences between news and commercials.

ICA provides a way to automatically extract features which, as shown later, capture the essentials of a genre. As opposed to other feature extraction studies [5,16], our approach eliminates human involvement and the trial-and-error cycles in finding features with good expressive power.

4.2 Classification

Our classification method is based on the idea of compression, that is, the set of AV-Bases which compresses a video clip most efficiently is that of the genre class to which the clip belongs. To measure the "goodness" of compression, we compare the reconstruction errors of a video clip when it is encoded by the AV-Bases of different genre classes. The set of AV-Bases which gives the smallest reconstruction error is the one that compresses the clip best, and the clip is classified to the genre associated with that set. The existence of the reconstruction error comes from the fact that we are using fewer AV-Bases than the dimension of the data samples, and therefore cannot fully reconstruct the content of a clip. This idea is similar to the vector space model in traditional information retrieval, where the representing vocabularies are words and terms. In our case, the AV-Bases act as the vocabulary.

Figure 3 illustrates the idea of our classification approach. Two fictitious data sets representing VideoCubes (samples of visual content) from news and commercial clips are shown, where each point is a VideoCube. Every VideoCube can be represented as a linear combination of VideoBases.



(a) News (b) Commercial (c) Classification

Fig. 3. Representation of news and commercial clips. **N** is a VideoCube of a news clip and **C** of a commercial clip. $\mathbf{a_{N_1}}$ and $\mathbf{a_{N_2}}$ are VideoBases of news, and $\mathbf{a_{C_1}}$ and $\mathbf{a_{C_2}}$ are those of commercials. (a) $\mathbf{N} = \mathbf{N_1} + \mathbf{N_2} = s_{N_1}\mathbf{a_{N_1}} + s_{N_2}\mathbf{a_{N_2}}$ and (b) $\mathbf{C} = \mathbf{C_1} + \mathbf{C_2} = s_{C_1}\mathbf{a_{C_1}} + s_{C_2}\mathbf{a_{C_2}}$. ($s_{N_i}$'s and s_{C_i}'s, $i = 1, 2$, are weighting scalars.) (c) Only one basis from each genre ($\mathbf{a_{N_1}}$ and $\mathbf{a_{C_1}}$) is kept. VideoCube N is best represented by $\mathbf{a_{N_1}}$, that is, the reconstruction error Err_N is less than that when **N** is represented by $\mathbf{a_{C_1}}$ (Err_C). Therefore, **N** is classified as news.

Figure 4 gives the classification algorithm (**VCube**) using VideoBases. Using AudioBases to classify clips is similar, with the processing units changed from VideoCubes to the audio segments of duration d seconds (e.g., d=0.5).

4.3 Capture Local Activities

A video clip with many local activities will be mistakenly considered to have activities throughout the scene, if local scene changes are not quantified as local.

```
Input: Video track of a clip; VideoBases of $G$ genre classes
Output: Genre class of the clip

1. Pixels in I-frames are collected and stacked, and non-overlapped n-by-n-by-n VideoCubes are formed from these collected pixels. (e.g., n=12)
2. Initialize the sum of reconstruction error of class $c_i$, $sumErr_i = 0$, $i = 1, \dots, G$.
3. For each VideoCubes ($vc$),
    3.1 Encode $vc$ with the VideoBases of each genre class ($c_i, i = 1, \dots, G$).
    3.2 Compute the reconstruction error ($err_i, i = 1, \dots, G$).
    3.3 $sumErr_i = sumErr_i + err_i$
4. Return $c_k$, where $k = \underset{i}{argmin}\ sumErr_i$.
```

Fig. 4. Classification algorithm (VCube)

To better capture local activities, we model local scene changes separately, rather than summing all activities in the video frames as a global effect.

We divide each video frame into 9 rectangular regions of equal size (in 3-by-3 arrangement) and extract the VideoBases for each region separately. In other words, each genre now has 9 sets of VideoBases, one for each region. For news clips, activities happen mostly in certain regions, while activities are spread over the whole frame for commercial clips.

The classification algorithm **VCube_P** ("P" stands for "Position") which considers local effects at different positions is similar to algorithm **VCube** (Figure 4). The only difference is step 3.1, which is modified as

```
    3.1 Let $r$ be the region where $vc$ is, $1 \leq r \leq 9$. Encode $vc$ with the VideoBases ($VB_{i,r}$), for every genre class ($c_i$, $i = 1, \dots, G$).
```

5 Experimental Results

In this section, we describe our experiments on classifying news and commercial video clips. Specifically, we discuss the properties of the VideoBases and the AudioBases of the two classes and report our classification result.

5.1 Data Set

We divide our collection of video clips into the training set and the testing set. The training set contains 8 news clips and 6 commercial clips, each about 30 seconds long. The testing set contains 62 news clips and 43 commercial clips, each about 18 seconds long. The training set is used for constructing the VideoBases and the AudioBases of the two classes. We use the FastICA [8] package from the Helsinki University of Technology for ICA processing.

Video frames are divided into 9 rectangular regions of equal size, in a 3-by-3 matrix-like arrangement. As a result, each genre class has 9 sets of VideoBases (one for each region) and 1 set of AudioBases. VideoCubes of size 12x12x12 are sampled randomly from the training clips, and those located at the same region are used to compute the VideoBases of that region. We keep only the red channel of the color information (red, green, blue channels) at each pixel. In our experience, VideoBases based on any of the three channels give us similar result.

AudioBases are derived from 0.5-second long audio segments sampled randomly from the training clips. The audio segments are down-sampled by a factor of 10 as a trade-off between the data size and coverage period. That is, we want auditory features of longer segments to better capture auditory characteristics, but under the encoding frequency of 44.1kHz, a 0.5-second long audio segment without down-sampling has too much data and will hinder the subsequent computation.

The number of video clips in the training set is not an important factor in the quality of the derived ICA bases. What really matters is the amount of data samples used for extracting features. In our training phase, 10,000 12x12x12 VideoCubes and 7,000 0.5-second audio segments are used, and 160 VideoBases (per region) and 60 AudioBases are extracted for each genre class.

5.2 Rule Discovery

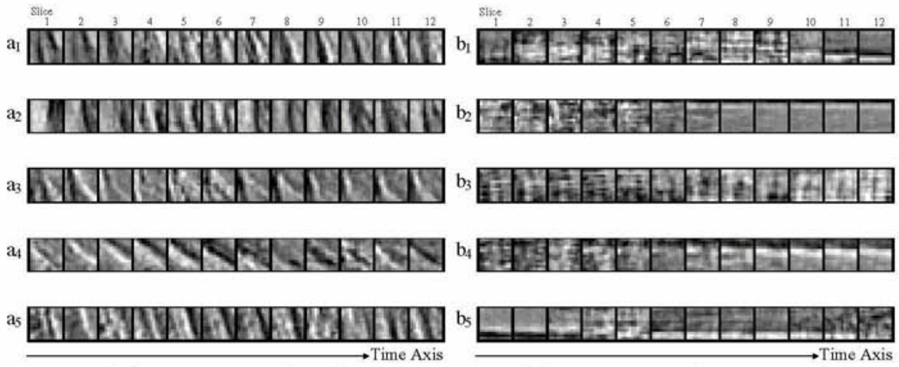
Figures 5 and 6 show several VideoBases and AudioBases. VideoCubes are 12x12x12 cubes which consist of pixels in both spatial and temporal dimensions. Therefore, VideoBases can be viewed as spatial-temporal filters (stacks of spatial filters at sequential time steps) of the same size (12x12x12) which capture spatial-temporal characteristics at once. In the following, we call the spatial filter of a VideoBasis at each time point a **slice**. In Figure 5, each row is a VideoBasis and its slices are arranged in their temporal order, from left to right.

The VideoBases of the two genre classes, news and commercials, reflect the major differences between these two classes. For example, in Figure 5(a), the $\mathbf{a_4}$ VideoBasis of news shows a (white) -45^o edge moving first from bottom-left to top-right (slices 1 to 4) and then reverse (slices 5 to 8) and then reverse again (slices 9 to 12). In Figure 5(b), the $\mathbf{b_4}$ VideoBasis of commercials shows a big transition between slices 6 and 7, while slices 1 to 6 have some random patterns and a clear edge pattern is moving downward in slices 7 to 12 .

Observation 1 ***VideoBasis.*** *VideoBases of news have clearer edge-like patterns in their slices, and have few transitions as time moves on. On the other hand, VideoBases of commercials are more noisy (no clear pattern) and have greater transitions/choppings between slices. The properties that these bases exhibit generally agree with our experiences with news stories and commercials.*

Figure 6 shows the AudioBases we extracted. These bases also coincide with our common knowledge about the sounds occurring in news stories and commercials.

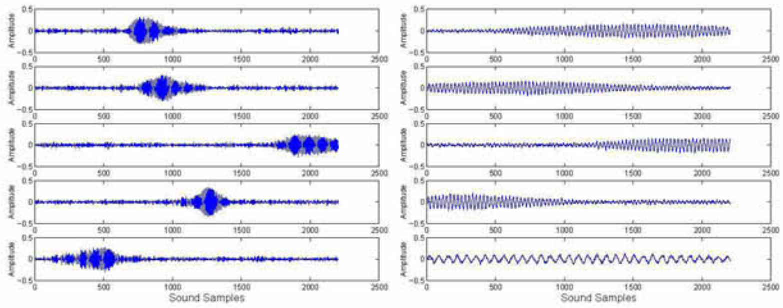
Observation 2 ***AudioBasis.*** *The AudioBases of news stories contain amplitude envelopes that are localized in time. The waveforms of these bases are intermediate between those of pure harmonic sounds and pure non-harmonic sounds,*



(a) VideoBases (news) (b) VideoBases (commercial)

Fig. 5. VideoBases. VideoBases shown here are from the central region of the 9 regions. Slices of a VideoBasis are arranged in a row from left to right in their temporal order. (a) Slices of news have clearer Gabor edge detector pattern moving slightly along the time axis. (b) Patterns within commercial VideoBases are not as clear, but the changes along the time axis are more significant.

and resemble those of human speech (mix of harmonic vowels and non-harmonic consonants) [13]. This agrees with our knowledge that the most frequent sound in news stories is human speech. As for the AudioBases of commercials, the waveforms are more harmonic and are similar to those of animal vocalizations and natural sounds [13]. This is due to the fact that in commercials, music is more dominant than speech.



(a) AudioBases (news) (b) AudioBases (commercial)

Fig. 6. AudioBases. AudioBases capture the sound characteristics of different video genres. (a) Bases for news resemble waveforms of human speech (mix of harmonic and non-harmonic sounds). (b) Bases for commercials resemble waveforms of natural sounds and animal vocalization (harmonic sounds).

Observation 3 ***Cross-media rule.*** *From observations 1 and 2, we find that*

- *a static scene usually corresponds to the human voice, and*
- *a dynamic scene usually corresponds to natural sounds (music, animal vocalization).*

5.3 Classification

The VideoBases and AudioBases give efficient coding of a video genre, due to the sparse coding property of ICA. In other words, a video clip is encoded most efficiently when it is encoded by the AV-Bases of its genre. We classify a video clip to a genre class whose AV-Bases give the smallest reconstruction error (Figure 3(c)).

In this study, VideoBases and AudioBases are used separately for classification, that is, we do two kinds of classification: one based on VideoBases and the other based on AudioBases. Table 1 lists our results on classifying a testing set of 62 news and 43 commercial clips.

Table 1. Results of classification using VideoBases and AudioBases

Class	Total	VideoBases			AudioBases		
		News	Commercial	Accuracy	News	Commercial	Accuracy
News	62	45	17	0.726	46	16	0.742
Commercial	43	3	40	0.930	4	39	0.907

In our experiments, classification using either VideoBases or AudioBases yielded similar results. Commercial clips are classified with higher accuracy, while news is classified less accurately. This is because news stories contain field reports. These reports have (a) more (background) motions and faster transitions which could confuse the classification process using VideoBases; (b) background sounds along with speech sounds, which confuse classification using Audiobases. Overall, we achieved classification accuracy of around 81%, which is comparable to previous studies noted in Section 2.

6 Conclusion

In this paper, we proposed VideoCube, a novel method for video classification. Its contributions are:

1. A **unified** approach to incorporate both spatial and temporal information, which works on both video and audio information.
2. An **automatic** feature extraction method based on *independent component analysis (ICA).*

3. The extracted features (*VideoBases* and *AudioBases*) are **natural** in that they are closely related to those used in human perceptual processing.

VideoBases and AudioBases successfully captured the major characteristics of video content. They found the following patterns:

- **News :** In news reports, fewer activities (motions) and editing transitions are present, and the major sound is human speech.
- **Commercials :** In commercials, more activities and fast editing transitions are present, and the major sounds are natural sounds, music and animal vocalizations.
- **Cross-media rule :** Static scenes correspond to human speech and dynamic scenes correspond to natural sounds.

Our experiments on classifying 62 news and 43 commercial clips using either VideoBases or AudioBases achieved good classification accuracy: around 73% for news and 92% for commercials. The overall accuracy is around 81%, which is comparable to previous studies.

References

1. Horace B. Barlow. Unsupervised learning. *Neural Computation*, (1):295–311, 1989.
2. Marian Stewart Bartlett, H. Martin Lades, and Terrence J. Sejnowski. Independent component representations for face recognition. *Proceedings of SPIE; Conference on Human Vision and Electronic Imaging III*, January 1998.
3. Anthony J. Bell and Terrence J. Sejnowski. The “independent components” of natural scenes are edge filters. *Vision Research*, (37):3327–3338, 1997.
4. Nevenka Dimitrova, Lalitha Agnihotri, and Gang Wei. Video classification based on hmm using text and faces. *ACM Multimedia*, 2000.
5. Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. Automatic recognition of film genres. *The 3rd ACM International Multimedia Conference and Exhibition*, 1995.
6. Andreas Girgensohn and Jonathan Foote. Video classification using transform coefficients. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 6.
7. Patrik O. Hoyer and Aapo Hyvarinen. A probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. of Am. A: Optics, Image Science, and Vision*, March 1999.
8. Aapo Hyvarinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 1999.
9. Aapo Hyvarinen. Survey on independent component analysis. *Neural Computing Surveys*, 2:94–128, 1999.
10. Aapo Hyvarinen. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.
11. Jong-Hwan Lee, Ho-Young Jung, Te-Won Lee, and Soo-Young Lee. Speech feature extraction using independent component analysis. *International Conference on Acoustics, Speech, and Signal Processing*, in press, June 2000.
12. Te-Won Lee, Mark Girolami, Anthony J. Bell, and Terrence J. Sejnowski. A unifying information-theoretic framework for independent component analysis. *International Journal on Mathematical and Computer Models*, in press, 1999.

13. Michael S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, April 2002.
14. Rainer Lienhart, Christoph Kuhmunch, and Wolfgang Effelsberg. On the detection and recognition of television commercials. *Proceedings of the International Conference on Multimedia Computing and Systems*, pages 509–516, 1996.
15. Zhu Liu, Jincheng Huang, and Yao Wang. Classification of tv programs based on audio information using hidden markov model. *Proc. of 1998 IEEE Second Workshop on Multimedia Signal Processing (MMSP'98)*, pages 27–31, December 1998.
16. Zhu Liu, Yao Wang, and Tsuhan Chen. Audio feature extraction and analysis for scene classification. *Journal of VLSI Signal Processing*, Special issue on multimedia signal processing:61–79, October 1998.
17. Bruno A. Olshausen and David J. Field. Wavelet-like receptive fields emerge from a network that learns sparse codes for natural images. *Nature*, (381):607–609, 1996.
18. Matthew J. Roach and John S. Mason. Video genre classification using audio. *EuroSpeech*, 2001.
19. Matthew J. Roach, John S. Mason, and Mark Pawlewski. Video genre classification using dynamics. *Int. Conf. on Acoustics, Speech and Signal Processing*, 2001.
20. Kim Shearer, Chitra Dorai, and Svetha Venkatesh. Local color analysis for scene break detection applied to tv commercials recognition. *Proc. 3rd. Intl. Conf. on Visual Information and Information Systems (VISUAL'99)*, pages 237–244, June 1999.
21. Ba Tu Truong, Svetha Venkatesh, and Chitra Dorai. Automatic genre identification for content-based video categorization. *International Conference Pattern Recognition*, 4:230–233, 2000.
22. J. H. van Hateren and D. L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. Royal Society Lond. B*, (265):2315–2320, 1998.
23. Howard Wactlar, Michael Christel, Y. Gong, and A. Hauptmann. Lessons learned from the creation and deployment of a terabyte digital video library. *IEEE Computer*, 32(2):66–73, February 1999.
24. Xiang Sean Zhou, Baback Moghaddam, and Thomas S. Huang. Ica-based probabilistic local appearance models. *IEEE International Conference on Image Processing (ICIP)*, October 2001.

Developing Tsinghua University Architecture Digital Library for Chinese Architecture Study and University Education[*]

Chunxiao Xing[1], Lizhu Zhou[1], Zhiqiang Zhang[2], Chun Zeng [2], and Xiaodan Zhou

Department of Computer Science & Technology Tsinghua University,
Beijing, China 100084
[1] {xingcx,dcszlz}@mail.tsinghua.edu.cn
[2]{Zhangzhiqiang,Bobofu00}@mails.tsinghua.edu.cn

Abstract. The Tsinghua University Architecture Digital Library (THADL) has been developed as a prototype system with dual goals - to study the key techniques of digital libraries (DLs), and to provide rich, valuable resources for Chinese architecture research and education as well as a high level of service. This paper focuses on the following issues: 1) the novel architecture of THADL and its implementation; 2) the metadata specification--THADL Metadata 2.0; 3) the OAI PMH v2.0 implementation to make THADL interoperable with other digital resources; 4) the Web-based courseware environment for students majoring in architecture science; and 5) the personalized active service subsystem -- TH-PASS.

1 Introduction

With the rapid advancement of Internet technology and the knowledge economy, Digital Libraries have emerged as large scale, distributed information and knowledge centers to bring together collections, services, and people in support of the full life cycle of creation, dissemination, use, storage, and preservation, much as a digitized collection with information management tools [1]. Many advanced countries and research communities consider digital libraries to be the flagship research effort for their National Information Infrastructure (NII) – e.g. Digital Libraries Initiative (DLI-1, 2), British Library's Digital Library System (DLS), GALLICA 2000, The Canadian Initiative on Digital Libraries (CIDL), and German GLOBAL INFO. In university education, Digital Libraries play an important role to provide a learning environment for everyone, anywhere, anytime. A wealth of prior works, such as NSDL, NEEDS, IMS, LOM, and Earthscape [1-5], have addressed the research and development of education environments based on digital library.

[*] Supported by the National Natural Science Foundation of China under Grant No. 60221120146; the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 206-217, 2002.

As a comprehensive and national key university having disciplines in the sciences, engineering, management and social sciences, Tsinghua University is developing digital libraries to enhance its information infrastructure and education environment. The THADL project, which started formally in March 2000, maintains a balance between technology-focused research and content-based research. The project involves substantial cooperation among computer researchers, librarians, and subject specialists. The goals of THADL are as follows: 1) Exploring efficient methods and technologies for constructing future large-scale digital libraries by designing and evaluating the THADL prototype. 2) Building a Chinese architecture digital library that provides an intelligent, interactive, and collaborative learning environment on the Internet. 3) Presenting an efficient method to digitalize, index, and preserve most kinds of materials for Chinese architecture study. 4) Establishing metadata specifications and standards for Chinese architecture science. 5) Supporting friendly, active, and personalized services for different users including students, scholars, librarians, and ordinary users on the Internet. 6) Developing a simple, high-efficiency, light-weight interoperable protocol for exchanging, sharing, and integrating other digital library resources.

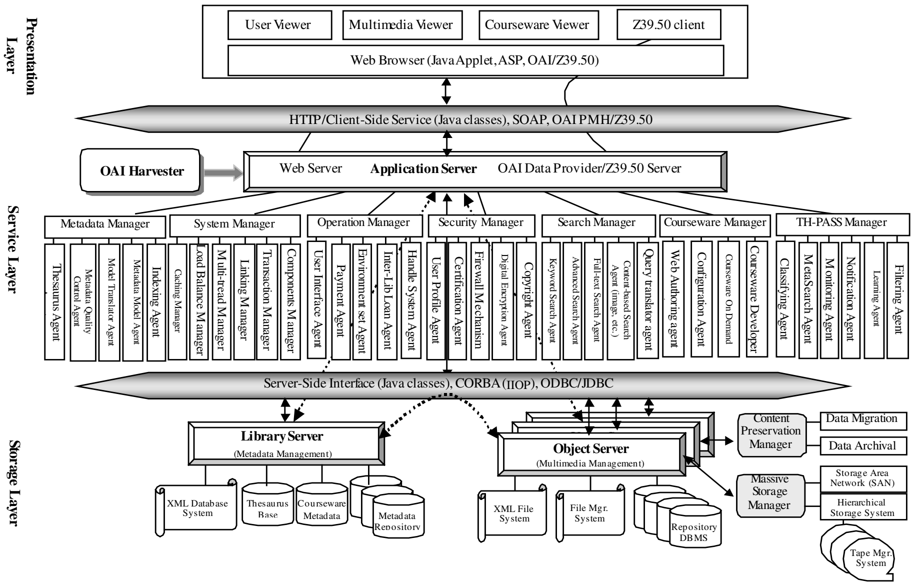


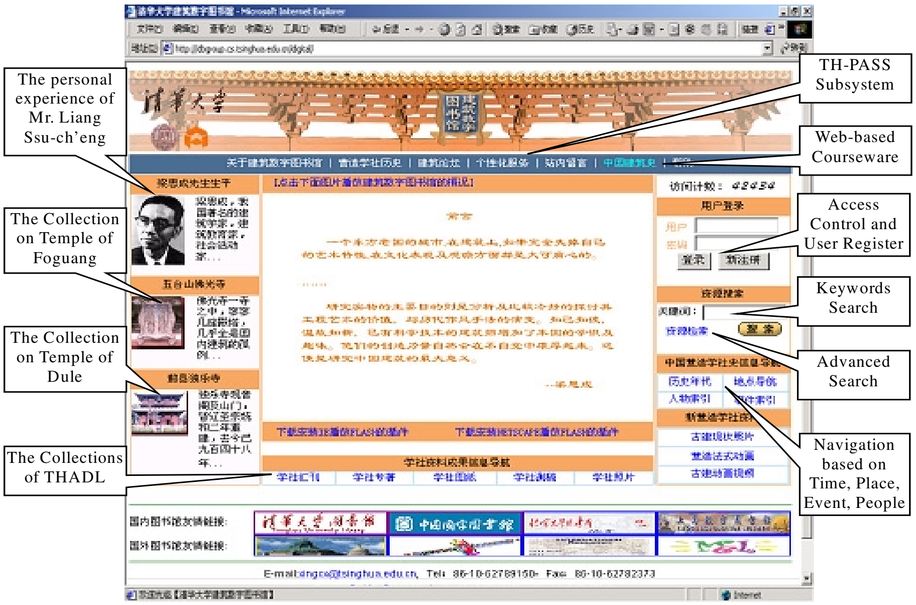
Fig. 1. THADL Architecture

The organization of this paper is as follows. Section 2 explains the design of the architecture of the THADL prototype system and its basic functions and services. Section 3 discusses the metadata specification and a related metadata editor based on Protégé 2000. Section 4 describes the implementation of THADL's interoperable protocol based on OAI PMH v2.0. Section 5 introduces THADL's Web-based learning resources and environment. Section 6 presents a personalized active service subsystem: TH-PASS, and Section 7 concludes and discusses future work.

2 Design of THADL Architecture

THADL's architecture consists of a presentation layer, a service layer, and a storage layer. The architecture is based on a triangular client/server model comprised of one Library server, one Application server, and one or more Object servers (Fig. 1). It evolved from THADL 1.0 [7], which complies with the reference model for an Open Archival Information System (OAIS) [8].

The **Presentation layer,** based on a Web browser (see Fig. 2), consists of a user viewer, a multimedia viewer such as video and e-book, and a courseware viewer for Chinese ancient architecture materials for distance learning.



The personal experience of Mr. Liang Ssu-ch'eng
The Collection on Temple of Foguang
The Collection on Temple of Dule
The Collections of THADL
TH-PASS Subsystem
Web-based Courseware
Access Control and User Register
Keywords Search
Advanced Search
Navigation based on Time, Place, Event, People

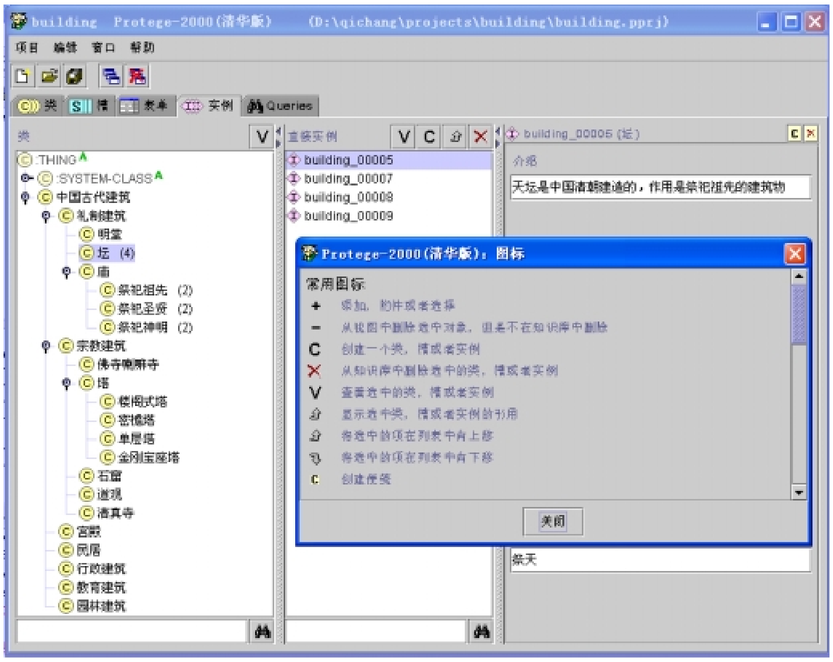
Fig. 2. THADL Homepage (http://dbgroup.cs.tsinghua.edu.cn/digital)

The **Service layer** is composed of major management and service components in the Application Server. These components include the metadata manager, the system manager, the operation manager, the security manager, the search manager, the courseware manager, and the TH-PASS (Tsinghua Personalized Active Service subsystem) Manager. To support the system's scalability, the functions of THADL are designed as distributed services linked by CORBA/IIOP objects. With regard to interoperability, we have developed a lightweight interoperable protocol based on the OAI PMH v2.0 (the Open Archives Initiative Protocol for Metadata Harvesting) for cross-library information discovery and retrieval. The services are performed by intelligent agents that can collaborate with each other (Fig. 1).

The **Storage layer** consists of the Library Server and the Object Server. The Library server is responsible for managing all the materials metadata, Courseware metadata, and thesaurus through the relational database management system (RDBMS) and the native XML database--Tamino. The object server manages all kinds of digital objects and repositories. The content preservation manager supports data migration and data archiving.

3 Establishment of THADL Metadata

The metadata framework of THADL Metadata 2.0 is an extension and improvement of THADL Metadata 1.0 [7], which aims to make valuable art treasures of Chinese ancient architecture accessible to users all over the world. We designed THADL Metadata 2.0 DTD based on XML/RDF and Dublin Core [9]. In the design of THADL Metadata, we take the following aspects into consideration: functions of metadata; fundamentals for designing metadata standard; workflow for designing metadata standard; content structure and the elements; semantic definition rules and related authorities; meta-language and metadata editor.



天坛是中国清朝建造的，作用是祭祀祖先的建筑物

+ 添加，创件或者选择

− 从视图中删除选中对象，但是不在知识库中删除

C 创建一个类，槽或者实例

× 从知识库中删除选中的类，槽或者实例

V 查看选中的类，槽或者实例

显示选中类，槽或者实例的引用

将选中的项在列表中向上移

将选中的项在列表中向下移

C 创建便笺

祭天

Fig. 3. Metadata Editor Based on Protégé-2000

3.1 Metadata Editor Based on Protégé 2000

The materials in the THADL repositories cover papers, journals, photographs, manuscripts, drawings/blueprints, animation, video, and audio on Chinese ancient architecture. The concepts and their relations on Chinese architecture are complex. Based on Protégé-2000 [12], we have developed a metadata editor for editing and organizing metadata on Chinese architecture. The metadata editor can help users to model domain-specific (such as Chinese architecture) knowledge and browse the contents of knowledge bases such as classes, instances of classes, slots representing attributes of classes and instances, and facets expressing additional information about slots. RDFS (Resource Description Framework Schema) is the representation language to express the concepts and their relations of the contents. The editor supports both Chinese and English. Fig. 3 is an example showing how to use the editor to build metadata.

4 OAI-PMH v2.0 Based Interoperable Protocol

OAI has a significant place in solving the digital library interoperability problem. By lowering the barrier for data providers to expose metadata, the technical framework of the OAI PMH (Metadata Harvesting Protocol) provides a uniform, highly efficient, and low-barrier approach for interoperability solutions. It has successfully positioned itself between high-functionality, complex federation schemes (such as Z39.50) and low-functionality, web crawled search services [13].

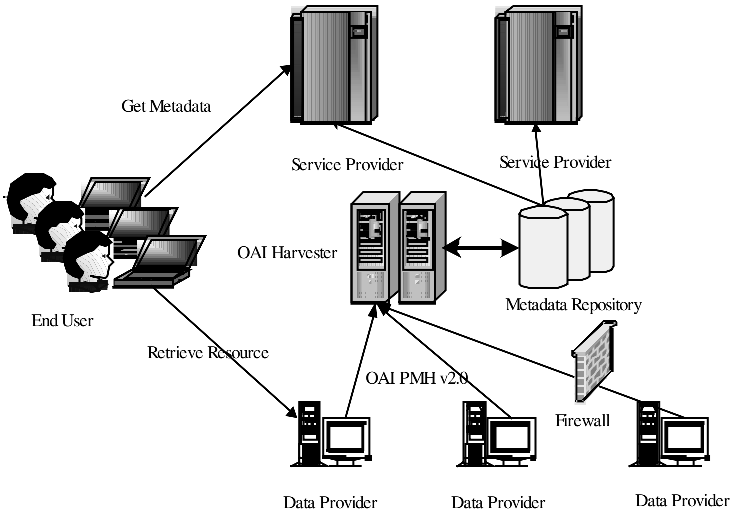
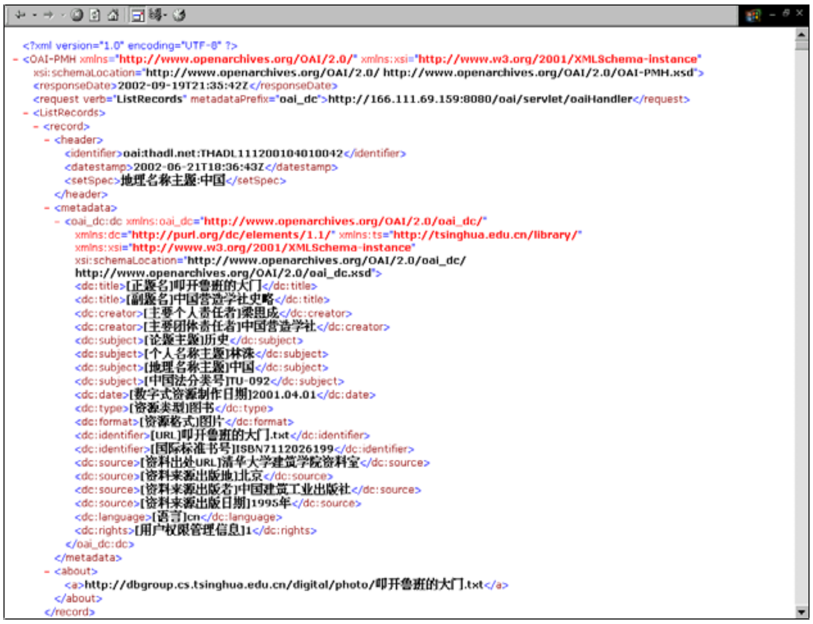


Fig. 4. The Framework of Interoperable Protocol Based on OAI-PMH v2.0 in THADL

In June 2002, the newest version of OAI-PMH v2.0 was released. There are two classes of participants in the OAI-PMH framework: Data Providers administer systems that support the OAI-PMH as a means of exposing metadata through an HTTP/XML based protocol; and Service Providers use metadata harvested via the OAI-PMH as a basis for building value-added services. A harvester is a client application that issues OAI-PMH requests, which is operated by a service provider as a means of collecting metadata from repositories (see Fig. 4).

THADL has a complete implementation of OAI-PMH v2.0. We use Dublin Core (DC) as the default metadata set, adopt Java and XML technology to implement the components of the OAI-PMH framework. It consists of six requests or verbs: *GetRecord, Identify, ListMetadataFormats, ListIdentifier, ListRecords, ListSets*. Fig. 5 shows the demo of the *ListRecords* request based on THADL Metadata 2.0. Our current implementation supports the relational database (Oracle), and the native XML database (Tamino) for metadata management.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2002-09-19T21:35:42Z</responseDate>
   <request verb="ListRecords" metadataPrefix="oai_dc">http://166.111.69.159:8080/oai/servlet/oaiHandler</request>
 - <ListRecords>
   - <record>
     - <header>
         <identifier>oai:thadl.net:THADL111200104010042</identifier>
         <datestamp>2002-06-21T18:36:43Z</datestamp>
         <setSpec>地理名称主题:中国</setSpec>
       </header>
     - <metadata>
       - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
           xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:ts="http://tsinghua.edu.cn/library/"
           xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
           http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
           <dc:title>[正题名]叩开鲁班的大门</dc:title>
           <dc:title>[副题名]中国营造学社史略</dc:title>
           <dc:creator>[主要个人责任者]梁思成</dc:creator>
           <dc:creator>[主要团体责任者]中国营造学社</dc:creator>
           <dc:subject>[论题主题]历史</dc:subject>
           <dc:subject>[个人名称主题]林洙</dc:subject>
           <dc:subject>[地理名称主题]中国</dc:subject>
           <dc:subject>[中国法分类号]TU-092</dc:subject>
           <dc:date>[数字式资源制作日期]2001.04.01</dc:date>
           <dc:type>[资源类型]图书</dc:type>
           <dc:format>[资源格式]图片</dc:format>
           <dc:identifier>[URL]叩开鲁班的大门.txt</dc:identifier>
           <dc:identifier>[国际标准书号]ISBN7112026199</dc:identifier>
           <dc:source>[资料出处URL]清华大学建筑学院资料室</dc:source>
           <dc:source>[资料来源出版地]北京</dc:source>
           <dc:source>[资料来源出版者]中国建筑工业出版社</dc:source>
           <dc:source>[资料来源出版日期]1995年</dc:source>
           <dc:language>[语言]cn</dc:language>
           <dc:rights>[用户权限管理信息]1</dc:rights>
         </oai_dc:dc>
       </metadata>
     - <about>
         <a>http://dbgroup.cs.tsinghua.edu.cn/digital/photo/叩开鲁班的大门.txt</a>
       </about>
     </record>
```

Fig. 5. OAI *ListRecords* request based on THADL Metadata 2.0

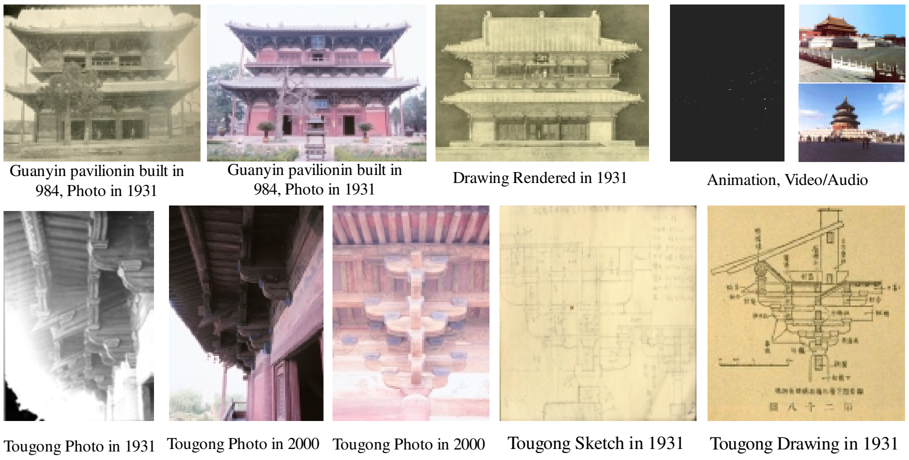
5 Web-Based Learning Resources and Environment

5.1 Digital Resources on Ancient Chinese Architecture

The major digital resources in THADL repositories include papers, journals, photographs, digital manuscripts, blueprints, drawings, page images, animation, video, and audio on ancient Chinese architecture (see Fig. 6). They are built from rare and invaluable materials of the Society for Research in Chinese Architecture (SRCA). SRCA was founded in the 1930s and was led by Zhu Qiqian, Liang Ssu-ch'eng, and Liu Dunzhen. It was the first research society on Chinese ancient architecture in China. It pioneered the scientific study of Chinese architectural development, which became the basis of Chinese architectural history. SRCA took about 15 years from 1931 to 1946 to investigate 2783 Chinese ancient buildings on sites all over China. We encourage faculty members to use web-based authoring tools (e.g. Director and Authorware) to create their courseware by making use of online digital resources of THADL.

5.2 Design of Web-Based Online Courseware

THADL provides a tool called Courseware Developer that offers shells to help subject experts plug-in their own digital materials and upload to THADL's Courseware Manager. Basic components of the web-based online interactive courseware of THADL consist of lecture materials including notes and reading references, class exercises, algorithm animations, computer programs and its run time support, and WWW links to other digital resources.



Guanyin pavilionin built in 984, Photo in 1931

Guanyin pavilionin built in 984, Photo in 1931

Drawing Rendered in 1931

Animation, Video/Audio

Tougong Photo in 1931

Tougong Photo in 2000

Tougong Photo in 2000

Tougong Sketch in 1931

Tougong Drawing in 1931

Fig. 6. THADL's Valuable Resources

6 TH-PASS: Personalized Active Service Subsystem

TH-PASS is a subsystem of THADL that provides users with a personalized filtering and notification service based on user modeling and profile learning. TH-PASS relies on many established techniques in information retrieval (IR), information filtering (IF), user modeling and machine learning, etc.

6.1 Framework of TH-PASS

In TH-PASS, the major components are developed as separate modules based on intelligent agents. The framework of TH-PASS is shown as Fig. 7. The framework consists of major components of User interface/WWW Browsers, TH-PASS Proxy Server, Information Filtering Agent, Learning/Discovering Agent, User model & User Profile, MetaSearch Engine, Metadata Extraction, Monitor Agent/Notification Agent, Notification Agent, and Distributed information Source.

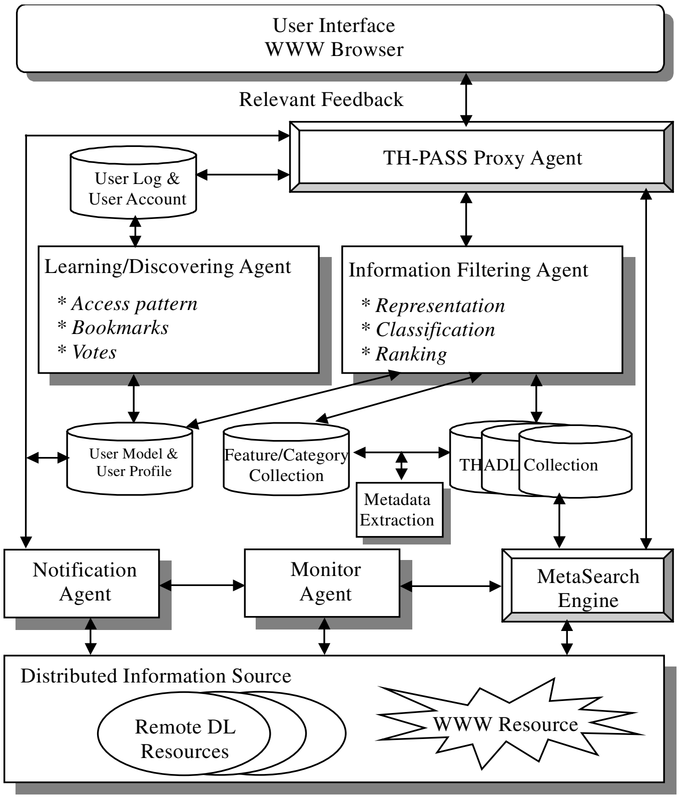


Fig. 7. Framework of TH-PASS Subsystem

6.2 Document Representation and Relevance Measure

The representation process in Information Filtering Agent involves converting the documents, requests, and user profiles to more efficiently structures so as to next relevance measure. We represent the document based on vector space model, which regards each document as a vector. The system uses term frequency inverse×document frequency (TFIDF)[10] to represent the document vectors in a collection of documents D. In this scheme the feature set F_d is a vector of word frequencies weighted by their rarity over *D*. Let *W* be the set of all words over D. In a document d, let the frequency of each word *s* be f_{ds} and let the number of documents in *D* having *s* be n_s. In document *d* let the highest term frequency be f_{dmax}. In one TFIDF scheme a word weight vector element w_{ds} is calculated [11] as:

$$w_{ds} = \frac{(0.5 + 0.5\frac{f_{ds}}{f_{d\max}})(\log\frac{|N|}{n_s})}{\sqrt{\sum_{j\in d}((0.5 + 0.5\frac{f_{dj}}{f_{d\max}})^2(\log\frac{N_D}{n_j})^2)}} \quad (1)$$

Where |*D*| is the total number of documents. F_d is the |*W*| dimensional vector of w_{ds} values. Once the feature vectors have been extracted for two documents, the distance between them may be calculated. Commonly, a dot product or Euclidean distance measure is used. The TFIDF relevance *R* between the documents *d* and candidate document *c* is a dot product of the two word vectors F_d and F_c given as:

$$R(d,c) = F_d \bullet F_c \quad (2)$$

Also we use citation relevance as another important features in scientific documents. TH-PASS uses common citations to make an estimate of document relevance. The measure that captures this idea of Relevance is called CCIDF (Common Citation Inverse Document Frequency) [11] and is partially analogous to the word vector based TFIDF. Let c_i be the frequency of a citation i in a collection of documents D, let $w_i=1/c_i$ be the inverse frequency, and let W_D be the vector of these inverse frequencies. Let c_{di} be a Boolean indicator of whether document d contains citation i, and let X_d be the resulting Boolean vector. The CCIDF Relevance between a document d and a document of interest d_i (specified by the user) is defined as:

$$R(d,d_i) = tr(X_d \times X_{d_i}) \bullet W_D \quad (3)$$

where *tr*(.) is the trace function and $\times$ is the outer product.

6.3 User Modeling and Profile Learning

In TH-PASS subsystem, the profile can represent the needs and interests of users and its role is to direct the search accordingly. We suggest that the information about a user should be processed through the way of explicit relevant feedback, profile learning and pattern discovery so that the system can establish complete and consistent user profiles. In general, user modeling is a process of discovering a user's patterns (e.g., a user's behavior pattern, knowledge pattern, cognitive pattern, etc.) based the context of the interaction. In TH-PASS this modeling is based on two judgment methods: explicit judgment models that are constructed explicitly (e.g., relevant feedback) by the user and implicit judgment models that are abstracted by the system learning and discovering on the basis of the user's behavior (e.g., access pattern and logs).

User profiles (interests) in user model contain two types of profiles for each user: cognitive and sociological. It assumes that users who share sociological parameters might also have common preferences and habits with respect to their information needs. This can be achieved by the formation of user stereotypes. Stereotypes are used in user modeling of TH-PASS to infer default beliefs about users until more accurate information is obtained. They may also serve as a shortcut for a user-model by inferring knowledge about users from their stereotypic belonging. The joining clustering algorithm can be used to partition the interviewed users to stereotypes. Based on user stereotypes, the system convert different user stereotype into 1-10 scale for filtering process.

We use stereotypes to infer knowledge about users, which is hard to infer directly from them. Specifically, we infer information usage and filtering patterns from stereotypic belonging. The users in our experiments are different user stereotypes (e.g. Professor, associate professor, lecturer, PhD student, M.S student, and etc.) in Tsinghua University.

While using TH-PASS Proxy Server, users contribute to their profiles either explicitly by manually editing the profile or implicitly by browsing the THADL Collection. Either action creates or modifies profile components, which represent users' research interests. A set of values representing features (often only a single feature or a few features) extracted from documents, which include request keywords, Hot URLs, citations, Term vectors, and citation vectors. Evidence suggests that this set is more powerful than any single representation. A user's profile consists of a set F of different types of features. In addition to a feature value, each feature f has a weight w_f corresponding to its influence.

TH-PASS adapts profiles to better represent users' interests by modifying the feature of user profile weights w_f. It does this in three ways: observing user behavior during database browsing, allowing manual adjustment, and learning from user responses to recommendations. After recommending document d_r, Learn/Discovering Agent observes the user's response and updates the weight for each feature f in profile D accordingly. The agent uses several types of user actions (e.g., Explicitly added to profile - Very high positive, Downloaded -High positive, Viewed details - Moderate positive, Ignored - Low negative, Removed from profile- ser to zero) as implicit indications of interest. The profile update rule is:

$$w_f \Leftarrow w_f + pa(b_{d_r})R_d(d, d_r) \qquad (4)$$

where $a(b_{d_r})$ denotes each user action interestingness value, and p is a learning rate.

6.4 Notification Service

TH-PASS represents a new document d_n in the THADL Collection with features corresponding to the union of the feature types in the user's profile *D*. Its Ranking Component compares d_n with those in the profile to find a level of similarity $R_D(d_n)$, which represents the document's relevance to the user. The calculation of interestingness or relevance is the weighted sum:

$$I_D(d_n) = \sum_{d\in D} w_f R_d(d,d_n) \tag{5}$$

where R_d (d, d_n) is the similarity or relatedness between features set of the user's profile and the new paper features *d**. We weight each relatedness measure by the profile feature's influence. Notification Agent recommends new documents with I_D (d_n) greater than a certain threshold. Once the profiles are created, Notification Agent periodically, or on demand, checks its database for new documents it should recommend to the user. It sends such recommendations by e-mail if the user desires so and presents them when the user logs in THADL.

7 Conclusion and Future Work

THADL is a digital library prototype system for the study of ancient Chinese architecture. The current version of THADL provides basic functions in storing digital collections, browsing of the collections, searching by keywords, etc. Our experience shows that providing better methods to organize information so as to allow students and educators to synthesize, manipulate, and recreate studied contents, is critical to the use of THADL and is therefore in high demand. The work of TH-PASS is preliminary towards this goal. We will undertake more research in this direction.

Acknowledgments. The authors would like to thank Chen Ching-Chih, professor of the Graduate School of Library and Information Science, Simmons College, Boston, USA. She is the chief consultant of the THADL Project and gives us excellent comments and helpful technical support. We also thank all who contributed to this work, especially colleagues and students of the Database Research Group of the Department of Computer Science & Technology, Tsinghua University.

Our work is partially supported by the National Natural Science Foundation of China under Grant No.60221120146; the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704.

References

1. Report of the Santa Fe Planning Workshop on Distributed Knowledge Work Environments: Digital Libraries. http://www.si.umich.edu/SantaFe/, (1997).
2. Arms, W. Y., Hillmann, D., Lagoze, C., Krafft, D., Marisa, R., Saylor, J. , Terrizzi, C. and Van de Sompel, H.: A spectrum of interoperability: The Site for Science prototype of the NSDL. D-Lib Magazine, Vol. 8, No. 1, (2002).
3. Muramatsu, B. and Agogino, A.M.: NEEDS—The National Engineering Education Delivery System: A Digital Library for Engineering Education. D-Lib Magazine, Vol.5, No.4 (1999).

4. IMS Globe Learning Consortium. The IMS Metadata Specification. http: //www.imsproject.org/meta-data.
5. IEEE 1484.12.1- 2002. Draft Standard for Learning Object Metadata (LOM). http://ltsc.ieee.org/wg12/, (2002).
6. Columbia Earthscape: An Online Resource On The Globe Environment. http://www.earthscape.org/
7. Xing Chunxiao, Wu Kaihua, Luo Deyin, Zhou Lizhu, Liu Guilin and Qin Youguo: THADL: A Digital Library for Chinese Ancient Architecture Study. The 12th International Conference on New Information Technology, Beijing, China, (2001) 373-382.
8. Reference Model for an Open Archival Information System (OAIS). Submitted as ISO ISO Draft International Standard, http://www.ccsds.org/documents/pdf/ CCSDS-650.0-R-2.pdf, (2001).
9. Xing Chunxiao, Zhou Lizhu, Wu Kaihua, Liu Guilin, Luo Deyin and Qin Youguo: Design and implementation of a Digital Library for Chinese Ancient Architecture Study and University Education. The Proceedings of Digital Library – IT Opportunities and Challenges in the New Millennium, Beijing, China, (2002) 220-236.
10. Salton, G. and McGill, M. J.: Introduction to modern information retrieval, New York: McGraw-Hill, (1983).
11. Bollacker, K., S. Lawrence, and C. Lee Giles. Discovering Relevant Scientific Literature on the Web, IEEE Intelligent Systems, (2000) 41-47.
12. Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W. Fergerson, and Mark A. Musen. IEEE Intelligent Systems. (2001) 60-71.
13. Lagoze C. and de Sompel, H. V.: The Open Archives Initiative: Building a low-barrier interoperability framework. I n Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries, (2001) 54-62.

Retrieving News Stories from a News Integration Archive

Arnulfo P. Azcarraga, Tat Seng Chua, and Jonathan Tan

Program for Research into Intelligent Systems (PRIS)
School of Computing, National University of Singapore
Lower Kent Ridge Road, Singapore 117543
{dcsapa+dcstsc+dcsjot}@nus.edu.sg

Abstract. The distinctive features of the Bveritas online news integration archive are as follows: 1) automatic clustering of related news documents into themes; 2) organization of these news clusters in a theme map; 3) extraction of meaningful labels for each news cluster; and 4) generation of links to related news articles. Several ways of retrieving news stories from this Bveritas archive are described. The retrieval methods range from the usual query box and links to related stories, to an interactive world map that allows news retrieval by country, to an interactive theme map. Query and browsing are mediated by a Scatter/Gather interface that allows the user to select interesting clusters, out of which the subset of documents are gathered and re-clustered for the user to visually inspect. The user is then asked to select new interesting clusters. This alternating selection/clustering process continues until the user decides to view the individual news story titles.

1 Online News Integration Websites

Most of the major dailies in the world now have online web-based versions of their printed newspapers. This Web-based set-up offers distinct advantages for the readers, aside from delivering news stories from all corners of the globe even when readers have no access to the printed newspapers. First, online news sites are able to deliver late breaking news throughout the day, while in printed news dailies, news of the day usually appear the next day. Second, online websites allow for searching of specific themes which allows readers to concentrate on just a small subset of the entire news collection of the day. Third, online news sites can tailor and personalize the form and organization of the website according to the interest and reading habits of the reader. Fourth, various news tracking services can be connected to the news site. For example, a user may want to be alerted by SMS (short message system) or by e-mail whenever a news item appears on some user-defined news event (e.g. merger of companies, latest football results).

There are a number of existing portals that provide access to news downloaded from numerous online sites. Some of them allow users to only browse the headlines and in most cases, users are directed to the original news sites when specific news titles are selected. These portals provide various levels of news categorization – either by region, or by general areas like business and sports. There are also various support

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 218-228, 2002.

features like links to related stories and other useful information. Among the most significant portals are the following:

1. **Moreover News Portal** - http://www.moreover.com This site collects news from more than 1600 sites. A very detailed topic and regional classification is provided. It is organized as multi-level tree and supported by a quite useful interface.
2. **Simply-info** - http://www.simply-info.co.uk/news This site claims to collect news from 2000 servers and newspapers. The site has mostly business and technology news in the United Kingdom. A fairly detailed news articles categorization is also provided.
3. **The Webnews - Webnews 360 room** - http://www.azreporter.com This has no sophisticated article classification and organization. The user is supposed to use this portal as a start point for browsing the original news sites.
4. **Newstation.com** - http://newstation.com This site pre-selects the most important headlines. For browsing of other news items, links to many news sites are provided.
5. **Columbia Newsblaster** – http:// www.cs.columbia.edu/nlp/newsblaster This site clusters collected news and extracts multidocument summaries for each cluster.

The *Bveritas* system is such an online news integration archive. The title Bveritas is a play of two words – the Malay word "berita" which means "news", and the Latin word "veritas" which means "truth". The system is South-East Asia centric in that it integrates English news websites based in Singapore, Malaysia, Philippines, and Thailand. News stories from these countries are supplemented by downloaded news stories from CNN and Reuters International.

Like most news portals, our system provides both topic and regional categorization, with a combination of headlines and late-breaking news. Differing from most of these existing systems, however, the Bveritas system employs a self-organizing map (SOM) to automatically organize the news articles in a theme map. Through the theme map, the system supports a wide range of services such as retrieval of news articles by keywords, intuitive browsing of news archive by themes, integration of searching and browsing, and the personalization of the news organization to suit the browsing style and interests of readers. Furthermore, our system provides such useful services as news summarization.

This paper describes the general design of the *Bveritas* system and describes the various ways by which news stories can be retrieved from the system's archives – by the use of a regular query box, by clicking on a specific country in an interactive world map, by selecting clusters of news stories via an interactive theme map, and by following links to related stories that are automatically generated whenever a specific news story is displayed on screen. Query and browsing are mediated by a Scatter/Gather interface that lets the user select a few interesting clusters, out of which the subset of documents are gathered and re-clustered for the user to visually inspect, and the user is asked to again select some interesting clusters, and so on.

The paper is organized as follows. The general architecture and main features of the Bveritas system are presented in section 2. Section 3 describes the various ways that news stories can be retrieved from the archives of the Bveritas system. Section 4 describes the interactive Scatter/Gather interface that facilitates searching and browsing. The paper ends with the conclusion in section 5.

2 Bveritas Online News Integration Archive

The *Bveritas* system has most of the elements of an online news site. Figure 1 shows the architecture of the system. Except for the user interface, all modules are performed in the background. The Database is the central part of the system, containing all original and processed news articles, extracted keywords, various special navigational links, etc. The Downloading service is concerned with analyzing homepages of selected news sites and retrieving news articles from URLs picked-up from the news sites. The Parser modules groups together all the various procedures involved in preparing the news articles for further processing, such as the removal of common words (stop words), stemming (i.e. taking the root form of each word [19]), and extraction of special fields like news source, title, and date. Newly parsed documents are passed on to the *classifier*, which analyzes the content of the news body and labels the article according to its story section (i.e. business news, sports news, and socio-political news) and country information (e.g. Singapore, USA, Japan). The SOM takes care of clustering the news documents into themes and assigning keywords that will facilitate query search. Related links among new stories are also supported by the SOM. Finally, all the interaction between the user and the database is handled by the Interface module.

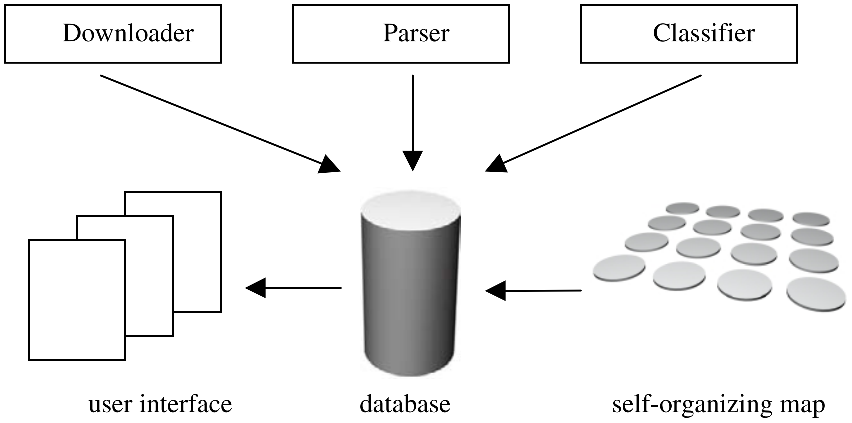


Fig. 1. System architecture of the Bveritas system

Differing from most existing news archiving systems, the Bveritas system employs a *self-organizing map* (SOM) to automatically organize the news articles in a theme map [1,2]. In the theme map, documents are grouped into clusters based on similarity of terms used. The SOM approach to document clustering is very appealing because on top of grouping similar document together to form clusters, the document clusters themselves are organized on a 2D rectangular grid in such a way that clusters of documents that are similar are put beside each other in the grid, while those that are very different are located in dispersed locations on the grid [8, 14-16]. The fact that documents are clustered and that clusters are organized on a 2D grid paves the way for an intuitive interface for browsing through the archive [17,18].

The SOM training algorithm [11-13] iteratively chooses at random a training pattern and determines the winning node whose reference vector registers the highest similarity measure with respect to the training pattern. We use the cosine of the angle between two vectors as similarity measure. Once the winning node is known, all the weights of all the nodes in the map are updated in such a way that nodes that are spatially closer to the winning node have larger weight changes than the nodes that are geographically farther away. The three-step process continues for a pre-set number of training cycles.

After training, the SOM is ready for archiving the entire text collection. Loading is done by assigning each text document to the node in the map that is most similar to it based on the same similarity metric used while training the SOM. As such, each node in the map will have a list of documents that are associated to it.

By virtue of the self-organization properties of SOMs, documents assigned to the same node or to neighboring nodes are similar. To test whether in fact this is true, we collected a total of 429 “related stories” that appeared on the Philippine Daily Inquirer (http://www.inq7.net) from January to June 2002. In this online news site, newspaper editors provide links to related stories for some of the main news reports so online readers can refer to same day or older news reports that have direct bearing on the news story being displayed on screen. There were 132 separate threads of related stories, some with just a pair of related stories while others have up to 10 stories in a single thread.

Table 1. Mean dispersion on the SOM grid of related news stories compared to their mean dispersion if they were allocated randomly to the different nodes in the map

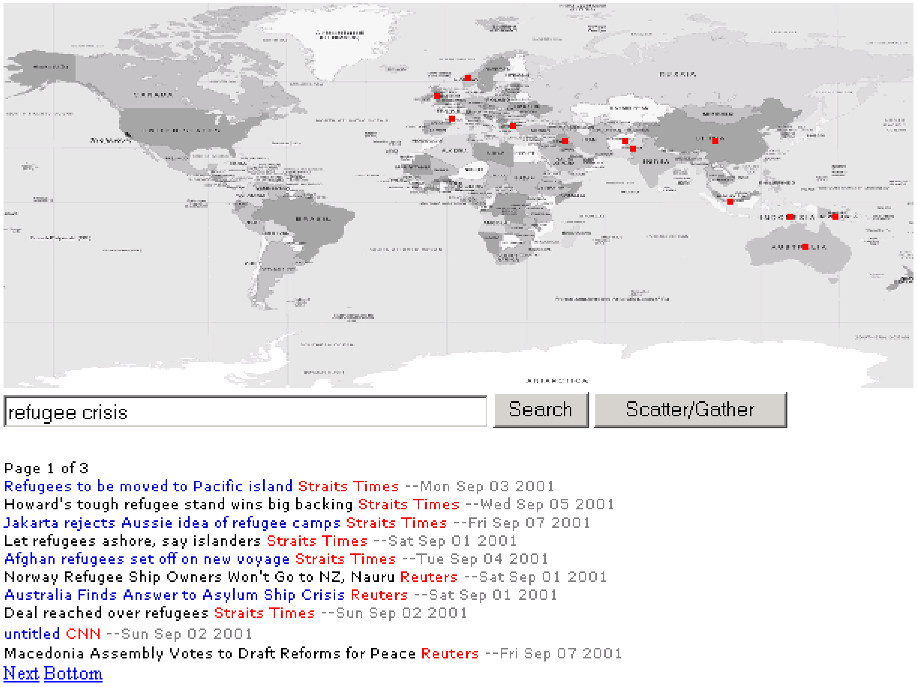
	Trained using related stories dataset (429 articles)			Randomly assigned documents to Nodes		
	8x8	10x10	12x12	8x8	10x10	12x12
at least 2(all)	0.686	0.906	0.913	2.747	3.522	4.127
at least 3	0.848	1.043	1.177	2.909	3.644	4.375
at least 4	0.813	0.961	1.023	2.901	3.631	4.408
at least 5	0.867	1.016	1.066	2.904	3.578	4.383
at least 6	1.116	1.275	1.492	2.873	3.630	4.322
at least 7	1.472	1.630	1.929	2.879	3.588	4.276
at least 8	1.603	1.809	2.277	2.967	3.470	4.313
at least 9	2.101	1.950	2.250	2.915	3.284	3.853
at least 10	2.101	1.950	2.250	2.915	3.284	3.853

Table 1 shows the mean dispersion on the trained SOM grid of the related news stories compared to their dispersion if they were to be allocated randomly to the different nodes in the map. Dispersion is measured as the standard deviation of all distances to the “centroid” location in the map of the different nodes to which the news stories of a single news thread (cluster of related stories) are associated . Thus, if all news stories of a given cluster of related stories are associated to the same node, dispersion is 0. Dispersion increases as the news stories are associated with nodes that are farther apart in the grid. All figures are averages over 5 different SOM's (initial weights are random) for each of the assigned map dimensions of 8x8, 10x10, and 12x12. The figures for the random allocation of news stories to nodes are also averages over 5 tries. Based on the results shown in Table 1, the trained SOMs does seem to cluster related stories (as judged by newspaper editors) under a single node or spatially close nodes in the SOM grid.

3 Alternative Ways of Retrieving News Stories

3.1 Using the Query Box and an Interactive World Map

The system interface allows for the search of news articles through entry of keywords or themes via the usual query box. Responses to queries are returned as a list of news story titles as well as red dots on a picture of the world map (cf. Figure 2). The retrieved list of pertinent news items is shown on the response window at the bottom half of the screen. If there are numerous titles retrieved, these are shown in batches (pages). The user can retrieve a specific news article by clicking on the title found in



refugee crisis

Page 1 of 3
Refugees to be moved to Pacific island Straits Times --Mon Sep 03 2001
Howard's tough refugee stand wins big backing Straits Times --Wed Sep 05 2001
Jakarta rejects Aussie idea of refugee camps Straits Times --Fri Sep 07 2001
Let refugees ashore, say islanders Straits Times --Sat Sep 01 2001
Afghan refugees set off on new voyage Straits Times --Tue Sep 04 2001
Norway Refugee Ship Owners Won't Go to NZ, Nauru Reuters --Sat Sep 01 2001
Australia Finds Answer to Asylum Ship Crisis Reuters --Sat Sep 01 2001
Deal reached over refugees Straits Times --Sun Sep 02 2001
untitled CNN --Sun Sep 02 2001
Macedonia Assembly Votes to Draft Reforms for Peace Reuters --Fri Sep 07 2001
Next Bottom

Fig. 2. World map interface of the Bveritas system

the list. In the figure, the query that was entered was “refugee crisis”, and the list of titles in the first page out of 3 pages are displayed.

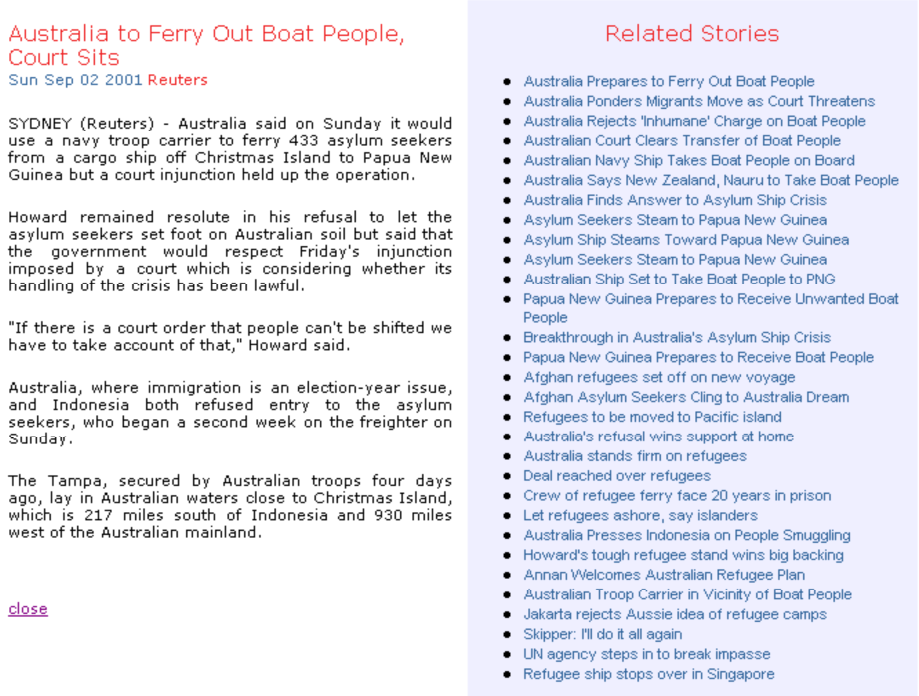
The red dots appear at the approximate locations in the world map of the countries involved in the various news stories in the returned title list. Note the positions of the red dots in the world map, which highlight countries such as Norway (owner of a ship that saved Afghan refugees from the high seas off Australia), Australia, Afghanistan, Indonesia (last take-off point of refugees), plus other countries like France, Great Britain, and China which have various refugee problems as well. The spread of the red dots in the world map gives a succinct indication of the scope of the news stories that are relevant to the query words or themes entered by the user. The user can click on the red dot centered on a specific country, and the response list is reduced to include only those news items that are related to the selected country. At any given time, the user may opt to go on a Scatter/Gather mode by clicking on the Scatter/Gather button. This mode of searching/browsing through the archive will be discussed in detail in section 4.

3.2 Following Links to Related Stories from the Reader’s Corner

When a user clicks on the title of a specific news story, the Reader’s Corner window is brought up, as shown in Figure 3. For every news article that is retrieved, the system automatically determines all the related news stories based on common use of various terms in the news bodies. Links are generated and the related news stories are listed on the right side of the Reader’s Corner window, starting with the most related story down to the least related.

In the current system, the search and ranking of related stories are done on the fly. Searching for all related stories can be done fast because the theme map has already pre-grouped the documents into clusters. Documents that are in the same cluster as the news story being displayed are automatically included as candidates for “related stories”. Likewise, since the theme map is a self-organized map, those documents associated with nodes in the immediate vicinity of the node where the displayed news story is found are also included in the set of candidate related stories. We are confident based on the experiments leading to the results results shown in Table 1 that indeed, related stories are mostly within one or two nodes away in the grid. The search for related stories from only a small region in the map is the reason why our method can do the search very quickly without indexing everything before hand.

All candidate related stories are subjected to a finer similarity measurement using the cosine of the angle between document vectors representing the displayed news story and those of the candidate related stories. The similarity measure ranges in values from 0 to 1, with 1 denoting a perfect match (i.e. identical documents). This will be the basis for ranking the documents from the most related to the least related. We only list those news stories with a similarity measure of at least 0.75. This cut-off threshold has been chosen based on empirical findings which tend to show that news stories with lower similarity measures may share various words and phrases in common with the displayed news item, but talk of quite distinct news events.



Australia to Ferry Out Boat People, Court Sits

Sun Sep 02 2001 Reuters

SYDNEY (Reuters) - Australia said on Sunday it would use a navy troop carrier to ferry 433 asylum seekers from a cargo ship off Christmas Island to Papua New Guinea but a court injunction held up the operation.

Howard remained resolute in his refusal to let the asylum seekers set foot on Australian soil but said that the government would respect Friday's injunction imposed by a court which is considering whether its handling of the crisis has been lawful.

"If there is a court order that people can't be shifted we have to take account of that," Howard said.

Australia, where immigration is an election-year issue, and Indonesia both refused entry to the asylum seekers, who began a second week on the freighter on Sunday.

The Tampa, secured by Australian troops four days ago, lay in Australian waters close to Christmas Island, which is 217 miles south of Indonesia and 930 miles west of the Australian mainland.

close

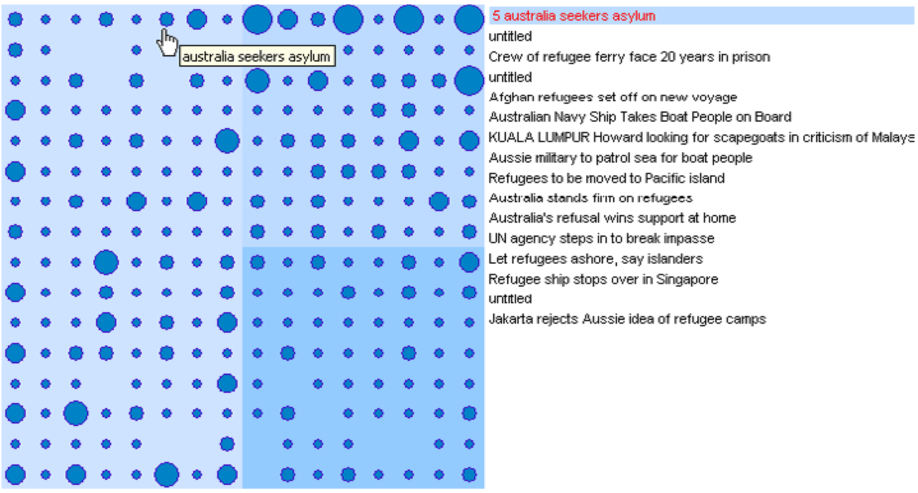
Related Stories

- Australia Prepares to Ferry Out Boat People
- Australia Ponders Migrants Move as Court Threatens
- Australia Rejects 'Inhumane' Charge on Boat People
- Australian Court Clears Transfer of Boat People
- Australian Navy Ship Takes Boat People on Board
- Australia Says New Zealand, Nauru to Take Boat People
- Australia Finds Answer to Asylum Ship Crisis
- Asylum Seekers Steam to Papua New Guinea
- Asylum Ship Steams Toward Papua New Guinea
- Asylum Seekers Steam to Papua New Guinea
- Australian Ship Set to Take Boat People to PNG
- Papua New Guinea Prepares to Receive Unwanted Boat People
- Breakthrough in Australia's Asylum Ship Crisis
- Papua New Guinea Prepares to Receive Boat People
- Afghan refugees set off on new voyage
- Afghan Asylum Seekers Cling to Australia Dream
- Refugees to be moved to Pacific island
- Australia's refusal wins support at home
- Australia stands firm on refugees
- Deal reached over refugees
- Crew of refugee ferry face 20 years in prison
- Let refugees ashore, say islanders
- Australia Presses Indonesia on People Smuggling
- Howard's tough refugee stand wins big backing
- Annan Welcomes Australian Refugee Plan
- Australian Troop Carrier in Vicinity of Boat People
- Jakarta rejects Aussie idea of refugee camps
- Skipper: I'll do it all again
- UN agency steps in to break impasse
- Refugee ship stops over in Singapore

Fig. 3. Reader's Corner with links to related stories

3.3 Retrieving News Stories from a Theme Map

The theme map of Figure 4 is trained using news articles covering socio-politics, business and sports news for the period Sept 1-7, 2001. A cluster of related news stories are associated to each node in the SOM. The size of the blue circle indicates the relative number of news stories associated with the node. The user may glide the mouse over the nodes of the grid on the screen, and for each node, a set of labels appears on screen and representative news story titles of documents belonging to the node are shown on the right side of the map. The labels are extracted by our own label extraction method [1, 2] similar to the label extraction method of [18]. Clicking on a node, as opposed to just a "mouse over", would retrieve the complete list of document titles associated to the node and clicking on a specific news title brings up the Reader's corner as mentioned earlier. The user may also click on the "Activate Scatter/Gather" button and then start clicking on nodes to collect the initial "focus set". All documents in the focus set are subjected to the interactive Scatter/Gather session to be discussed in the next section.



5 australia seekers asylum
untitled
Crew of refugee ferry face 20 years in prison
untitled
Afghan refugees set off on new voyage
Australian Navy Ship Takes Boat People on Board
KUALA LUMPUR Howard looking for scapegoats in criticism of Malays
Aussie military to patrol sea for boat people
Refugees to be moved to Pacific island
Australia stands firm on refugees
Australia's refusal wins support at home
UN agency steps in to break impasse
Let refugees ashore, say islanders
Refugee ship stops over in Singapore
untitled
Jakarta rejects Aussie idea of refugee camps

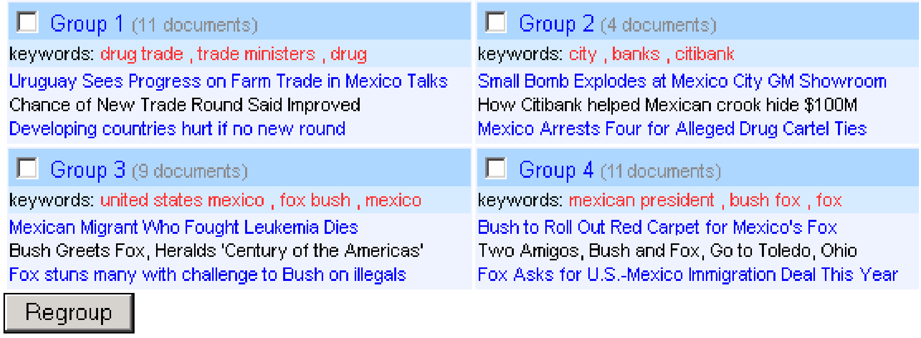
Fig. 4. Sample theme map in the Bveritas system

4 Using Scatter/Gather for Browsing and Searching

The Bveritas system uses Scatter/Gather as a single interface to the various ways of accessing the contents of our digital archive. Scatter/Gather was first proposed in 1992 as a browser interface for large document archives [3]. A few years later, at about the time when the World Wide Web had just gained immense popularity, Scatter/Gather was resurrected [5-7] - this time as a tool for navigating through retrieval results from Web search engines.

Scatter/Gather is an interactive clustering method that provides an intuitive and effective interface for browsing and searching document archives. In *scatter* mode, the document collection is clustered into several groups of documents. A description of each cluster is then extracted and shown to the user for visual inspection. The user can then select one or more of the given clusters that he/she thinks are relevant. The system then switches to *gather* mode. In *gather* mode, the documents in the chosen clusters are collected and re-scattered into a fresh set of clusters. The alternating scatter/gather process continues until the number of documents in the selected clusters is small enough that the user can just view the individual document titles.

For illustration purposes, suppose the user is interested in finding news articles about the visit to the US of Vicente Fox, President of Mexico, and in particular the news stories highlighting his discussions with George Bush concerning the immigration of Mexicans to the US. In this example, the user can start the session by clicking on a number of nodes in the theme map where such keywords as “Fox” and “Mexico” are found. Figure 5 shows the Scatter/Gather interface using gathered documents from the theme map which were used as the initial “focus set”. There were a total of 35 documents. Scanning through the descriptions of the resultant clusters after “scatter”, the contents of some clusters become fairly evident. Groups 3 and 4 contains news stories about the meeting between Bush and Fox while other miscellaneous socio-political news about Mexico are found in clusters 1 and 2.



□ Group 1 (11 documents)
keywords: drug trade , trade ministers , drug
Uruguay Sees Progress on Farm Trade in Mexico Talks
Chance of New Trade Round Said Improved
Developing countries hurt if no new round

□ Group 2 (4 documents)
keywords: city , banks , citibank
Small Bomb Explodes at Mexico City GM Showroom
How Citibank helped Mexican crook hide $100M
Mexico Arrests Four for Alleged Drug Cartel Ties

□ Group 3 (9 documents)
keywords: united states mexico , fox bush , mexico
Mexican Migrant Who Fought Leukemia Dies
Bush Greets Fox, Heralds 'Century of the Americas'
Fox stuns many with challenge to Bush on illegals

□ Group 4 (11 documents)
keywords: mexican president , bush fox , fox
Bush to Roll Out Red Carpet for Mexico's Fox
Two Amigos, Bush and Fox, Go to Toledo, Ohio
Fox Asks for U.S.-Mexico Immigration Deal This Year

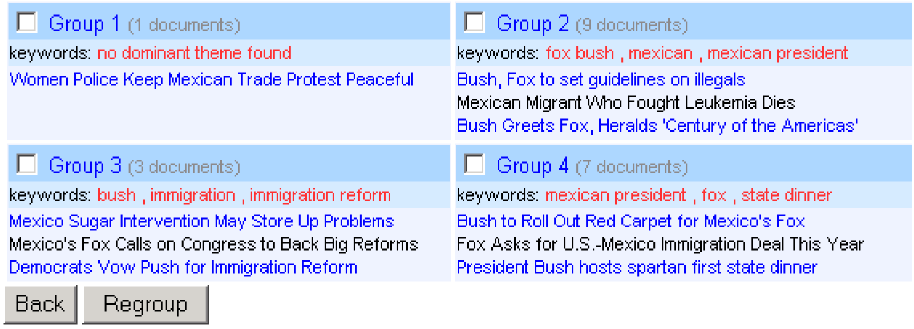
Fig. 5. Sample Scatter/Gather interface on initial focus set from gathered document list from the clicked–on nodes of the theme map

For the next iteration, suppose the user selects clusters 3 and 4, as these are the most relevant with respect to the topic of interest. Figure 6 shows the result of re-clustering using the smaller focus set of only 20 documents out of the initial 35 documents. The stories about the meeting between Bush and Fox, focused on the hospitality accorded to the visiting Mexican president, are found in cluster 4. The documents of cluster 2 are focused on the immigration discussions per se, while cluster 3 focuses on the role of the US congress. The lone document of cluster 1, an outlier, is kept separate. Given the sample news titles of the various groups, the decision is left to the user to probe any of the clusters. In this case, he/she may want to start with cluster 2, and then move on to some related news stories found in cluster 3.

There are three ways by which the Scatter/Gather interface comes into play with the rest of Bveritas system. The first is as an interactive interface for navigating through search retrieval results. As depicted in Figure 2, query results are returned as red dots printed over a world map, displaying the distribution of countries represented in the query results. At this point, the user would have two options. One is to click on a specific red dot, and the subset of the list of only those news stories related to the clicked-on country will be displayed at the bottom half of the screen. The other option is to click on the Scatter/Gather button, which then launches the Scatter/Gather interface, using whatever is on the retrieved search list as the initial focus set.

The second way that Scatter/Gather is used in the Bveritas system is through the world map when no search query is issued. Red dots on the world map reflect the distribution of all the news stories contained in the entire archive. When the user clicks on a specific red dot, then all those news stories related to the chosen country are listed on the bottom half of the screen, the same way it is done as when it is with a search query. At this point, the user again has the option of clicking on the Scatter/Gather button to start the interactive clustering session.

The third way that Scatter/Gather is used is for browsing the archive. The entry point to the archive is the SOM-based theme map. The user clicks on various nodes of the map and the Scatter/Gather session is activated. All the documents associated with all the clicked-on nodes are *gathered* to constitute the first “focus set”. Then Scatter/Gather proceeds as described earlier.



□ Group 1 (1 documents)
keywords: no dominant theme found
Women Police Keep Mexican Trade Protest Peaceful

□ Group 2 (9 documents)
keywords: fox bush , mexican , mexican president
Bush, Fox to set guidelines on illegals
Mexican Migrant Who Fought Leukemia Dies
Bush Greets Fox, Heralds 'Century of the Americas'

□ Group 3 (3 documents)
keywords: bush , immigration , immigration reform
Mexico Sugar Intervention May Store Up Problems
Mexico's Fox Calls on Congress to Back Big Reforms
Democrats Vow Push for Immigration Reform

□ Group 4 (7 documents)
keywords: mexican president , fox , state dinner
Bush to Roll Out Red Carpet for Mexico's Fox
Fox Asks for U.S.-Mexico Immigration Deal This Year
President Bush hosts spartan first state dinner

Fig. 6. Scattered clusters based on a smaller focus set gathered from clusters 3 and 4 of the previous set of clusters.

We have improved on Scatter/Gather in three respects. First, instead of truncating the document term vectors to only select the top *n* words (in one implementation of Scatter/Gather, they used n=50 [4]), we use random projection that reduces the number of dimensions of the resultant document encoding without sacrificing the quality of the clustering results [10,14,15]. Note that when truncating document term vectors, similarity measures between documents are no longer as accurate. Second, we use the “organized” layout of SOM-based document archives as the initial clustering. In [5], there is a mention of clusters that are laid out in a “semantic continuum”. We believe that the organization of document clusters of our SOM-based theme map, where clusters that are similar are positioned spatially near each other in the 2D SOM grid, would be a natural starting point for Scatter/Gather. Third, the extraction of cluster labels and document summaries are very crucial components of Scatter/Gather. Since we use random projection, our label extraction for the various clusters must likewise adapt to this new document encoding scheme [1,2].

5 Conclusion

The Bveritas system is a South-East Asia centric online news archive that integrates and organizes news articles from English news websites based in Singapore, Malaysia, Philippines, and Thailand. News stories from these countries are supplemented by downloaded news stories from CNN and Reuters International. The system is supported by a Self-Organizing Map (SOM) which gives it the following distinctive features: 1) automatic clustering of news stories into groups of related news articles from different news sites; 2) automatic organization of these news clusters in a theme map, 3) automatic extraction of meaningful labels for each cluster of news articles; and 4) automatic generation of links to related news articles.

We described several ways of retrieving news stories from the Bveritas archive. The retrieval methods discussed include the following: 1) using of usual query box; 2) following links to related stories starting from some initial news story; 3) using an interactive world map that allows news retrieval by country; and 4) using an interactive theme map. Query and browsing are mediated by a Scatter/Gather interface that allows the user to select interesting clusters, out of which the subset of documents are gathered and re-clustered (scattered) for the user to visually inspect, and the user is asked to select again some interesting clusters. This alternating Scatter/Gather process continues until the number of documents in the selected clusters is small enough so that the user can simply view all the individual document titles.

References

1. Azcarraga, A., and Yap, T. Jr. (2001) Extracting Meaningful Labels for WEBSOM-Based Text Archives. *10th ACM International Conference on Information and Knowledge Management (CIKM 2001)*, Atlanta, USA.
2. Azcarraga, A., Yap, T. Jr., and Chua, T.S. (2002) Comparing Keyword Extraction Techniques for WEBSOM Text Archives. *International Journal of Artificial Intelligence Tools*, Vol. 11, No 2.
3. Cutting, D. et al. (1992) Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. *Proc 15th ACM/SIGIR*, Copenhagen.
4. Cutting, D. et al. (1993) Constant Interaction-Time Scatter/Gather Browsing of Large Document Collections. *Proc 16th ACM/SIGIR*, Pittsburg.
5. Hearst, M. et al. (1995) Scatter/Gather as a Tool for the Navigation of Retrieval Results. *Proc 1995 AAAI Fall Symposium on Knowledge Navigation*.
6. Hearst, M. et al. (1996) Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. *Proc 19th ACM/SIGIR*, Zurich.
7. Hearst, M. et al. (1996) Four TREC-4 Tracks: the Xerox Site Report. *Proc 4th Text REtrieval Conference (TREC-4)*, Nov 1-3, Arlington, VA.
8. Honkela, T. et al. (1997) WEBSOM-Self-Organizing Maps of Document Collections. *Proc WSOM'97,* Espoo, Finland.
9. Kaski, S. et al. (1998) Statistical Aspects of the WEBSOM System in Organizing Document Collections. *Computing Science and Statistics*. Vol. 29. pp. 281-290.
10. Kaski, S. (1998) Dimensionality reduction by random mapping: Fast similarity computation for clustering. *Proc IJCNN'98, International Joint Conference on Neural Ne*tworks, Vol. 1, Piscataway, NJ.
11. Kohonen, T. (1982) Analysis of a Simple Self-Organizing Process, *Biological Cybernetics*, Vol. 44, pp. 135-140.
12. Kohonen, T. (1988) *Self-Organization and Associative Memory*. Series in Information Sciences, Second Edition. Berlin, Springer-Verlag.
13. Kohonen, T. (1995) *Self-Organizing Maps*. Berlin, Springer-Verlag.
14. Kohonen, T. (1998) Self-Organization of Very Large Document Collections: State of the Art. *Intl Conference on Artificial Neural Networks, ICANN98*. Skovde, Sweden.
15. Kohonen, T. et al. (2000) Self Organization of a Massive Document Collection, *IEEE Trans on Neural Networks*, Vol. 11, no 3, pp. 574-585.
16. Merkl, D. and Rauber, A. (2000). Uncovering the Hierarchical Structure of Text Archives by Using an Unsupervised Neural Networks with Adaptive Architecture. *PAKDD'2000*. Kyoto, Japan.
17. Lagus et al, (1999) WEBSOM for Textual Data Mining. *Artificial Intelligence Review*, Vol. 13, pp. 345-364.
18. Rauber, A. and Merkl, D. (1999). Mining Text Archives: Creating Readable Maps to Structure and Describe Document Collections. *PKDD99*.
19. Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA.

A Data Mining Approach to New Library Book Recommendations[*]

San-Yih Hwang[1] and Ee-Peng Lim[2]

[1]Department of Information Management
National Sun Yat-Sen University
Kaohsiung 80424, Taiwan
syhwang@mis.nsysu.edu.tw

[2]Centre for Advanced Information Systems
School of Computer Engineering
Nanyang Technological University
Nanyang Avenue, Singapore 639798
aseplim@ntu.edu.sg

Abstract. In this paper, we propose a data mining approach to recommending new library books that have never been rated or borrowed by users. In our problem context, users are characterized by their demographic attributes, and concept hierarchies can be defined for some of these demographic attributes. Books are assigned to the base categories of a taxonomy. Our goal is therefore to identify the type of users interested in some specific type of books. We call such knowledge *generalized profile association rules*. In this paper, we propose a new definition of rule interestingness to prune away rules that are redundant and not useful in book recommendation. We have developed a new algorithm for efficiently discovering generalized profile association rules from a circulation database. It is noted that generalized profile association rules can be applied to other kinds of applications, including e-commerce.

1 Introduction

Book recommendation in the digital library context is similar to product recommendation in electronic commerce. In the past few years, we have seen the emergence of many recommendation systems that provide personalized recommendation of various types of products and services, including news (GroupLens), web pages and research articles (citeseer.nj.nec.com), books (amazon.com), albums (CDNow.com), and movies (MovieFinder.com) [SKR01]. The basic idea behind recommendation techniques is to recommend products according to the users' preferences, which are either explicitly stated by the users or implicitly inferred from previous transaction records, web logs,

[*] This research was supported by National Science Council, ROC, under grant NSC 90-2213-E-110-022.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 229-240, 2002.

or cookies data. Most recommendation techniques fall into two categories, namely *content-based filtering* and *collaborative filtering* [CACM92]. The content-based filtering technique characterizes product items by a set of content features and users' interest profiles by a similar feature set. A similarity function is then defined to measure the relevance of a product item and a user's interest profile. Product items having high degrees of similarity with a user's interest profile are then recommended to the user. This method assumes that common features exist between the product items and users in order to define a reasonable similarity function. Unfortunately, this assumption does not hold for many applications where the features of user interest profiles are incompatible with the products' features. Even when such common content features exist between product items and users, the fact that only product item with content features similar to that of a user will imply no surprising recommendation can ever be found by the content-based techniques. The *collaborative filtering* (also called *social filtering*) techniques address this problem by taking into account the given user's interest profile and the profiles of other users with similar interests [SM95]. Specifically, the collaborative filtering techniques look for similarities among users by observing the ratings they assign to products. Given a target user, the nearest–neighbor users are those who are most similar in terms of product rating assignments. These users then act as the "recommendation partners" for the target user, and a collaborative filtering technique will recommend product items that appear in the transactions of these recommendation partners but not in the target user's transactions. To realize collaborative filtering, many measures have been proposed to predict the rating a person will give to an un-rated product item, based on either simple calculation on the rating matrix, e.g., correlation, cosine and regression, or a more complicated probability model, e.g., Bayesian classifier and Bayesian network [BHK98], or a combination of both [PHLG00]. It has been shown that collaborative filtering techniques yield more effective recommendations [Pazz99, MR00].

However, while the idea of recommending to a given user those products in which his peers have shown interest has demonstrated its usefulness in many applications, it has its limitations. First, it provides no or limited interpretation for a recommendation. Second, the prediction is poor if the data is sparse. In other words, unless there is a sufficient number of common items rated by users, the prediction may be unreliable. Finally, collaborative techniques fail to recommend newly introduced products that have not yet been rated by users.

Example. Consider the new book recommendation system of the National Sun Yat-sen University (NSYSU) library. The NSYSU library currently houses over 600,000 volumes of books and bound periodicals, and this amount is increasing at the pace of 6% per year. In other words, each month there are about 3,000 new books, which is a long list and unlikely to be browsed by any individual. As there are only 6,000 students enrolled at NSYSU, statistics from the circulation system show that the check-out figures are very low. The average number of books ever borrowed by a patron is around 30, and 75% of the library's books have never been checked out. Also, the circulation system records a wide variety of patrons' demographic information, including address, gender, birthday, degree sought, program majored, work unit, and academic status. In the library domain, each book is well classified by experts according to some classification scheme. The NSYSU library adopts a Chinese classification scheme and the Library of Congress classification scheme to classify oriental books and western books, respectively. It is an important task to recommend a small number of new books to patrons based on their interests derived from past circulation (check out) records.

The above example shows that it may not be appropriate to infer the interests of a patron from his or her check-out records and that it is important for recommendation systems in the context of new library books to incorporate demographic information of patrons and/or genres of books. In [Pazz99], Pazzani derived demographic information of students from their home pages and used classification techniques to identify the characteristics of users who liked a particular restaurant. In [Kim01], Kim et al. used decision tree techniques to infer the characteristics of customers who liked a particular product category, as opposed to an individual item as considered in [Pazz99]. However, aggregating demographic attribute values was not explored by either study.

Our work takes into account a wide variety of information in making new book (or product) recommendations, including customers' demographic information, book (product) attribute values, customers' borrowing (purchase) records, and the concept hierarchies on demographic and book (product) attributes. Specifically, our approach starts with the discovery of generalized association rules that determine the associations between types of customers and book types. Less interesting or redundant rules are then pruned to form a concise rule set. The above two steps are time consuming and conducted off-line. In step 3, the resulting rule set is then used for on-line promotion of new books (products). Due to space constraints, this paper will only focus on the first 2 steps of mining interesting generalized association rules. To give a more general discussion, we will use the terms 'book' and 'product' interchangeably.

This paper is structured as follows. In Section 2, we will formally define the problem of mining generalized profile association rules for new product recommendations. Section 3 describes the data mining algorithm we developed. Section 4 presents our interestingness measure and the approach we used to prune redundant rules. Finally, Section 5 concludes with a summary and discussion of future research directions.

2 The Problem

Our data mining approach to the new product recommendation is to identify a set of strong associations between types of customers and genres of products that frequently appear in a transaction database, followed by the recommendations by using these association rules. Suppose there are k demographic attributes with domains being $D_1,\ldots, D_k$ respectively. Let $P = \{p_1, p_2, ..., p_r\}$ be the set of product items. An aggregation hierarchy on the i'th demographic attribute, denoted $H(D_i)$, is a tree with the set of leaves being equal to D_i, and an internal node represents a demographic type. A taxonomy on P, denoted $H(P)$, is a tree with the set of leaves being equal to P and internal nodes indicate product categories. A link in H represents an is-a relationship.

Each transaction in the transaction database may records the identifier of a customer, the products (books) s/he has purchased (checked out), and the time of the transaction. To facilitate mining generalized profile association rules, we group transactions of the same customer and include the customer's demographic information, resulting in a new type of transaction called *demographic-product transaction*. Specifically, the demographic-product transaction of the *i*'th customer is represented as a tuple $t_i = < d_{i,1}, d_{i,2}, ..., d_{i,k}, p_{i,1}, p_{i,2}, ..., p_{i,s} > (1 \le i \le n, k \ge 1, s \ge 1)$, where $d_{i,j}$ is a leaf in $H(D_j)$ that represents the *j*th demographic attribute value of the *i*th customer, and $p_{i,t}$ is a leaf in $H(P)$ that represents the *t*th product item that the *i*th customer has ever purchased. In the following discussion, unless otherwise stated, when we say a transaction we actually refers to a demographic-product transaction. Since our goal is to identify the associations between customer demographics types and product categories, the demographic values and product items presented in each transaction must be converted into demographic types and product categories respectively, resulting in a so called *extended transaction* [SA95]. Here we simply include all demographic types of each demographic value and all product categories of each product item appeared in the transaction. Therefore, the *i*'th transaction can be translated to the extended transaction $t_i' = <d_{i,1}', d_{i,2}', ..., d_{i,u}', p_{i,1}', p_{i,2}', ..., p_{i,m}'> (1 \le i \le n, u \ge 1, m \ge 1)$, where $d_{i,j}'$, $1 \le j \le u$, and $p_{i,j}'$, $1 \le j \le m$, are internal nodes in $H(D_j)$ and $H(P)$ respectively. We say that the transaction t_i supports a demographic type $d' = (d_1, d_2, \ldots, d_l)$ if $\{d_1, d_2, \ldots, d_l\} \subset t_i'$, where t_i' is the extended transaction of t_i. Similarly, we say that t_i supports a product category c if $c \in t_i'$. A *generalized profile association rule* is an implication of the form $X \rightarrow Y$, where X is a demographic type and Y is a product category. The rule $X \rightarrow Y$ holds in the transaction set T with a confidence $c\%$ if c percent of the transactions in T that support X also support Y. The rule $X \rightarrow Y$ has support $s\%$ in the transaction set T if s percent of the transactions in T support both X and Y. Therefore, given a set of transactions T and several demographic aggregation hierarchies $H(D_1)$, $H(D_2)$, …, $H(D_k)$ (each one representing the generalization of one demographic attribute), and one product taxonomy $H(P)$, the problem of mining generalized profile association rules from transaction data is to discover all rules that have support and confidence greater than the user-specified minimum support (called Min_{sup}) and minimum confidence (called Min_{conf}). These rules are named *strong* rules.

3 Identifying Generalized Profile Association Rules

Since the goal is to identify generalized profile association rules, the itemsets that will interest us are of the following form $< d_{i_1}, d_{i_2}, ..., d_{i_l}, p >$, where $d_{i_j} \in$ is an internal node in $H(D_{i_j})$ and p is an internal node in $H(P)$. Such itemsets are called *demographic-product itemsets*. By finding large (or frequent) demographic–product itemsets, one can easily derive the corresponding generalized profile association rules. In the following, we present our proposed **GP-Apriori** algorithm for generating frequent itemsets.

GP-Apriori is a slight modification to the original Apriori algorithm proposed in [SA95] for mining generalized association rules. Consider the classical problem of discovering generalized frequent itemsets from market basket databases, where all items in an itemset are product items and a taxonomy for all product items is given [SA95, HF95]. It is possible to directly employ the existing techniques to discover the generalized demographic–product itemsets. In other words, a (demographic-product) transaction can be visualized as a market basket transaction by treating both demographic attribute values and product items homogeneously as ordinary items. However, this straightforward approach is inefficient and may generate many useless rules with antecedent and consequent being of the same type (products or demographic attributes). This problem of unwanted rules can be easily addressed by modifying the way candidate itemsets are generated. Let L_k denote the frequent itemsets of the form $< d_{i_1}, d_{i_2}, ..., d_{i_k}, p >$. A candidate itemset C_{k+1} is generated by joining L_k and L_k in a way similar to the Apriori candidate generation algorithm [AS94], except that the *k* join attributes must include one product (*p*) and the other *k*-1 demographic attribute values (from $d_{i_1}, d_{i_2}, ..., d_{i_k}$).

Specifically, this modified approach works as follows. We first extend each transaction $t_i =< d_{i,1}, d_{i,2}, ..., d_{i,k}, p_{i,1}, p_{i,2}, ..., p_{i,s} > (1 \le i \le n, k \ge 1, s \ge 1)$ in T as described above. The set of extended transactions is denoted *ET*. After scanning the data set *ET*, we obtain large demographic 1-itemsets $L_1(D)$ and large product 1-itemsets $L_1(P)$. If an item is not a member of $L_1(D)$ or $L_1(P)$, it will not appear in any large demographic–product itemset and is therefore useless. We delete all the useless items in every transaction of *ET* in order to reduce its size. The set C_1 of candidate 1-itemsets is defined as $L_1(D) \times L_1(P)$. Data set *ET* is scanned again to find the set L_1 of large demographic-product 1-itemsets from C_1. A subsequent pass, say pass k, is composed of two steps. First, we use the above-mentioned candidate generation function to generate the set C_k of candidate itemsets by joining two large ($k-1$)-itemsets in L_{k-1} on the basis of their common $k - 2$ demographic attribute values and the product attribute value. Next, data set *ET* is scanned and the support of candidates in C_k is counted. The set L_k of large k-itemsets are itemsets in C_k with minimum support. This algorithm is called “GP-Apriori” because it is an extension of Apriori algorithm for finding Generalized Profile association rules. The pseudo-code is eliminated for brevity

4 Pruning Uninteresting Rules

From the large demographic-product itemsets derived from the GP-Apriori algorithm described in the previous section, it is trivial to derive the generalized profile association rules that satisfy both Min_{sup} and Min_{conf}. However, some of the strong generalized profile association rules could be related to each other in either the demographic itemset part (the antecedent) or the product itemset part (the consequent), and therefore the existence of one such rule could make some others not interesting. There has been a lot of work for measuring the interestingness of association rules on items [AL99, LHM99 PT00, SK98, JS02]. A rule $A \rightarrow C$ is said to be a sub-rule of another rule $B \rightarrow C$ if $A \subset B$. A rule that has confidence close to one of its sub-rules is considered not interesting. Many approaches that try to identify such rules and prune them have been proposed in the literature. With respect to generalized association rules, Srikant and Agrawal defined an interestingness measure that is used to prune descendant rules given an ancestor rule[1] [SA95]. In their work, a rule R is interesting if and only if for every close ancestor rule R', the support of R is at least γ times higher than the expected support derived from R', or the confidence of R is at least γ times higher than the expected support derived from R', where γ is a user-specified threshold. The intuition is that if the support and confidence of a rule can be derived from any of its ancestor rules, this rule is considered uninteresting and can be pruned.

All the previous work described above favors more general rules because they have wider scope of application, and the more specialized rules will not be picked unless they are much stronger in terms of support or confidence. Take the library circulation data mining, to which our approach has been applied, as an example. The existence of a rule such as R_1: "engineering students" → "mathematics books" will make a more specialized rule R_2: "CS students" → "mathematics books" not interesting unless the later is much stronger in terms of support and confidence. While this approach is useful in some cases, it falls short in identifying those ancestor rules that are strong simply because some of the descendant rules are strong. Furthermore, to recommend product items, specificity of rules should be considered. That is, if a descendant rule has adequate support and confidence, it will make a better rule for product recommendation than its ancestor rules that have slightly higher or similar support and confidence. Suppose that the following rule is strong: R_1: "CS students" → "computer books". Then the rule, R_2: "CS students" → "mathematics books", must also be strong because every computer book is classified as a mathematics book by the library classification scheme. However, although R_2 is more general, this rule may not be interesting if most transactions that support R_2 also support R_1. We see R_2 as interesting only when many students who major in CS have also been issued non-computer mathematics books. In this case, it makes sense to recommend non-computer mathematics books to CS students. Consider another association rule R_3: "engineering students" → "computer books". If R_1 is strong, then R_3 must satisfying minimum support. Again, R_3 is not interesting if most transactions that support R_3 come from those supporting R_1. In contrast, we will consider R_3 as interesting if a sufficient number of engineering students who are not CS majors have also been issued with computer books.

[1] As defined in [SA95], a rule $X \rightarrow Y$ is an ancestor of another $X' \rightarrow Y'$ if $X' \subseteq X$ and $Y' \subseteq Y$. Given a set of rules, a rule R is called a close ancestor of R' if there does not exist a distinct rule R'' in the set such that R is an ancestor of R'' and R'' is an ancestor of R'. R' is said to be a descendant of R if R is an ancestor of R'.

Based on the above observation, we develop our "interestingness" measure as follows. Let Π be the set of all demographic attribute types, i.e., $\Pi = \text{internal nodes}(H(D_1)) \cup \text{internal nodes}(H(D_2)) \cup \cdots \cup \text{internal nodes}(H(D_k))$. For a given constant γ, $0 \le \gamma \le 1$, rule $D \rightarrow p$ is called γ-confident if its confidence is no less than $\gamma \cdot Min_{conf}$. We call a rule R_1: $D' \rightarrow p_1$, where $D' \subseteq \Pi$ and $p_1 \in P$, a D-ancestor of another rule R_2: $D'' \rightarrow p_2$ where $D'' \subseteq \Pi$ and $p_2 \in P$, if $p_1 = p_2$ and $\forall d_1 \in D', \exists d_2 \in D''$, such that d_1 is equal to or an ancestor of d_2 in the associated demographic concept hierarchy (i.e., D'' is more specific than D'). Similarly, we call a rule R_1: $D' \rightarrow p_1$ a P-ancestor of another rule R_2: $D'' \rightarrow p_2$ if $D' = D''$ and p_1 is equal to or an ancestor of p_2 in the product taxonomy. Also, R_2 is called a D-descendant (P-descendant) of R_1 if R_1 is a D-ancestor (P-ancestor) of R_2. For example, both (CS students)→(mathematics books) and (male, engineering students) →(mathematics books) are D-descendants of (engineering students) →(mathematics books), and (engineering students) →(computer books) is a P-descendant of (engineering students) →(mathematics books).

Given a set of strong rules and a given constant γ_1, $0 \le \gamma_1 \le 1$, a generalized profile association rule R: $D \rightarrow p$ is downward-interesting if

- R does not have any D-descendant or for all close D-descendants of R, $R_1 : D' \rightarrow p, R_2 : D'' \rightarrow p, \cdots, R_l : D^{(l)} \rightarrow p$, $D - (D' \cup D'' \cup \cdots \cup D^{(l)}) \rightarrow p$ (called *D-deductive rule*) is γ_1-confident, and
- R does not have any P-descendant or for all close P-descendants of $R, R_1 : D \rightarrow p_1, R_2 : D \rightarrow p_2, \cdots, R_{l'} : D \rightarrow p_{l'}$, $D \rightarrow p - (p_1 \cup p_2 \cup \cdots \cup p_{l'})$ (called *P-deductive rule*) is γ_1-confident.

Note that in the above definition, to determine whether a rule R: $D \rightarrow p$ is downward-interesting, we do not consider the more specialized rule R': $D' \rightarrow p'$, where $D' \subset D$ and $p' \subset p$. This is because if R': $D' \rightarrow p'$ is strong, so is $D' \rightarrow p$, a D-descendant of R. Therefore, it suffices to consider only the D-descendants and P-descendants when it comes to determining downward-interestingness. The intuition behind the downward-interestingness measure is that a more general rule will interest us only when it cannot be represented collectively by some less general rules (i.e., D-descendants or P-descendants). In other words, if the deduction of a rule and its D-descendants (P-descendants) still present sufficiently high confidence (γ_1-confident), then this rule should be preserved. The downward-interestingness measure favors more specialized rules, and a more general rule is selected only if it can be generally applied to the specializations of its antecedents. It is important to prune out rules that are not downward-interesting because it is misleading to apply these rules for making recommendations. For example, if the rule R: $D \rightarrow p$ is not downward-interesting because its P-deductive rule $D \rightarrow p - (p_1 \cup p_2 \cup \cdots \cup p_{l'})$ is not γ_1-confident, it does not make sense to recommend a product of category $p - (p_1 \cup p_2 \cup \cdots \cup p_{l'})$ to a customer characterized by D. However, when the set of descendant rules can be indeed fully represented by a more general, downward-interesting rule, the rule and its descendant rules will be preserved. Although the existence of such descendant rules will not affect the effectiveness of the product recommendation, the large number of rules may impact performance. Therefore we propose to combine both downward-interestingness and the measure proposed in [SA95] (we call it upward-interestingness) and define a hybrid new interestingness measure as follows:

Given a set of strong rules and a given constant γ_2, $\gamma_2 \geq 1$, a generalized profile association rule R: $D \rightarrow p$ is interesting if

- R is downward interesting.
- For each close ancestor R' of R that are downward interesting, the confidence of R is at least γ_2 times the expected confidence based on R'.

In this definition, in addition to being downward interesting, a rule must present sufficiently high confidence with respect to each of its ancestor in order to be considered interesting. Note that the expected confidence of a rule $D \rightarrow p$ based on an ancestor rule $D' \rightarrow p'$ is represented as $conf(D' \rightarrow p') \cdot \frac{\sup(p)}{\sup(p')}$, where sup($p$) is the support of p. Also, unlike the work [SA95] of Srikant and Agrawal which considers a rule as interesting if it has higher value in either support and confidence, we focus only on confidence as the goal is to identify the association between demographics and products.

The set $\mathfrak{R}$ of all strong rules can be seen as a partial order set (POSET) ($\mathfrak{R}$, <), where $r_1 < r_2$, $r_1, r_2 \in \mathfrak{R}$ if r_1 is an ancestor of r_2. The constructive definition of our interestingness measure suggests a bottom-up traversal (for identifying downward interesting rules), followed by a top-down traversal (for identifying upward interesting rules). However, the difficulties of identifying downward interesting rules lie in the computation of the confidences of D-deductive and P-deductive rules. We approximate the confidence of a D-deductive rule by using the following theoretic results:

Lemma 1[2]**.** Let $D', D'', \ldots, D^{(l)}$ be mutually disjoint, the confidence of $D - (D' \cup D'' \cup \cdots \cup D^{(l)}) \rightarrow p$ is $\frac{\sup(D,p) - \sup(D',p) - \sup(D'',p) - \ldots - \sup(D^{(l)},p)}{\sup(D) - \sup(D') - \sup(D'') - \ldots - \sup(D^{(l)})}$.

Theorem 1. Without loss of generality, let $D', D'', \ldots, D^{(i)}, 1 \leq i \leq l$, be mutually disjoint. Assume that $Conf((D^{(i+1)} \cup \ldots \cup D^{(l)}) - (D' \cup \ldots \cup D^{(i)}) \rightarrow p) \geq \gamma_1 \cdot Min_{conf}$. If $D - (D' \cup D'' \cup \cdots \cup D^{(l)}) \rightarrow p$ is γ_1-confident,

$$\frac{\sup(D,p) - \sup(D',p) - \sup(D'',p) - \ldots - \sup(D^{(i)},p)}{\sup(D) - \sup(D') - \sup(D'') - \ldots - \sup(D^{(i)})} \geq \gamma_1 \cdot Min_{conf}.$$

Note that to apply Theorem 1, the equation $Conf((D^{(i+1)} \cup \ldots \cup D^{(l)}) - (D' \cup \ldots \cup D^{(i)}) \rightarrow p) \geq \gamma_1 \cdot Min_{conf}$ must hold. Refer to Figure 1, since $D_4 \rightarrow p$ and $D_5 \rightarrow p$ both have confidences higher than Min_{Sup}, it is very likely that $Conf((D_4 \cup D_5) - (D_1 \cup D_2 \cup D_3) \rightarrow p) \geq \gamma_1 \times Min_{Conf}$, where $\gamma_1 < 1$. $((D_4 \cup D_5) - (D_1 \cup D_2 \cup D_3)$ is shown in shaded area in Figure 1.)

[2] We have skipped the proofs of all lemmas and theorems due to space constraints.

Therefore, this equation will hold in many cases. When computing the confidence of a D-deductive rule *r*: $D-(D'\cup D''\cup\cdots\cup D^{(l)})\to p$, we first find a (maximum) set of mutually disjoint domains $D', D'', \ldots, D^{(i)}, 1\le i\le l$, and compute the confidence of $D-(D'\cup D''\cup\cdots\cup D^{(i)})\to p$. If the confidence is less than $\gamma_1\cdot Min_{conf}$, we drop the rule because it is likely that *r* is not γ_1-confident.

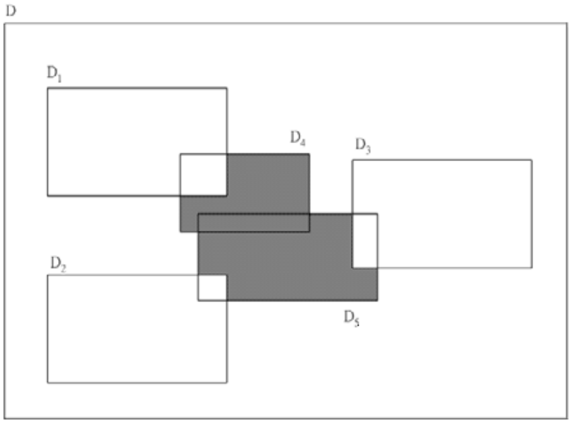


Fig. 1. Pictorial representation of *D* and its five sub-domains D_1, D_2, D_3, D_4, and D_5.

Now we discuss how to compute the confidence of a P-deductive rule *r*: $D\to p-(p_1\cup p_2\cup\cdots\cup p_{l'})$. The transactions that support *r* must have included products that fall outside $p_1\cup p_2\cup\cdots\cup p_{l'}$. We say product categories p_j and p_i are siblings if they have a common parent in the respective concept hierarchy. Let *NoSiblingTrans*(*D*, p_i) denote the set of transactions that support (*D*, p_i) but does not support any sibling of p_i. Obviously, none of the transactions in *NoSiblingTrans*(*D*, p_i) supports *r*. Let *NoSiblingSup*(*D*, p_i) denote the ratio of the number of transactions in *NoSiblingTrans*(*D*, p_i) to the total number of transactions in the database. To calculate *NoSiblingSup* (*D*, p_i), we associate a flag *NoSibling* on each product category of an extended transaction. *NoSibling*(*p*, *et*), where *p* is a product category and *et* is an extended transaction, is equal to 1 if there exists no sibling of *p* in *et* and 0 otherwise. Therefore, *NoSiblingSup* (*D*, p_i)= $\dfrac{\sum_{et \text{ supports } (D,p_i)} NoSibling(p_i, et)}{n}$, where *n* is the total number of transactions.

Theorem 2. If *r*: $D\to p-(p_1\cup p_2\cup\cdots\cup p_{l'})$ is γ_1-confident,

$$\frac{\sup(D,p)-\sum_{1\le i\le l'} NoSiblingSup(D,p_i)}{\sup(D)}\ge\gamma_1\cdot Min_{conf}\,.$$

NoSiblingSum(*D*, p_i) for a demographic-product itemset (*D*, p_i) can be computed when counting the support for (*D*, p_i) by GP-Apriori described in Section 3, and such a computation causes negligible overhead. Theorem 2 shows that $\frac{\sup(D,p) - \sum_{1\le i\le l} NoSiblingSum(D,p_i)}{\sup(D)}$ is an upper bound of the confidence of $D \to p - (p_1 \cup p_2 \cup \cdots \cup p_{l'})$. Therefore, if the upper bound is less than $\gamma_1 \cdot Min_{conf}$, we drop the rule because it cannot be γ_1-confident.

For example, consider the four strong rules shown in Table 1. The bottom-up traversal starts with the rule "CS students→computer books", which is downward interesting because it does not have any D-descendant or P-descendant. Then we determine that the rule "Engineering students→computer books" is not downward interesting because the D-deductive rule "Non CS-majored engineering students →computer books" has low confidence 1/18 as shown below:

$$\frac{|E,comp| - |CS,comp|}{|E| - |CS|} = \frac{sup(E,comp) - sup(CS,comp)}{sup(E) - sup(CS)} = \frac{sup(E,comp) - sup(CS,comp)}{\frac{sup(E,comp)}{conf(E \to comp)} - \frac{sup(CS,comp)}{conf(CS \to comp)}}$$

$$= \frac{1/18 - 1/20}{1/6 - 1/15} = 1/18 < Min_{conf} \cdot \gamma = 20\%$$

The rule "CS students→math books" is not downward interesting either because the P-deductive rule "CS students→(math – comp) books" has confidence no higher than 13/110 as shown below:

$$\frac{sup(CS,math) - NoSiblingSup(CS,comp)}{sup(CS)} = \frac{4/75 - 1/22}{1/15} = 13/110 < Min_{conf} \cdot \gamma = 20\%$$

The rule "Engineering students→math books", however, is downward interesting because both the D-deductive rule "Non CS-majored engineering students→math books" and the P-deductive rule "Engineering students→ (math-comp) books" have high confidences as shown below:

Conf(Non CS-majored engineering students→math books)

$$= \frac{|E,math| - |CS,math|}{|E| - |CS|} = \frac{sup(E,math) - sup(CS,math)}{sup(E) - sup(CS)}$$

$$= \frac{sup(E,math) - sup(CS,math)}{\frac{sup(E,math)}{conf(E \to math)} - \frac{sup(CS,math)}{conf(CS \to math)}} = \frac{1/8 - 4/75}{1/6 - 1/15} = 43/60 > Min_{conf} \cdot \gamma = 20\%$$

Conf(Engineering students→ (math-comp) books)

$$= \frac{sup(E,math) - NoSiblingSup(E,comp)}{sup(E)} = \frac{1/8 - 1/40}{1/6} = 3/5 > Min_{conf} \cdot \gamma = 20\%$$

In the subsequent top-down traversal (for testing upward-interestingness), no rules will be pruned. Therefore, at the end of traversal, only two rules remain: "Engineering students→math books" and "CS students→computer books". Note that if we simply adopt upward-interestingness, it is likely that all four rules will be preserved (because the confidences of "CS students→math books" and "Engineering students→computer books" could be higher than the estimated confidences derived from "Engineering students→math books"). As a result, ineffective recommendations, such as recommending pure math books to CS students or computer books to non-CS engineering students, will be subsequently made.

Table 1. Four example rules

Minconf = 25% γ= 0.8

Strong rules	**confidence**	**support**	**NoSiblingSup**
CS students→computer books	75%	1/20	1/22
CS students→math books	80%	4/75	Don't care
Engineering students→computer books	33.3%	1/18	1/40
Engineering students→math books	75%	1/8	Don't care

5 Conclusion

We have examined the problem of mining generalized profile association rules for recommending new books (products) that have no circulation (transaction) records and have proposed a novel approach to this problem. This approach starts with the identification of the associations between demographic types and product categories. We have developed an algorithm for this task. The obtained generalized profile associations rules are pruned by a new interestingness measure that favors special rules over general rules. We are in the process of evaluating the application of the discovered rules to recommend new books in the context of university libraries. Preliminary performance results will be shown during the conference presentation.

References

[AL99] Y. Aumann and Y. Lindell, "A statistical theory for quantitative association rules," *Proc. of the 5'th ACM SIGKDD Int'l. Conf. on Knowledge Discovery and Data Mining*, pp. 261-270, 1999.

[AS94] R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," *Proc. of the 20th VLDB Conf.*, pp. 478–499, Sept. 1994.

[BHK98] J. S. Breese, D. Heckerman and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Tech. Report*, MSR-TR-98-12, Microsoft Research, Oct. 1998.

[CACM92] Special issue on information filtering, *Communications of the ACM*, 35(12), Dec. 1992.

[HF95] J. Han and Y. Fu, “Discovery of multiple-level association rules from large databases,” *Proc. of the 21st VLDB Conf.*, pp. 420–431, 1995.

[JS02] S. Jaroszewicz and D. A. Simovici, “Pruning redundant association rules using maximum entropy principle,” *Proc. of 6’th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD2002), Taipei, Taiwan, 2002.

[Kim01] J. W. Kim, B. H. Lee, M. J. Shaw, H. L. Chang, and M. Nelson, “Application of Decision-tree Induction Techniques to Personalized Advertisements on Internet Storefronts,” *International Journal of Electronic Commerce*, 5(3), pp. 45-62, 2001.

[LHM99] B. Liu, W. Hsu, and Y. Ma, “Pruning and summarizing the discovered associations,” *Proc. of the 5’th ACM SIGKDD Int’l. Conf. on Knowledge Discovery and Data Mining*, pp.125-134, N.Y. Aug., 1999.

[MR00] R. J. Mooney and L. Roy, “Content-based book recommending using learning for text categorization,” *Proc. Of the 5’th ACM Conf. on Digital Libraries*, pp. 195-240, June 2000.

[Pazz99] M. J. Pazzani, “A framework for collaborative, content-based and demographic filtering,” *Artificial Intelligence Review*, pp. 393-408, 1999.

[PHLG00] D. Pennock, E. Horvitz, S. Lawrence and C. L. Giles, “Collaborative filtering by personality diagnosis: a hybrid memory- and model-based approach,” *Proc. of the 6’th Conf. on Uncertainty in Artificial Intelligence* (UAI-2000), pp. 473-480, 2000.

[PT00] B. Padmanabhan and A. Tuzhilin, “Small is beautiful: discovering the minimal set of unexpected patterns,” *Proc. of the 6’th ACM SIGKDD Int’l. Conf. on Knowledge Discovery and Data Mining*, pp.54-63, Aug. 2000.

[SA95] R. Srikant and R. Agrawal, “Mining generalized association rules,” *Proc. of the 21st VLDB Conf.*, pp. 409–419, 1995.

[SK98] E. Suzuki and Y. Kodratoff, “Discovery of surprising exception rules based on intensity of implication,” *Proc. of PKDD-98*, France, p.10-18, 1998.

[SKR01] J. B. Schafer, J. A. Konstan, and J. Riedl, “E-Commerce Recommendation Applications,” *Data Mining and Knowledge Discovery*, 5(1), pp. 115-153, 2001.

[SM95] U. Shardanand and P. Maes, “Social Information Filtering: Algorithms for Automating ‘Word of Mouth,” *Proc. Of the Conference on Human Factors in Computing Systems* (CHI’95), pp. 210-217, 1995.

Grouping Web Pages about Persons and Organizations for Information Extraction

Shiren Ye, Tat-seng Chua, Jimin Liu, and Jeremy R. Kei

School of Computing, National University of Singapore, Singapore, 117543
{yesr,chuats,liujm,jkei}@comp.nus.edu.sg

Abstract. Information extraction on the Web permits users to retrieve specific information on a person or an organization. As names are non-unique, the same name may be mapped to multiple entities. The aim of this paper is to describe an algorithm to cluster Web pages returned by search engines so that pages belonging to different entities are clustered into different groups. The algorithm uses named entities as the features to divide the document set into direct and indirect pages. It then uses distinct direct pages as seeds of clusters to group indirect pages into different clusters. The algorithm has been found to be effective for Web-based applications.

1 Introduction

Information Extraction (IE) is a hot area of research. As defined in the Message Understanding Conference[1], IE involves the extraction of named entities (NEs), scenarios, and events from input documents. The application of effective IE technologies on the Web enables retrieval of not just documents as is currently done by search engines, but more specific information. For example, a query on a person name on the Web should return a summary of all information related to the person, rather than a ranked list of Web pages containing one or more words in that person's name. Through the IE system, the user will submit a name of a person or an organization as a query, and the system will search the Web, collect all relevant Web pages, and extract a summary of desired information.

Fig. 1 shows the overall process of information extraction on the Web. As the submitted name is typically non-unique, it may be mapped to multiple persons or organizations. To resolve this problem, we need to perform clustering of the Web pages returned so that pages belonging to different entities (person or organization) are clustered into different groups. We can then concentrate on extracting information relating to a specific entity from each cluster. Search engines incorporating clustering can also return clusters to help users browse the results.

This paper presents our work on the clustering of Web pages containing names of persons and organizations (PnO), to support the overall information extraction process. The aim is to cluster pages belonging to different entities (persons or organizations) into different clusters. We employ our tools to identify named entities in all the returned pages. We then use a combination of named entities, URL, and links as the features to perform the clustering. Our testing indicates that it is effective. The main contribution of this research is in providing an effective clustering tool for PnO pages. To the best of our knowledge, there is no other related work on this topic.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 241-251, 2002.

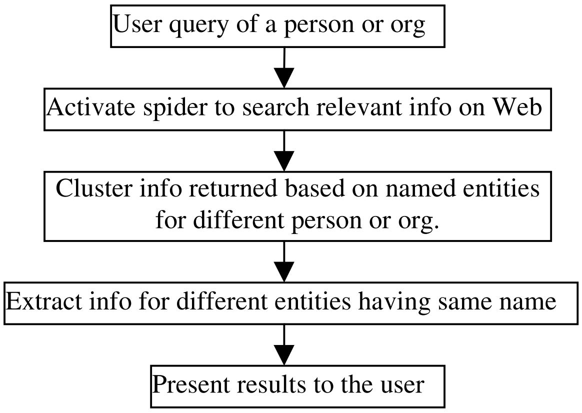

User query of a person or org
Activate spider to search relevant info on Web
Cluster info returned based on named entities for different person or org.
Extract info for different entities having same name
Present results to the user

Fig. 1. The process of Web-based information extraction

In section 2, we discuss the issues in clustering Web pages on PnO. Section 3 discusses document features based on named entities; section 4 presents the algorithm to identify seeds of clusters. The method of delivering indirect pages into clusters is explained in section 5. The results of our experiments and conclusion are, respectively, presented in section 6 and section 7.

2 Clustering of Web Pages

Document Clustering algorithms attempt to identify groups of documents that are more similar to each other than the rest of the collection. In a typical document clustering algorithm, each document is represented as a weighted attribute vector, with each word in the entire document collection being an attribute in this vector (see *vector-space model* [2]). Besides probabilistic technique such as Bayesian, a priori knowledge for defining the distance or similarity measures among them is used to compare similarities between documents.

Information foraging theory [3] notes that there is a trade-off between the value of information and the time spent in finding it. The vast quantity of Web pages returned as the result of a search means that some form of clustering or summarization of the results is essential. Several new approaches have emerged to group or cluster Web pages. Those include association rule hyper-graph partitioning, principal direction divisive partitioning [4], Suffix Tree Clustering (STC) [5]. Scatter/Gather [6] (www.parc.xerox.com) clusters text documents according to their similarities and automatically computes an overview of the documents in each cluster. Unfortunately, most search engines at present do not use clustering as a regular procedure during Information Retrieval.

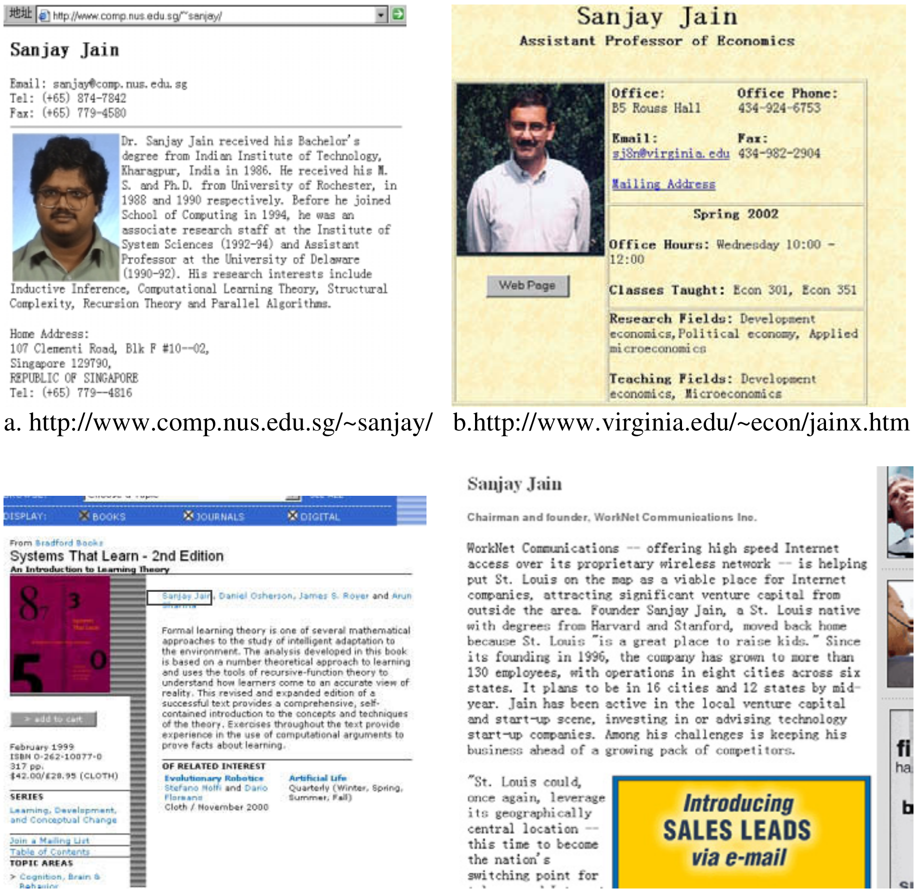
PnOs are common query requests during Internet surfing. Users frequently submit personal or organization names to search engines, which return results that are usually quite good and normally include the target among the top ranked results. However, there are still many problems with the search results as outlined below.

- Most users have patience to browse only 10-20 pages. The number of pages returned can reach thousands.
- Search results may contain several persons and/or organizations whose names are the same as the query string. If the search results could be grouped into different clusters, with the pages about different entities grouped separately, then users can concentrate on more promising clusters.
- Some retrieved pages are completely irrelevant but are displayed nonetheless because they contain phrases which are similar to the name. For example, a fable page or AI research page may appear when the user is trying to find information about the software company “Oracle Co.”.
- The low-ranking pages listed at the end of a results list may often be of only minor importance or could even be just tangentially related. However, they are not always useless. In some cases, novel or unexpectedly valuable results can be found in these pages. For instance, a report of a company involved in a fraud case may be ranked at the bottom of thousands of returned pages, but pages such as this will be significant to users trying to evaluate the worthiness of the company.

As shown in Figure 2, when we submitted the query "*Sanjay Jain*" to Google, at least ten persons named "*Sanjay Jain*" were returned. Here, pages (a) and (b) are the homepages about two different persons. Page (c) is an introduction of a book authored by the person in page (a). Page (d) is the description of another person, but its style is different from that of pages (a) and (b). It can be seen that the search engine returns a great variety of both correct and incorrect results. If we are able to identify and partition the corresponding results into clusters (in this example, into three clusters for three different persons), it will facilitate users in browsing the results.

3 Document Features Based on Named Entities

In most clustering approaches, similarity (distance) between a pair of documents is computed as the cosine of the angle between the corresponding vectors in the feature space [7]. The feature vectors should be scaled to avoid skewing the result by different document lengths or possibly by how common a word is across many documents. Many techniques such as TFIDF and stop word list [7] have been proposed for these problems. However, they do not work well for PnOs. Consider two resume pages about two different persons. It is highly possible that they will be grouped into one because they have many similar words and phrases, such as graduate, university, work, degree, join, employment, department and so on. This is especially so when their style, pattern and glossary are also similar. On the other hand, it is difficult to group a news page and resume page together even though the same person is mentioned in these pages. This is because their lengths and glossaries are very different. To solve this problem, it is essential that the right set of features be used to identify pages about a PnO.



Sanjay Jain

Email: sanjay@comp.nus.edu.sg
Tel: (+65) 874-7842
Fax: (+65) 779-4580

Dr. Sanjay Jain received his Bachelor's degree from Indian Institute of Technology, Kharagpur, India in 1986. He received his M. S. and Ph.D. from University of Rochester, in 1988 and 1990 respectively. Before he joined School of Computing in 1994, he was an associate research staff at the Institute of System Sciences (1992-94) and Assistant Professor at the University of Delaware (1990-92). His research interests include Inductive Inference, Computational Learning Theory, Structural Complexity, Recursion Theory and Parallel Algorithms.

Home Address:
107 Clementi Road, Blk F #10--02,
Singapore 129790,
REPUBLIC OF SINGAPORE
Tel: (+65) 779--4816

a. http://www.comp.nus.edu.sg/~sanjay/

Sanjay Jain
Assistant Professor of Economics

Office:
B5 Rouss Hall

Office Phone:
434-924-6753

Email:
sj8n@virginia.edu

Fax:
434-982-2904

Mailing Address

Spring 2002

Office Hours: Wednesday 10:00 - 12:00

Classes Taught: Econ 301, Econ 351

Research Fields: Development economics, Political economy, Applied microeconomics

Teaching Fields: Development economics, Microeconomics

b.http://www.virginia.edu/~econ/jainx.htm

From Bradford Books
Systems That Learn - 2nd Edition
An Introduction to Learning Theory

Sanjay Jain, Daniel Osherson, James S. Royer and Arun Sharma

Formal learning theory is one of several mathematical approaches to the study of intelligent adaptation to the environment. The analysis developed in this book is based on a number theoretical approach to learning and uses the tools of recursive-function theory to understand how learners come to an accurate view of reality. This revised and expanded edition of a successful text provides a comprehensive, self-contained introduction to the concepts and techniques of the theory. Exercises throughout the text provide experience in the use of computational arguments to prove facts about learning.

February 1999
ISBN 0-262-10077-0
317 pp.
$42.00/£28.95 (CLOTH)

SERIES
Learning, Development, and Conceptual Change

Join a Mailing List
Table of Contents

TOPIC AREAS
> Cognition, Brain & Behavior

OF RELATED INTEREST
Evolutionary Robotics
Stefano Nolfi and Dario Floreano
Cloth / November 2000

Artificial Life
Quarterly (Winter, Spring, Summer, Fall)

c. http://mitpress.mit.edu/catalog/item /default.asp?ttype=2&tid=7277

Sanjay Jain

Chairman and founder, WorkNet Communications Inc.

WorkNet Communications -- offering high speed Internet access over its proprietary wireless network -- is helping put St. Louis on the map as a viable place for Internet companies, attracting significant venture capital from outside the area. Founder Sanjay Jain, a St. Louis native with degrees from Harvard and Stanford, moved back home because St. Louis "is a great place to raise kids." Since its founding in 1996, the company has grown to more than 130 employees, with operations in eight cities across six states. It plans to be in 16 cities and 12 states by mid-year. Jain has been active in the local venture capital and start-up scene, investing in or advising technology start-up companies. Among his challenges is keeping his business ahead of a growing pack of competitors.

"St. Louis could, once again, leverage its geographically central location -- this time to become the nation's switching point for

Introducing
SALES LEADS
via e-mail

d. http://www.bizjournals.com/stlouis /stories/2000/01/31/focus37.html

Fig. 2. Typical pages when "Sanjay Jain" is submitted to Google

In general, we observe that the occurrences of PnO related named entities (NEs) in the web pages about PnOs is higher than in other types of pages. Here, PnO NEs include person, location and organization name, time and date, fax/phone number, currency, percentage, e-mail, URL, Link and so on. For simplicity, we will call these entities collectively as NEs. We could therefore use NEs as the basis to identity PnO pages. To support our claims, we collected and analyzed 1,000 pages for PnOs and 1,000 other pages from the Web. We found that the percentage of NEs in PnO pages is at least 6 times higher than in other pages, if we ignore NEs of type number and percentage.

The finding is quite consistent with intuition, as NEs play important roles in semantic expression and can be used to reflect content of the pages, especially when human activities are depicted. The typical number of NEs appearing in the results of a search is only in the hundreds or thousands, which means that it is feasible to use them as the features of search results about PnOs. Our analysis also showed that NEs are good at partitioning pages belonging to different persons or organizations, while frequent phrases and words (such as degree, education, work) are not.

NE recognition is a fundamental step to many natural language processing tasks. It was a basic task of the Message Understanding Conference (MUC) [1] and has been studied intensively. Recent research on NE recognition has been focused on the machine learning approach, such as hidden Markov model [8], decision tree [9], collocation statistics [10], and maximum entropy model [11]. As reported in MUC-7 [1], the accuracy and recall of NE recognition in English is above 95% and is as good as what human experts can achieve. The best reported results in NE recognition in Chinese are above 90% [12]. Thus we can accurately extract NEs from the pages and then use them to reflect the content of those pages. In our system, we use the decision tree to detect English NEs and the Rationality computation and default decision tree [12][13] to detect Chinese NEs.

4 Identifying Seeds of Clusters

It is generally observed that there are two types of pages returned by search engines when a user makes a query on a PnO. The first type is almost entirely about the users' focus. Examples of such pages include homepages, profiles, resumes, biographies and memoirs of PnOs. These pages contain a large number of NEs, such as graduation schools, contact information (phone, fax, e-mail, address), working organizations and experiences (time and organizations). We call such pages *Direct Pages*. The second type of pages are *Indirect Pages*, where the concept related to the user's query string is just mentioned briefly. For instance, the person's name may appear in a page about the attendees of a conference, staff of a company, record of a transaction, or the homepage of his friend.

Since a large proportion of the content of direct pages depicts items related to the query string, the relevance between direct pages and query is larger than that between the query and the indirect pages. Direct pages could provide more information than the indirect pages that satisfies the users' need. Therefore, we should choose the direct pages as the candidates of clusters' seeds. Of course, if there is more than one direct page about a target entity, then only the best one is selected as the seed for clustering.

To select the best direct pages of a target entity, we need to solve two problems. First we must be able to identify a direct page from the indirect pages. Second, in the case of multiple direct pages for the same target entity, we must be able to select the best one. In this paper, we employ the following measures to identify the direct pages that can be used as the seeds for target entities. We observe that the number and percentage of NEs in the direct pages are much larger than those in the indirect pages do. Suppose that the number of NEs is N_{NE}, the number of tokens in pages is N_{token}. The percentage of NEs of a page is

$$f=N_{NE}/N_{token} \tag{1}$$

In our experiment, NEs within the HTML tags are not accounted into N_{token}. We use a measure that combines N_{NE} and f to provide a balance between both quantities to identify direct pages as

$$\theta = N_{NE} * f = N_{NE}^{2} / N_{token} \qquad (2)$$

For example, if there are 7 NEs in a 100-token page, then θ=7*7/100=0.49. The page is considered a direct page if θ is larger than a threshold τ_1.

Next, if there is more than one direct page found for a target entity, we need to find the best candidate as the seed. We observe that if both the homepage and resume of *John Smith* are selected as direct page, those two pages will share many similar NEs related to John Smith, such as the university that he graduated. Thus we could evaluate the similarity between two direct pages by examine their overlaps in instances of unique NEs. Here we use TFIDF to estimate the weight of each unique NEs as follows.

$$W_{i,j} = tf_{i,j} * \log(N/df_i) \qquad (3)$$

where $tf_{i,j}$ is number of NE i in page j; df_i is the number of pages containing NE i; and N is the number of pages.
The similarity of direct page p_i and p_j could be expressed by their cosine distance as.

$$sim(p_i, p_j) = \frac{\sum_k (w_{k,i}{}^{c} * w_{k,j})}{\sqrt{\sum_k (w_{k,i})^2 * \sum_k (w_{k,j}{}^{c})^2}} \qquad (4)$$

If $sim(p_i,p_j)$ is larger than a pre-defined threshold τ_3, then p_i and p_j are considered to be similar. The page that has more NEs will be used as the seed and the other will be removed. Because the number of direct pages is a small fraction of the search results, and the number of NEs in direct pages is usually less than hundreds, thus the computational cost in eliminating redundant direct pages is acceptable.

Third, because not all pages with high number of NEs, like member list of a conferences and stock price lists etc, are not direct pages. We should further check the roles of target entities those appear in the query in the text. In general, if the target entity appears in important locations, such as in HTML tags <title>, <H1> and <H2>, or it appears frequently, then the corresponding pages should be really direct pages and their topic is about the users' target. We could detect the trace of page topic using technology like wrapper rules [14].

According to the above discussion, the procedure to identify seeds of clusters is summarized as following:

```
Detect_seed(page_set)
{
  set seed_set=null;//the collection of candidate seeds
  for each (page in page_set){
    sum up $N_{token}$, $N_{NE}$ and $N_{topic}$;
    //where $N_{topic}$ is number of query strings
    //appearing in a page
    $\theta$= $N_{NE}^2/N_{token}$;
    if ($\theta>\tau_1$ && ($N_{topic} > \tau_2$ ||query_string is
      in title)) add page from page_set into seed_set;
  }
```

```
    for each pair $p_i$, $p_j$ in seed_set:
      if (Sim($P_i$,$P_j$)> $\tau_3$){
        if ($N_{NE}$ in $p_i$ >$N_{NE}$ in $P_j$)
          move $p_j$ from seed_set into page_set;
        else
          move $p_i$ from seed_set into page_set;
      }
    return seed_set;
  }
```

At the end of the process, the pages remaining in the collection seed_set could be used as the seeds for clusters---they are representatives of entities named in the query string. The titles of seeds could be regarded as labels of the clusters. The NEs in titles or heads could be used as alternatives to labels.

5 Delivering Indirect Pages to Clusters

Compared to direct pages, indirect pages provide less information about the target entity. Nevertheless, it does not mean that they are less important. Actually, the information extracted from indirect page may be more novel and provide more valuable information to the users. For example, your classmate may list your name in his homepage, though you have not contacted him for many years and do not know of such a page in Web. You must be surprised to find this page and feel that it is very useful. Generally, indirect page could:

- Provide additional information such as activity or experience of the target entity.
- Support or oppose the content in direct page whether they are consistent or not.
- Provide critical or negative content which may not appear in the direct page. It thus provides important information to evaluate the target entities fairly and integrality.

Therefore, we must explore an approach to link direct pages and indirect pages properly. In other words, we want to add indirect pages into clusters which are created by the seeds (direct pages). As mentioned in the above example, some of the NEs (such as your name, graduate school, or even period and degree) in your classmate's homepage is similar to those in your direct page. We can thus use the similarity between these NEs to link them together. In other words, we could compute their similarity based on a selected set of NE features.

Besides NEs in pages, URL and links in pages could also be used as heuristics to select and rank indirect pages with respect to a seed page. If the roots of URLs are same (such as www.comp.nus.edu and www.comp.nus.edu/~pris), or components of URLs are similar (such as www.nus.edu.sg and www.comp.nus.edu.sg), there should have some associations among them. Similarly, if there is a link between the direct page and the indirect page, they should not be separated. To avoid complicating the question, we suppose that URL, links and NEs have same weight, namely, URL and links are regarded as other types of NEs.

We use the algorithm below to select and link indirect pages to a seed page.

```
Arrange_indirect_page(page_set,cluster_set)
//clusters are presented by their seeds
{
  set unknown_set=null; //collection of useless pages
  foreach (page_i in page_set)
  {
    $j = \arg\max_j sim(page_i, seed_j)$ //see equation 4
    if ($j > \tau_4$)
      add pagei into cluster_j;
    else
      add page_i into unknown_set;
  }
}
```

6 Experiments and Discussion

Grouping web pages about PnOs is a pivotal component in our Web-based Information Extraction System (see Figure 1) [15]. Here search results about PnOs are segmented into different clusters according to their target entities, and then the pages in different clusters are used to fill in different templates which are related to different entities.

Experiment of web information processing is a time-consuming task, where each search typically returns hundreds, or even thousands of pages. Moreover, evaluating the effectiveness of clustering is notorious even though there are many guidelines such as entropy [5], clustering error [16], and average precision [6] to measure the quality of clustering. We obtained the primary results according to following steps. Because of lack of comparable results and standard test data, we just provide our preliminery results.

a) We collect the names of 30 persons and 30 names of organization (such as companies, governments and schools) in English from Yahoo (www.yahoo.com), MSN (www.msn. com). We control PnOs that belong to large companies and famous persons (such as *Microsoft or George W. Bush*), since there would be too many pages in the search results. For example, Google returns 2,880,000 pages for Microsoft, and first hundreds of pages are about only one special target. To ensure sufficient data for the analysis, we also excluded those PnOs whose returning pages are less than 30.

b) We use every PnO name in the above collection as query string and submit to Google. The results of the searches are downloaded. If the returning pages are more than 1,000, we downloaded just the first 1,000 pages. We also filter out the files whose formats are not HTML and plain, such as PDF, PS, PPT formats, and those whose lengths are less than 100 or more than 10,000. The average number of pages returned is 227.

c) NEs are detected from the downloaded Web pages. We remove the numbers that are used to list the items in the pages. We, however, include e-mail addresses and telephone number as a part of NEs. The average number of NEs for each page is 15.78.

d) Cluster seeds are then detected from each set of search results. The number of clusters depends on the parameters used in the algorithm detect_seed. If τ_1 and τ_2 are smaller, the algorithm will produce more candidates of seeds. However, most of them will be removed during the step of eliminating redundancy. In our experiment, the candidates vary from 1 to 30s, where τ_1 is set to 0.64 and τ_2 to 5. The seeds that detect_seed outputs vary from 0 to 11. The average number of seeds is 3.5. The number of seeds for person is larger than that for organization. This means that the number of persons with the same name is larger than that of organizations.

e) Indirect pages are added into different clusters. The average number of indirect pages in each cluster is 38. The indirect pages in each cluster about organizations are considerably more than that about persons.

f) The quality of seeds is pivotal because it controls the distribution of segmentation. Missing a seed will entail some indirect pages being assigned into wrong or unknown-set and a cluster missing. If there are redundant seeds, the direct pages about the same target may be delivered into different clusters. Fortunately, it is quite easy to differentiate between direct pages and indirect pages by using our algorithm.

The detailed performance of detecting seeds is shown in Table 1. Here the missing is the number of direct pages which should not be removed from seed-set. The average number of clusters for persons and organization is 4.57 and 2.17 respectively. Precision is correct / (correct + incorrect + redundant); recall is correct / (correct + missing + redundant + incorrect). The performance of assigning indirect pages to clusters is shown in Table 2. There are only more than 50% indirect pages could be delivered to clusters and the other are discarded into unknown set. The latter are dispersed pages or lack the evidence to group their seeds. Because the number of indirect pages is large, we do not check the quality of their delivering, such as missing and incorrect. We will focus this in the further research.

We evaluated the performance of our clustering approach according to two aspects. One important factor is the quality of seeds. Missing a seed will reduce the number of clusters. While conserving a redundant seed will incur the pages about the same target entity scattering in more than one clusters. As shown in Table 1, the average ratio of the missing clusters and the redundant clusters is lower than 10%, which indicates that the seeds are stable and reliable. The other factor is the quality of entire clusters, as listed in Table 2. There are about 40% of pages that cannot be clustered into exact targets and are assigned to unknown-set. This may be caused by missing heuristic information, when some target entities do not have direct pages and the contents related to them are sparse in the Web.

Table 1. The performance of detecting seeds or direct pages for distinct target entities

Seeds	total	correct	incorrect	missing	redundant	precision	recall
Person	137	127	8	3	2	92.7%	94%
Org.	65	61	4	3	4	84.7%	89.7%
Overall	202	188	12	6	6	88.7%	91.8%

Table 2. The performance of assigning indirect pages

page	total num.	avg. num. in clusters	unknown	ratio of delivering
Person	3,600s	70s(*30)	1,500s	58.3%
Org.	9,800s	220s (*30)	3,200s	67.3%

The performance of this PnOs grouping approach is discussed as follows:

- **Effectiveness**: Compared with the results of manual clustering, outputs of computer do not always match human expectation[17]. Different people may produce different segmentations, and the case by computers is worse. For many clustering, though we could find out the reason for such segmentation, some of them still are specious. However, our approach is natural and its measurement is clear and relatively objective: pages about the same target should be grouped together. The seeds of clusters (direct pages) cover many items of targets and the relevance between direct page and query is highest. According to this measurement, there is little variable when search results are segmented. Hence, users could comprehend our clusters very well and accept it. Of course, once errors occur, especially when pages about different targets are arranged together, they could figure them out explicitly. In spite of that, the extensibility of this clustering approach is limited. For pages beyond PnOs, such as research paper, financial report, etc., the frequency of NEs is much lower. If there were represented by NEs, only few features should be nonzero, most of them could not be grouped together. At the same time, it is difficult to distinguish direct pages from indirect pages, so the number of clusters is unstable. On the contrary, for PnOs, there is obvious difference among them and detecting seeds is not sensitive to parameters. In a word, the testing users are satisfied with the segmentation by our approach.
- **Snippet or document**: If the clustering result on snippets returned by search engine is as good as that on document, we could save quite a little time and cost for downloading original pages. Like most clustering algorithm, our approach is sensitive to length of source and improper for snippets. This may be caused by snippets being short of NEs and fake NEs emerging. For example, if there is string *"…University of Singapore"* in the boundary of a snippet, a NE *University of Singapore* will be detected. In fact, it should be *National University of Singapore*.
- **Computational cost**: Account for response to users in time, the speed of clustering Web pages is pivotal. The computing time of our approach is nearly linear to the number of Web pages, the number of NEs among them (including URL and links) and the number of clusters, when the number of clusters is further less than that of pages. Furthermore, NEs could be recognized in the procedure of downloading pages. Determining direct page or indirect page could be proceeded at the same time if NEs are not normalized. In our experiment, the average time for clustering 1000 pages about PnOs using non-optimal algorithm is 34 seconds (PIII 900 and Men 512M).

7 Conclusion

PnOs are common query questions of users surfing the Internet. Our clustering approach, based on NEs can be used to group search results according to their target entities, with distinctive direct pages as the seeds of clusters. Although it is difficult to extend this approach to domains other than PnOs, it is certainly an effective approach for users to summarize information about targets and track their activity.

Further research is centered on: (a) Improving the effectiveness of the clustering method; (b) Incorporating learning to fine tune the parameters of the clustering approach; (c) Using the clustering results to perform information extraction for PnOs, and to facilitate user browsing; (d) Extracting the techniques to extract information in other domains.

References

[1] Elaine Marsh, Dennis Perzanowski, MUC-7 Evaluation of IE Technology: Overview of Result, (1998), http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7 _toc.html
[2] G. Salton, Automatic Text Processing. Addison-Wesley, New York, (1989).
[3] Pirolli, P. & Card, S. Information foraging in information access environments. In: Proc. of the Conf. on Human Factors in Computing Sys., (1995) 51-58.
[4] Daniel Boley, et al, Partitioning-based Clustering for Web Document Categorization, in: Decision Support System 27, (1999) 329-341.
[5] Oren Zamir and Oren Etzioni. Web document clustering: A feasibility demonstration. In: Proc. ACM SIGIR'98, (1998) 46-54.
[6] Oren Zamir, Oren Etzioni, Grouper: a dynamic clustering interface to Web search results, in: Computer Networks l31, (1999)1361-174.
[7] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, (1983).
[8] Bikel D.M., Schwartz R. & Weischedel R.M. An Algorithm that Learns What's in a Name, in: *Machine Learning,* 34(1-3), (1999)211-231.
[9] Sekine S. NYU: Description of The Japanese NE System Used for MET-2, in: MUC-7, (1998).
[10] Lin D. Using collocation statistics in information extraction, in: MUC-7, (1998).
[11] Borthwick A. *A Maximum Entropy Approach to Named Entity Recognition*. Ph.D. Thesis, New York Univ. (1999).
[12] Shiren Ye, Tat-seng Chua, Jimin Liu, Learning Pattern Rules for Chinese Named-Entity Extraction, COLING 02, Taiwan, (2002).
[13] Tat-seng, Jimin Liu, Learning Pattern Rules for Chinese Named Entity Extraction, AAAI02, Canada (2002).
[14] Mark Craven, Dan DiPasquo, et al, Learning to Extract Symbolic Knowledge from the WWW, in: Proc. of AAAI-98, Madison, USA, (1998) 509-516.
[15] http://www.comp.nus.edu.sg/~pris, (2002).
[16] Dmitri Roussinov and J. Leon Zhao, Automatic Discovery of Similarity Relationships through Web Mining, in: Decision Support Systems: Special Issue on Web Retrieval and Mining, (2002).
[17] D. Roussinov, H. Chen, Document Clustering for Electronic Meetings: An Experimental Comparision of Two Techniques, in: Decision Support Sys., (27)1-2, (1999) 67-79.

Personalized Services for Digital Library[*]

Chun Zeng[1], Xiaohui Zheng[2], Chunxiao Xing[1], and Lizhu Zhou[1]

[1]Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084, P. R. China
bobofu00@mails.tsinghua.edu.cn,
{xingcx, dcszlz}@tsignhua.edu.cn
[2]Library, Tsinghua University, Beijing, 100084, P. R. China
zhengxh@mail.lib.tsinghua.edu.cn

Abstract. In this paper, we describe a prototype system, called PASS (Personalized Active Service System), which provides personalized services for digital libraries. User profiles are represented as probabilistic distributions of interests over different domains. The system realizes content-based filtering by computing the similarity of probabilistic distributions between documents and user profiles. In addition the system realizes collaborative filtering by clustering similar user profiles. Experimental results show its performance satisfactory.

1 Introduction

A Digital Library is a massive source of information. It is a cumbersome task for a user to find the information relevant to his/her interests. Personalized service techniques solve this problem by providing different services for different users. There are generally three types of personalization systems: rule-based systems, content-based filtering systems and collaborative filtering systems [1]. In this paper, we describe a prototype, called PASS (Personalized Active Service System), which can essentially be viewed as a content-based collaborative system.

In practice, collaborative systems suffer from two fundamental problems: data sparsity and scalability. In this paper, we present our solution to the first problem. We manually build a number of training domains, and then classify all data items into these domains. In this way, the dimension of data decreases and the density of data increases.

2 User Profile

Associating a user with a set of profiles is a common technique in personalized information. Usually, user profiles are based on interests or behavior. In our system, user profiles are represented as probabilistic distributions of interests over different

[*] Supported by the National Natural Science Foundation of China under Grant No. 60221120146; the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 252-253, 2002.

domains. In contrast to vector space models, the probability model is more effective. In fact, vector space models pay attention to strict matching between words using cosine measure. To avoid strict matching, we compute the similarity between probabilistic distributions over a domain structure.

Suppose C={c_1, c_2, …, c_n} is the set of domain structures, where c_j is the jth domain. Then the user profile of a user u_i is defined as: Profile=[p(c_1 | u_i), p(c_2 | u_i), …, p(c_n | u_i)]. To compare a document and a user profile, the representation of the document is consistent to the user profile. We use relative entropy or Kullback-Leibler distance to compute the similarity between probabilistic distributions.

When a user profile is created, the system dynamically modifies the profile by tracking the user's behavior. When retrieved results are provided for the user, he/she can browse interesting results, add bookmarks, and so on. Then the system will adjust the probabilistic distribution of the user's profile.

3 Performance Evaluation

We load into the system more than 5000 abstracts of papers on computer science from the INSPEC database. We first build a domain structure of computer science using 2000 abstracts. The size of the domain is 100. Then we classify all documents into the domain structure. We invite more than 20 students from our lab to do the experiment. When a user enters the system, he/she can manage his/her personal information and bookmarks. The user can do personalized searching and ask for recommendations. By tracking the user's behavior, the system dynamically modifies his/her profile. Fig. 1 shows the comparison between two types of representation of user profiles. In contrast to the vector space model, the probability model is more effective.

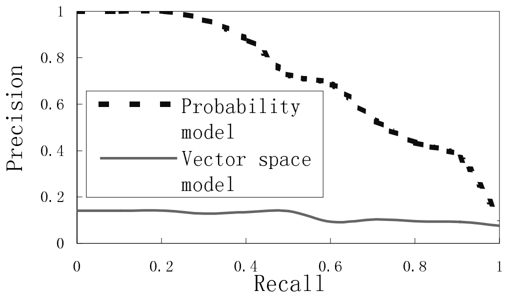


Fig. 1. Comparison between two types of representation of user profiles

References

1. Mobasher, B., Colley, R., and Srivastava, J. Automatic personalization based on Web usage mining. *Communications of the ACM*, Aug. 2000, 43(8), 142-152.

Automatic References: Active Support for Scientists in Digital Libraries

Harald Krottmaier

Institute for Information Processing and Computer Supported new Media (IICM)
Inffeldgasse 16c, A-8010 Graz, Austria
hkrott@iicm.edu

Abstract. Scientists need automatic support from digital library systems when writing papers about a new topic. In this paper we introduce a concept to be integrated in the Journal of Universal Computer Science (J.UCS), where scientists will be able to upload an abstract or paper outline he has written, and the system will return a list of "must-read" papers related to the topic of the uploaded document.

1 Introduction, Problems, and a Possible Solution

Traditional libraries have a few very sophisticated features still not available in most digital libraries. While some problems of traditional libraries are solved in many digital library systems, such as searching in fulltext or personalization of the material stored in those systems, the most sophisticated "feature" in traditional libraries is still not available: the librarian!

The role of librarians is often underestimated by most of the users. Especially for young scientists, who are trying to explore a new topic, librarians are very helpful. They are able to answer fuzzy questions or help with formulating "search queries" by asking related questions. This interactive discussion is very helpful for a scientist when trying to find the right "starting-point" of a new work, i.e. to form the right list of references. Knowledge management tools are widely available and must be integrated in existing digital library environments.

In the following, we list tools to find the right references. Please note the big problem with common search engines: they index freely available resources, not high quality publications which are often available only to subscribers of the corresponding journals. Therefore many high quality publications are not accessible via search engines.

- *Search engine* usage is an every day task. It is possible to formulate a search query, but it is very difficult to formulate the right query. There are very often thousands of results or none. It is well known that search engines are very powerful if someone knows exactly what to search for but this is not the case in our application. We want to find the right references when the topic is not clearly stated.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 254-255, 2002.

- *GoogleScout* is a very powerful tool implemented in the google search engine (Google, 2002). GoogleScout is an implementation of the "search for similar pages" feature without having to worry about selecting the right keywords.
- *Related conference proceedings and SIGs* are very often the right starting point of researching a topic and finding the "big players" in the topic. These proceedings are very often also available in electronic form and are therefore easily searchable. Nevertheless, current publishing systems provide the user with a feature like "find related articles to an already published article" but not "find related articles to some arbitrary abstract or document outline"!
- *Digital libraries* like ACM/DL or IEEE/Xplore do support registered users with a "find related/similar articles"-feature. They consider already published articles but not arbitrary, user-defined abstracts or outlines.
- *ResearchIndex* also supports scientists in finding the right reference list with rated resources (Lawrence et al., 1999).

We are going to prepare an uploading area in the Journal of Universal Computer Science (J.UCS, 2002). Scientists will be able to upload an article abstract or outline in different electronic formats (such as PDF, PostScript, MS-Word, HTML, XML, etc.). The system will find related articles using similarity matching algorithms implemented in the used search engine.

At the moment there are approximately 600 peer-reviewed articles available in the system. These articles were published during the last 8 years. Because of strict peer-reviewing in J.UCS, the quality of the articles is very high compared to most of the freely available systems. We are convinced that scientists working in the area of computer science will consider this unique feature valuable, and hope that other digital library systems will incorporate this feature into their systems.

2 Conclusion and Future Work

An every day problem of information exploration for scientists is to find related or similar documents to some given article outline. Many digital library systems provide registered users with similarity searching to already published articles. Currently, it is not possible in well known publishing systems to find related articles to an arbitrary, user-defined abstract. We suggest incorporating this feature into currently available systems, and are going to implement a prototype in the Journal of Universal Computer Science (J.UCS).

References

1. Google (2002). http://www.google.com .
2. J.UCS (2002). Journal of Universal Computer Science. http://www.jucs.org .
3. Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6), 67-71.

Organizing and Maintaining Dynamic Digital Collections

M. Nikolaidou[1], D. Anastasiou[2], D. Anagnostopoulos[2], and M. Hatzopoulos[1]

[1] Department of Informatics,
University of Athens
Panepistimiopolis, 15771 Athens, Greece
{mara,mike}@di.uoa.gr
[2] Department of Geography
Harokopion University of Athens
70 El. Venizelou Str, 17671, Athens, Greece
{danasta,dimosthe}@hua.gr

1 Introduction

To provide access to research material for researchers and physicians of the Athens Medical School (AMS), a Digital Library system (AMS DL) was developed. It maintains medical images produced by the laboratories of the School. End-users access the system through the Web. Images are catalogued and processed partially by laboratory scientific staff and partially by cataloguers in the Central Library of Health Sciences. The AMS DL system is based on a multi-tiered client-server model and is implemented using Java and the *IBM Content Manager* platform [1]. Since the system should be open, it supports existing standards regarding the metadata scheme and communication interfaces. Thus, the metadata scheme adopted is an extension of Dublin Core, while results are obtained as XML pages [2]. As laboratory requirements differ significantly, it was decided to develop a different collection for each laboratory. Since the number of collections needed is not static and predefined, we have identified two requirements: a. the need to easily create new collections and b. the need to extend or modify *collection description* (e.g. the metadata information used to characterize the images included in each collection). We introduced the term *dynamic collection management* to denote the support of automated collection definition and management within an integrated digital library environment.

2 Dynamic Collection Management

This concerns the creation of new collections and the modification of collection descriptions. For each collection added in the digital library environment, the corresponding object structure and metadata model must be defined. Collection descriptions can be derived from existing collections by extending the object structure and metadata model, e.g. a collection description can be defined as the descendent of an existing collection description, while additional object parts and metadata fields can also be defined. This feature allows flexibility during collection definition and facilitates collection description in a simplified manner. In the case of AMS DL, each collection consists of the medical images produced by a specific laboratory. Thus, they are characterized by common general metadata and domain specific metadata, useful for researchers in the specific domain. The general metadata model and a basic

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 256-257, 2002.

object structure corresponding to *medical image objects* are used to describe the *Medical Image Collection*. The Medical Image Collection Description is used to easily define collections corresponding to each laboratory as its descendents, while the collection is practically empty. In order to efficiently support dynamic collection management, AMS DL facilitates dynamic interface creation for processing and cataloguing and collection search. The same interface is used for all collections, while screens are dynamically formed based on collection descriptions.

3 Data and Metadata Representation

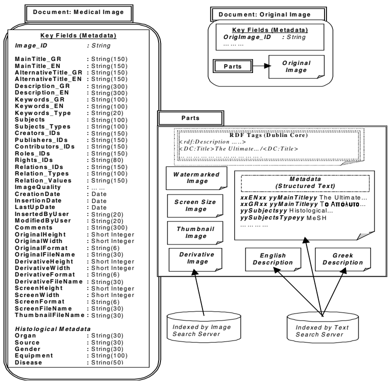
The following are included in the *medical image objects* in the collections: *Original Image*, a high quality image with strong copyright protection; *Derivative Image*, produced from the original image to be accessed over the Web; *Watermarked Image*; *Screen Size and Thumbnail Image*; and *Image Description* in Greek and English. The original image and the description are produced by the researcher, while all other formats are produced by the cataloguer.

The Dublin Core metadata scheme is used to describe general metadata information [2]. Additional fields are used to represent domain-specific metadata. Since these fields are collection specific, we adopted a DC-like structure for their representation in XML, where the collection is depicted similar to a DC qualifier:

```
<AMS:local_field_name>
  <AMSq:collection>collection_name/< AMSq:collection>
    <rdf:value>local_field_value/<rdf:value>
/< AMS:local_field_name>
```

Medical Image data and metadata internal representation using Content Manager constructs are presented in Figure 1. The digital object used to represent Medical Image Objects consists of all derived images and image descriptions. Since the system must support both exact and approximate search in combined multi-valued metadata fields, the searching capabilities of a relational database are too poor to ensure Collection Search application performance. Thus, database search is applied for exact numerical and date metadata field search, while exact string and approximate search are performed using free text search capabilities. Metadata information is stored both in the underlying database and within a tag-structured text part in the corresponding Medical Image digital object *(metadata part)*. Different tags are used to support Greek and English, while all properties of a specific metadata field are included within one tag.

1. Darmoni SJ et.al: CISMeF. A Structured Health Resource Guide. Methods Information Medicine 39 (2001)
2. Weibel S., Hakala J.:DC-5. The Helsinki Metadata Workshop. D-Lib Magazine 4 (1998).



Key Fields (Metadata)

Image_ID	*: String*
MainTitle_GR	: String(150)
MainTitle_EN	: String(150)
AlternativeTitle_GR	: String(150)
AlternativeTitle_EN	: String(150)
Description_GR	: String(300)
Description_EN	: String(300)
Keywords_GR	: String(100)
Keywords_EN	: String(100)
Keywords_Type	: String(20)
Subjects	: String(100)
Subjects_Types	: String(100)
Creators_IDs	: String(150)
Publishers_IDs	: String(150)
Contributors_IDs	: String(150)
Roles_IDs	: String(150)
Rights_IDs	: String(80)
Relations_IDs	: String(150)
Relation_Types	: String(100)
Relation_Values	: String(150)
ImageQuality	: … …
CreationDate	: Date
InsertionDate	: Date
LastUpDate	: Date
InsertedByUser	: String(20)
ModifiedByUser	: String(20)
Comments	: String(300)
OriginalHeight	: Short Integer
OriginalWidth	: Short Integer
OriginalFormat	: String(6)
OriginalFileName	: String(30)
DerivativeHeight	: Short Integer
DerivativeWidth	: Short Integer
DerivativeFormat	: String(6)
DerivativeFileName	: String(30)
ScreenHeight	: Short Integer
ScreenWidth	: Short Integer
ScreenFormat	: String(6)
ScreenFileName	: String(30)
ThumbnailFileName	: String(30)
Histological Metadata	
Organ	: String(30)
Source	: String(30)
Gender	: String(30)
Equipment	: String(100)
Disease	: String(50)

Fig. 1. Medical Image Data and Metadata Representation

Navigation, Organization, and Retrieval in Personal Digital Libraries of Email

Benjamin M. Gross

University of California Berkeley, SIMS
University of Illinois Urbana-Champaign, GSLIS
bgross@uiuc.edu

Abstract. Electronic mail remains the dominant application for the Internet and is the most ubiquitous type of personal digital library collection. Existing research on email includes studies of use, visualization, categorization and retrieval. [1] [4] [5] [2] This paper makes two contributions. First, I describe user interviews that revealed problems in current email systems, including role conflict, high cognitive overhead associated with organization and retrieval, inability to navigate conversations, and difficulties in addressing messages. Next, I describe a prototype system that addresses many of these problems through improvements in the message store, query interfaces, authority control, and support for identity and roles.

1 Interviews of Email Use

I conducted in-depth user interviews that revealed problems in current email systems. I interviewed twelve users (six novice, six expert; five male, seven female; with education ranging from high school diploma to Ph.D. candidate). Most users in my study maintain multiple email addresses in order to "act" in multiple "roles," ranging from two to dozens. For example, the majority of users maintain separate email addresses for work and personal communication. The most typical mechanism to cope with a large number of email addresses is to forward multiple addresses to a smaller number of addresses. One difficulty is "role conflict." For example, users consistently reported being embarrassed by mailing a professional contact from a personal address.

Nearly all classification mechanisms require users to place messages into fixed categories. This creates a burden on the user to choose the "correct" categories to file messages and to then remember later, in order to retrieve the messages. Users typically locate messages through sorting by name or date and then browsing to find the desired item. Users want to search for messages by a person's name, not his email address. A problem is that there is no way to reference an individual consistently over time, as his email address and name may change.

Because sent mail is saved in a separate folder, a message and its response are hard to display together. To reconstruct conversations, users typically must go back and forth between their sent mail, inbox and folders to correlate messages.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 258–259, 2002.

2 System Design and Future Work

The prototype system I am developing has an underlying message store that adds substantial improvements for organization, retrieval, addressing and navigation. Messages are stored in a database with a full text index. Rather than creating a folder on the file system, categories are created by queries. Most queries are handled by a simple interface with an advanced interface for creating complex queries and for editing queries. The benefit is that users no longer have to manually categorize messages in order to organize them. Instead of filing messages into fixed categories, users simply add metadata to create categories. Categories are views of the collection, allowing messages to be in multiple and overlapping categories.

Most modern email applications have "roles" or "personality" functionality. However, these personalities and roles cannot be used for organization or retrieval. The prototype system includes a notion of an individual person, each with a locally unique identifier within the email collection. The unique identifier allows senders and recipients to have a consistent "identity" comprised of multiple facets: name forms, email addresses, roles, contact information, notes, etc. This form of authority control is useful for mapping multiple name and address entries into one entry for retrieval [3]. Since the most common method to file messages is by sender, the prototype system automatically generates a category for each identity. This reduces the amount of categorization users have to do.

By attaching role information to an identity, the system is able to perform "role matching." For example, if a user sends a message using a personal role to someone who is both a friend and a coworker, the application will use the recipient's personal email address by default. Another advantage of using a canonical identity is that it improves the reconstruction and display of threads by allowing messages both sent to and received from a person to be viewed at once.

In the future, I will conduct performance evaluation, including user studies, to test improvements in the prototype. Performance evaluation will include precision, recall, speed and efficiency comparisons. User studies will include an analysis of organization versus retrieval time, ease of use, and effectiveness of interfaces.

References

1. Olle Bälter. *Electronic Mail in a Working Context.* PhD thesis, Royal Institute of Technology, IPLab, NADA, KTH, 10044 Stockholm, 1998.
2. Aleksandra Jovicic. Retrieval issues in email management. Master's thesis, University of Toronto, 2000.
3. F. W. Lancaster. *Vocabulary Control for Information Retrieval.* Information Resources Press, Arlington, VA, second edition, 1986.
4. Ann Lantz. Heavy users of electronic mail. *International Journal of Human-Computer Interaction*, 10(4):361–379, 1998.
5. Steve Whittaker and Candace Sidner. Email overload: Exploring personal information management of email. In *Proceedings of ACM CHI 96*, pages 276–283, 1996.

A Work Environment for a Digital Library of Historical Resources[1]

Dion Goh, Lin Fu, and Schubert Foo

Division of Information Studies
School of Communication and Information
Nanyang Technological University
Singapore 637718
{ashlgoh,p148934363,assfoo}@ntu.edu.sg

Abstract. A work environment for the National Archives of Singapore (NAS) digital library is described. The system provides an integrated platform for supporting scholarly tasks such as browsing, querying, organizing and annotating of resources using a spatial metaphor. When complete, the system will allow users to build and share personal collections of resources as well as author new resources of various types through application plug-ins.

1 Introduction

The NAS digital library project [1] aims to build a web-based digital library of historical resources offering a range of services (e.g. searching) and applications (e.g. virtual exhibitions) to users and staff of the NAS. The interfaces currently being developed are HTML-based to cater to most browsers in use today and hence allow more users access to the digital library. A shortcoming is that user interaction is limited, and scholarly tasks [2] such as annotating and organizing of resources become cumbersome. We are thus exploring a direct manipulation interface that will form the basis of a work environment for users of the NAS digital library.

A prototype of the work environment is shown in Figure 1. There are three major components of the system. The *work area* is the focus of the system. It employs a spatial metaphor and functions like a desktop in which users organize their resources. Users may also add annotations and view resource metadata by right-clicking on a resource and selecting an appropriate option from the popup menu. The size of the work area is arbitrarily large and panning and zooming is supported for navigation.

The *browsing tree* provides access to the collection through a hierarchical organization of resources in which leaf nodes represent resources while non-leaf nodes represent categories and subcategories to an arbitrary depth. Resources may be viewed by double-clicking or dragging-and-dropping on the work area. The *query area* supports a variety of search options through a single text field. It is deliberately kept simple to reduce clutter on the interface. A list of search results appear below the text field and resources are viewed in a similar fashion as in the browsing tree.

[1] Resources used in this project were obtained in collaboration with the National Archives of Singapore.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 260-261, 2002.

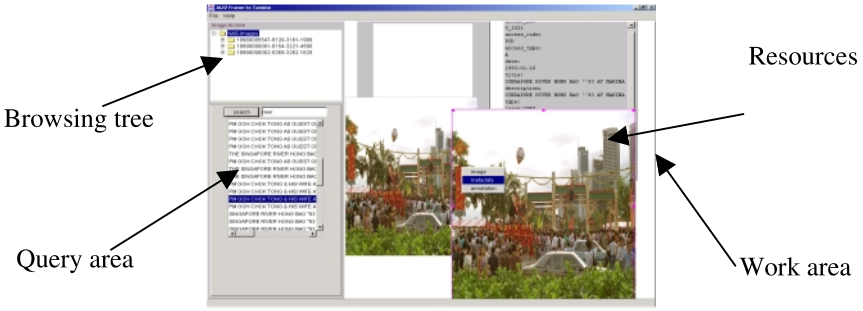


Fig. 1. The NAS digital library work environment

The work environment is an integrated platform for users to browse, query, organize and annotate resources. In addition, users can build personal collections of resources through the saving and reloading of the contents of their work areas. Various authoring tasks may also be performed. These are done through various *application plug-ins* which are components that allow users to author application-specific resources. Finally, resources may also be shared with other users. Most interfaces to digital libraries support individual access and manipulation of resources. We are designing the work environment to be collaborative so that users may share and manipulate personal collections with other users, and in the process, create communities within the digital library.

2 Conclusion and Future Work

The work environment is an initial step in the development of an integrated platform for supporting a variety of scholarly tasks in the NAS digital library. Support for basic digital library services is nearly complete. The next phase in the development of the system will focus on the design and implementation of a framework for application plug-ins. Testing of the initial design of the work environment will also be conducted in parallel to assess the usability of the system in supporting scholarly work.

References

1. Goh, D., and Foo, S. (2002). Developing a digital library of reusable historical artifacts. Proceedings of the Second Joint Conference on Digital Libraries (Portland, OR, July 2002), 394.
2. Goh, D., and Leggett, J. (2000). Patron-augmented digital libraries. In Proceedings of the Digital Libraries 2000 Conference (San Antonio, TX, June 2000), 153-163.

A Personalized Collaborative Digital Library Environment

M. Elena Renda and Umberto Straccia

I.S.T.I. – C.N.R.
Via G. Moruzzi,1 I-56124 Pisa (PI) ITALY
{renda,straccia}@iei.pi.cnr.it

Abstract. We envisage a Digital Library not only as an information resource where users may submit queries to satisfy their information needs, but also as a collaborative working and meeting space. We present a personalized collaborative Digital Library environment, where users may organise the information space according to their own subjective view, become aware of each other, exchange information and knowledge with each other, build communities and get recommendations based on preference patterns of other users.

1 Introduction

It is widely recognised that the internet is growing rapidly in terms of the number of users accessing it, the number of *Digital Libraries* (DLs) created and accessible through it, and the number of times users use them in order to satisfy their information needs. This has made it increasingly difficult for individuals to control and effectively look for information among the potentially infinite number of DLs available on the internet.

Typically, DLs provide a search service to the web community at large. A common characteristic of most of these retrieval services is that *they do not provide any personalized support to individual users*, or poorly support them. Indeed, they are oriented towards a generic user, as they answer queries crudely rather than, learn the long-term requirements of a specific user. In practice, users use the same information resource again and again, and would benefit from customization. The time consuming effort that the user puts in searching for documents and possibly downloading them from the DL is often forgotten and lost. Later, the user may wish to perform a search on the same topic to find relevant documents that have since appeared. This requires a repetition of the same laborious searching and browsing, just like the first time.

Additionally, users are highly interested in being able to organize the information space according to their own subjective perspective (see *e.g.* [8,13]). The requirement of personalized search in the context of DLs is already known and some DLs provide related functionality (see *e.g.* [4,8,11,13,14,17,19,21]). Many of them fall in the category of *alerting services*, i.e. services that notify a user (through e-mail), with a list of references to new documents deemed as

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 262–274, 2002.

relevant. Typically, all these services are based on the notion of *user profile* (a machine representation of the user's information need). It can be created either automatically (by user-system interaction) or set-up manually (by the user). The combination of a user profile and the successive matching of documents against it, in order to *filter out* the relevant ones, is known as *Information* or *Content-based Filtering* [2,12].

Very seldom, except *e.g.* [8], DLs can also be considered as *collaborative meeting* or *working places*, where users may become aware of each other, open communication channels, and exchange information and knowledge with each other or with experts. Since a user usually accesses a DL in search of some information, it is quite probable that some other user(s) may have information needs and interests that overlap with his. Such users may well profit from each other's knowledge, opinions, experiences and advice. Some users may enter into long-term relationships and eventually evolve into a community. Such a service may be important for a DL to offer as it supplies very focussed information. Concerning the information seeking task, the *recommendation* of items based on the preference patterns of other users is probably the most important one. The use of opinions and knowledge of other users to predict the relevance value of items to be recommended to a user in a community is known as *Collaborative* or *Social Filtering* [3,5,15,16,20]. Such filtering is based on the assumption that a good way to find interesting content is to find other users who have similar interests, and then recommend items that those other users like. In contrast to information filtering methods, collaborative filtering methods do not require any content analysis as they are based on aggregated user ratings of these items.

Both types of filtering share the common goal of assisting in the users' search for items of interest, and thus attempt to address one of the key research problems of the information age: locating relevant information in a haystack that is growing rapidly. Providing personalized information organisation and search in the context of a collaborative DL environment, as additional services to the uniform and generic information searching offered today, is likely to be an important step to make relevant information available to people with minimal user effort [1].

The contribution of our paper is as follows. We will (*i*) formalise an abstract collaborative DL environment, where users and communities may search, share and organize their information space according to their own views; (*ii*) present an instance of such an environment within the EU funded project CYCLADES[1]; and (*iii*) for completeness, sketch out the recommendation algorithms. The underlying techniques used for recommendation fall in the afore mentioned filtering methods.

The outline of the paper is as follows. In the next section we will formalise the main concepts of our personalized collaborative DL environment. In Section 3, we will present CYCLADES, while in Section 4, the recommendation methods will be presented briefly. Finally, Section 5 concludes.

[1] www.ercim.org/cyclades

2 A Personalized Collaborative DL Environment

Our personalized collaborative DL environment is made up of several concepts: *actors*, *objects* and *functionality*. Actors will be able to act on objects by means of the DL's functionality. At first, we will give a brief overview of the environment we have in mind and then move on to its formalisation. Roughly, our collaborative environment is based on the *folder paradigm*, *i.e.* users and communities of users may organise the information space into their own folder hierarchy, as may be done with directories in operating systems, bookmark folders in Web browser and folders in e-mail programs. The idea of organising the information space within DLs into folders is not new. For instance, in [8] users are allowed to define folders of bookmarks (*i.e.* URLs). A folder becomes a holder of information items, which are usually semantically related and, thus, implicitly determines what the folder's topic is about. Therefore, rather than considering user profiles, we will deal with *folder profiles*, *i.e.* representations of what folders are *about*. The user's set of folder profiles represents the set of topics she is interested in.

2.1 Actors

We will distinguish between two types of *actors*: the set $\mathcal{U}$ of *users* u and the set $\mathcal{C}$ of *communities* C. A community may be seen as a set of users sharing a common (scientific, professional) background or view of the world. In particular, communities are characterised by a shared interest in the information made available. We postulate that a community $C \in \mathcal{C}$ has a membership function $\mu_C{:}\mathcal{U} \to \{0,1\}$, where $\mu_C(u) = 1$ (for ease $u \in C$) indicates that the user u belongs to the community C. We do not require a user to belong to at least one community, *i.e.* we assume that it is a user's choice to join a community or to leave it. A user may belong to different communities as well. It is not our purpose to address the issues of how a community may be created and which are the policies to join or to leave it. We simply assume that there is a *community administrator* (a user $u^C \in \mathcal{U}$) for each community $C \in \mathcal{C}$, who is in charge of defining these policies. (We will also not address the issue of becoming a community administrator within the environment.)

2.2 Objects

We will consider three types of *objects* which may be managed within the environment by users and communities: data items, collections and folders. The objects are organised according to a multilevel model (see Figure 1).

Data Items. At the lowest level, we have the set $\mathcal{D}$ of *data items* d. $\mathcal{D}$ is the information space and the data items are the information resources that a user is usually interested in discovering or searching for within the DL. The data items may be *e.g.* papers, reports, journals, proceedings, notes, annotations, discussions, URIs. A data item may also be just a metadata record, which consists of a set of attributes and related values specifying features of a document, according to a specific schema, *e.g.* Dublin Core [9]. The set of data items $\mathcal{D}$ may well be distributed, heterogeneous in content, format and media (video, audio).

Collections. At the next higher level, we allow the data items $d \in \mathcal{D}$ to be grouped into *collections*. A collection may be seen as a set of data items, which are grouped together according to some relatedness criteria, *e.g.* the set of data items created within the same year, or those created by the same author, or those about the same topic (say "collaborative digital libraries"), or, more obviously, the set of data items belonging to the same digital archive. We assume that there is a set $\mathcal{L}$ of collections L and a membership function $\mu_L \colon \mathcal{D} \rightarrow \{0,1\}$, where $\mu_L(d) = 1$ (for ease $d \in L$) indicates that the data item d belongs to the collection L. We also assume that there is at least one collection in $\mathcal{L}$, called *universal collection* and denoted $L_\top$, which includes all the data items $d \in \mathcal{D}$. Note that a data item may belong to several collections. Furthermore, we do not specify whether the collections are materialised or are just "views" over $\mathcal{D}$. This does not play a significant role in our context. Finally, as for communities, we will assume that for each collection $L \in \mathcal{L}$ there is a *collection administrator* (a user $u^L \in \mathcal{U}$), who is in charge of defining both the collection L and the access policies to it.

Folders. At the third level, we have *folders*. A folder is a container for data items. A folder should be seen as the main environment in which users will carry out their work. Folders may be organised by users according to their own folder hierarchy, *i.e.* a set of hierarchically organised folders, each of which is a repository of the user's selected data items. Each folder typically corresponds to one subject (or discipline, or field) the user is interested in, so that it may be viewed as a thematic repository of data items. In order to accomplish a truly personalized interaction between user and system, this correspondence is implemented in a way which is fully idiosyncratic to the user. This means that *e.g.* a folder named `Knowledge Representation and Reasoning` and owned by user `Tim` will not correspond to any "objective" definition or characterisation of what "knowledge representation and reasoning" is, but will correspond *to what Tim means by* "knowledge representation and reasoning", *i.e.* to his personal view of (or interest in) "knowledge representation and reasoning". As we will see later on, this user-oriented view of folders is realised by learning the "semantics of folders" from the current contents of the folders themselves. We will allow two types of folders: (*i*) *private folders*, *i.e.* a folder owned by a user only. This kind of folder can only be accessed and manipulated by its owner. For other users, they are invisible; and (*ii*) *community folders*, which can be accessed and manipulated by all members of the community who owns the folder. Community folders are used to share data items with other users and to build up a common folder hierarchy. Community folders may also contain *discussion forums* (a kind of data item), where notes may be exchanged in threaded discussions (similar to news groups). Formally, we assume that there is a set $\mathcal{F}$ of (either private or community) folders F. For each user u, with $\langle \mathcal{F}^u, \preceq^u \rangle$, we indicate the user's folder hierarchy, where $\mathcal{F}^u \subseteq \mathcal{F}$, $\preceq^u$ is a tree-like order on $\mathcal{F}^u$ and with $F^u_\top$ we indicate its *home folder* or *top folder*, *i.e.* the root folder of the hierarchy

$\langle \mathcal{F}^u, \preceq^u \rangle$. Furthermore, given a folder $F \in \mathcal{F}$, we assume that (i) there is a membership function $\mu_F\colon\mathcal{U} \to \{0,1\}$, where $\mu_F(u) = 1$ (for ease $F \in u$) indicates that the folder F belongs to the folder hierarchy of user u, *i.e.* $F \in \mathcal{F}^u$; (ii) there is a membership function $\mu_F\colon\mathcal{C} \to \{0,1\}$, where $\mu_C(d) = 1$ (for ease $F \in C$) indicates that the folder F is a community folder and belongs to the community C; and (iii) there is a membership function $\mu_F\colon\mathcal{D} \to \{0,1\}$, where $\mu_F(d) = 1$ (for ease $d \in F$) indicates that the data item d belongs to the folder F. Figure 1 shows an example of community, users and object organisation. In it, users u_1 and u_2 belong to the same community C_1. User u_2 has no private folders, while F_4 and F_5 belong to the same community C_1.

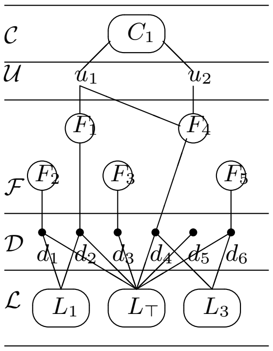


Fig. 1. Personalized information space organisation.

2.3 Actions

A user may perform a set of actions (see below), depending on whether she is a member of a community or not, and whether she is a collection or a community administrator. At any time, the user performs her actions with respect to (w.r.t.) the *current folder*. At the beginning, this is the user's home folder.

Folder management. A user can perform basic management actions on the folders she has access to. (i) w.r.t. "folder hierarchy", folder management operations include creating a new folder as a child of an existing folder, deleting, moving a folder from an existing parent folder to a new parent folder (community administrators are allowed to manage the folder hierarchy of a community). (ii) w.r.t. "folder content", folder management actions include saving data items from a search session in folders (see below), deleting, undeleting and destroying data items, moving and copying data items from one folder to another, rating, annotating, downloading and uploading data items.

Collection management. A collection administrator can create, edit, delete and define the access policies of collections. New collections may be defined in terms of others, *e.g.* using meet, join and refinement operators.

Collaborative support. Collaboration between users is supported through the possibility of sharing community folders along with their contents and folder structure. Discussion forums may be created within folders to allow informal exchange of notes and arguments. Rating and annotation of data items also may take the form of discussions among the members of a community. In order not to loose shared activity in the collaborative DL environment, mutual awareness may be supported through event icons (a kind of data item) displayed in the environment. Activity reports that are daily received by email may also be possible. Also, users may view the list of all existing communities so that they become aware of ongoing community activity. This does not mean that they can look inside communities, but can see only *e.g.* the title, the description and the identity of the community administrator. To become a member, users may directly join the community if this is allowed by the community's policy, or may contact the administrator to be invited to the community. In summary, collaboration support is concerned with inviting members to join or removing them from a community, leaving, viewing and joining a community (only for communities open to subscription), contacting community managers or other users (*e.g.* via email), creating discussion forums, adding notes to a discussion forum, editing event notification preferences (icons, daily report) and rating data items.

Searching of data items. The user can issue a query q, whose result is a partial order (the result list) on the data items $d \in \mathcal{D}$. The user is allowed to store selected data items of the result list within her folder hierarchy. In an *ad-hoc search*, a user u specifies a query q and the action of the system will be to look for relevant data items within a set of user specified folders $F_i \in \mathcal{F}^u$ she has access to, *i.e.* to search within $\{d \in \mathcal{D}\colon d \in F_i\}$, or to search within a specified collection C, *i.e.* $\{d \in \mathcal{D}\colon d \in C\}$. (We do not specify the syntax of queries, which depends on the indexing capabilities of the underlying DL.) We further allow a kind of *filtered search*. This is like an ad-hoc search, except that the user u specifies a query q and a folder $F \in u$, and the action of the system will be to look for data items $d \in \mathcal{D}$ such that d is relevant both to the query and to the folder F. For both types of search, there exists widely known methods. Adhoc searching is the usual task of information retrieval (see [22]), while filtered searching may be accomplished in at least two ways: (i) through techniques of query expansions [7], *i.e.* we expand the query q with significant terms of the folder profile f of F and then submit the expanded query; or (ii) we first issue the query q as an ad-hoc query, and then filter the result list w.r.t. the folder profile [2,6,12,18].

Recommendation. A user may get recommendations of *data items*, *collections*, *users* and *communities*, based on other users' (implicit or explicit) ratings, and on the perceived similarity between the interests of the user, as represented by a given folder, and the interests of other users, as represented by their folders. All recommendations are specific to a given user folder, *i.e.* they have always to be understood in the context not of the general interests of the user, but of the specific interest(s) (topic(s)) of the user represented by the folder.

Without doubt, the above set of actions provides us with an enhanced personalized collaborative DL environment. Several of the above items can be further investigated, but we will not address them further.

3 An Application: CYCLADES

The model of the personalized collaborative DL environment we have presented is currently under implementation in the CYCLADES system. The main goal of CYCLADES is the development of a system which provides an open collaborative virtual archive environment, which (among others) supports users, communities (and their members) with functionality for (*i*) advanced searching in *large, heterogeneous, multidisciplinary digital archives*; (*ii*) collaboration; and (*iii*) filtering and recommendation. With respect to the model described in Section 2, a main feature of CYCLADES is that it will use the protocol specified by the Open Archives Initiative[2] (OAI) to harvest and index metadata records from any archive that supports the OAI standard. As a consequence, the set $\mathcal{D}$ of data items includes the set of metadata records harvested from OAI compliant archives. As a reminder, the OAI is an agreement between several Digital Archives in order to provide interoperability. The specifications give *data providers* (individual archives) easy-to-implement mechanisms for making the documents' metadata records in their archives externally available. This external availability then makes it possible for *service providers* to build higher levels of functionality. The CYCLADES system is indeed such a service provider. From a logical point of view we may depict the functionality of the CYCLADES system as in Figure 2, which highlights the functionality related to *collaboration, search, filtering and recommendation* of data items grouped into collections. Figure 3 shows a mock-up of the user interface, while Figure 4 shows its architecture.

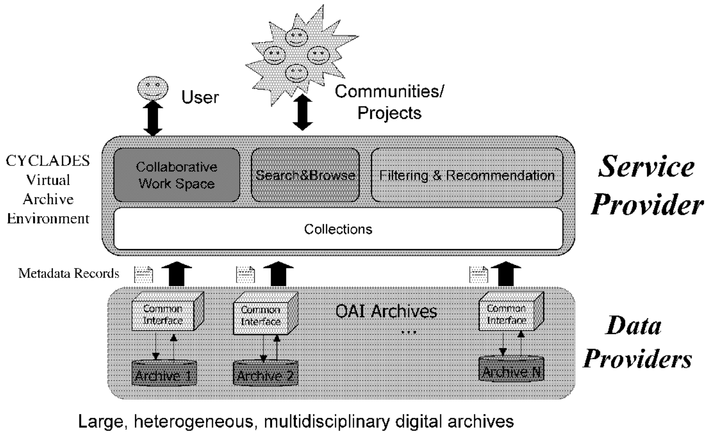


Fig. 2. Logical view of CYCLADES functionality.

It should be noted that from an architecture point of view, each box is a Web service distributed over the internet. The CYCLADES system, which will be accessible through Web browsers, provides the user with different environments, according to the actions the user wants to perform.

[2] www.openarchives.org

The *Collaborative Work Service* provides the folder-based environment for managing metadata records, queries, collections, external documents, ratings and annotations. Furthermore, it supports collaboration between CYCLADES users by way of folder sharing in communities.

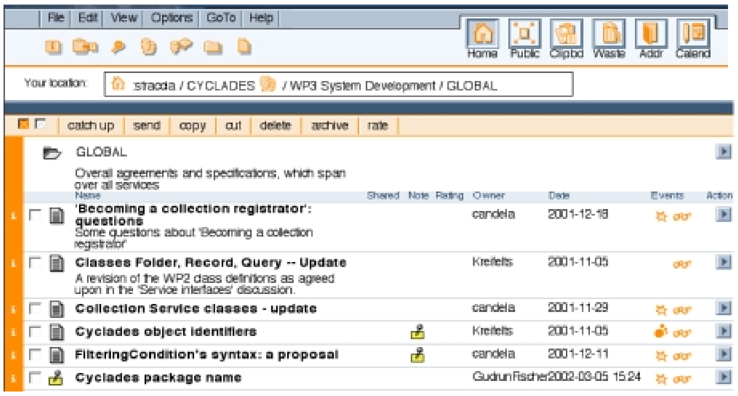
The *Search and Browse Service* supports the activity of searching records from the various collections, formulating and reusing queries, and browsing schemas, attribute values, and metadata records.

The *Access Service* is in charge of interfacing with the underlying metadata archives. In this project, only archives adhering to the OAI specification will be considered. However, the system can be extended to other kinds of archives by modifying the Access Service only.

The *Collection Service* manages collections (*i.e.* their definition, creation, and update), thus allowing a partitioning of the information space according to the users' interests, and making the individual archives transparent to the user.

The *Filtering and Recommendation Service* provides filtered search, recommendations of records, collections, users, and communities.

The *Mediator Service*, the entry point to the CYCLADES system, acts as a registry for the other services, checks if a user is entitled to use the system, and ensures that the other services are available only after proper authentication. All of these services interoperate in a distributed environment. Security and system administration will be provided for centrally (by the Mediator Service). The CYCLADES services can run on different machines, and will only need a HTTP connection to communicate and collaborate.



Your location: stracda / CYCLADES / WP3 System Development / GLOBAL

GLOBAL

Overall agreements and specifications, which span over all services

Name	Shared	Note	Rating	Owner	Date	Events	Action
'Becoming a collection registrator': questions Some questions about 'Becoming a collection registrator'				candela	2001-12-18		
Classes Folder, Record, Query -- Update A revision of the WP2 class definitions as agreed upon in the 'Service interfaces' discussion.				Kreifelts	2001-11-05		
Collection Service classes - update				candela	2001-11-29		
Cyclades object identifiers				Kreifelts	2001-11-05		
FilteringCondition's syntax: a proposal				candela	2001-12-11		
Cyclades package name				GudrunFischer	2002-03-05 15:24		

Fig. 3. User interface (mock-up).

4 Recommendation Algorithms

A consequence of our proposed enhanced DL environment is that, (*i*) by allowing users to organise the information space according to their own subjective view, and (*ii*) by supporting a collaborative environment, it is possible to provide a set of recommendation functionality that, to the best of our knowledge, have not yet been investigated. Indeed, the recommendations concern not only the data items and the collections made available by the DL, but also the users and communities. Due to space limitations, we will just sketch out the algorithms. The algorithms below are those implemented in the CYCLADES system.

Preliminaries. For ease of presentation, we will assume that data items are pieces of text (*e.g.* text documents). It is worth noting that our algorithms can be extended to manage data items of different media, like audio and video. By t_k, d_j, and F_i we will denote a text term, a data item, and a folder, respectively. Terms are usually identified either with the words, or with the stems of words, occurring in data items. For ease, following the well-known vector space model [22], a data item d_j is represented as a vector of *weights* $d_j = \langle w_{j1}, \ldots, w_{jm} \rangle$, where $0 \leq w_{jk} \leq 1$ corresponds to the "importance value" that term t_k has in the data item d_j, and m is the total number of unique terms in the indexed universal collection $L_{\top}$. The *folder profile* (denoted f_i) for folder F_i is computed as the *centroid* of the data items belonging to F_i. This means that the profile of F_i may be seen as a data item itself [2] (*i.e.* the mean, or prototypical, data item of F_i) and, thus, it is represented as a vector of weighted terms as well, *i.e.* $f_i = \langle w_{i1}, \ldots, w_{im} \rangle$. Of course, more complicated approaches for determining the folder profile may be considered as well, *e.g.* taking into account the hierarchical structure of the folders (see, *e.g.* [10]). Conceptually, they do not change much in our algorithm. Given a folder F_i, a data item $d_j \in F_i$ and a user $u_k \in \mathcal{U}$ such that $F_i \in u_k$, by $0 \leq r_{ijk} \leq 1$ we denote the *rating* given by user u_k to data item d_j relative to folder F_i. (A data item within a community folder, may be accessed, *e.g.* read, annotated and rated, by many different users.) We further assume that whenever a data item d_j belongs to a folder F_i of a user u_k, an *implicit* default rating $\breve{r}$ is assigned. Indeed, the fact that $d_j \in F_i \in \mathcal{F}^{u_k}$ is an implicit indicator of d_j being relevant to folder F_i for user u_k. Finally, we average the ratings r_{ijk} given by users u_k relative to the same data item–folder pair (i, j) and indicate it as r_{ij}.

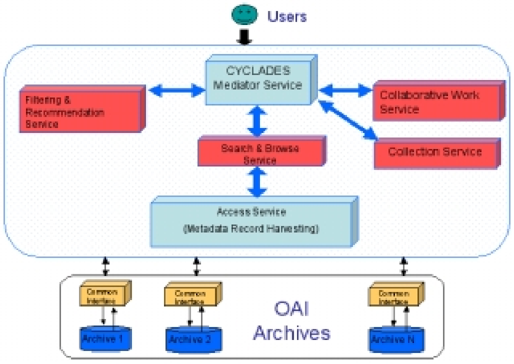


Fig. 4. Architecture.

In summary, we may represent (i) the data items as a 2-dimensional matrix, where a row represents a data item d_j and a column represents a term t_k. The value of the cell is the weight w_{jk} of term t_k in the data item d_j; (ii) the folder profiles as a 2-dimensional matrix, where a row represents a folder profile f_i and a column represents a term t_k. The value of the cell is the weight w_{ik} of term t_k in the folder profile f_i; and (iii) the ratings as a 2-dimensional matrix, where a row represents a folder F_i and a column represents a data item d_j. The value of the cell is the rating r_{ij}. The three matrixes are shown in Table 1, where $v = |\mathcal{F}|$ is the number of folders and $n = |L_\top|$ in the number of data items.

The *content similarity* (denoted $CSim(\cdot, \cdot)$) between two data items, or between a data item and a folder profile, or between two folder profiles, is a correlation coefficient (*e.g. cosine*) among two rows within the matrixes (a) and (b) of Table 1. Similarly, the *rating similarity* of two folders F_1 and F_2 (denoted $RSim(F_1, F_2)$) can be determined as a correlation [5,16] (*e.g. Pearson correlation coefficient*) between two rows of the matrix (c) in Table 1. Finally, the *similarity* (denoted $Sim(F_1, F_2)$ between two folders F_1 and F_2, which takes into account both the content and collaborative aspects, can be determined as a linear combination between their content similarity and their rating similarity.

Our recommendation algorithms follow a four-step schema described below. Let u be a user and let $F \in u$ be a folder (the *target folder*) for which the recommended items should be found. The algorithm schema is as follows: (i) select the set of most similar folders F_i to F, according to the similarity measure Sim; (ii) from this set, determine a pool of possible recommendable items; (iii) for each of the items in the pool compute a recommendation score; (iv) select and recommend a subset of items with the highest scores, and not yet recommended to F. We proceed now with a more detailed description of the above algorithm, specialised for the two cases of user recommendation[3] and of data items.

Recommendation of users. (i) Select the set $MS(F)$ of most similar folders to the target folder $F \in u$; (ii) for each folder $F_i \in MS(F)$, consider the users for which the folder F_i belongs to their folder hierarchy, *i.e.* compute the *pool of possible recommendable users* $P_U = \{u' \in \mathcal{U} \colon \exists F_i . F_i \in MS(F), F_i \in u'\} \setminus \{u\}$; (iii) compute the recommendation score for each possible recommendable user, *i.e.* for each user $u' \in P_U$ determine the *user hits factor* $h(u') = |\{F_i : F_i \in MS(F), F_i \in u'\}|$ (the number of folders F_i judged as similar to the target folder F belonging to user u'). For each user $u' \in P_U$ the *recommendation score* $s(F, u')$ is computed as follow: $s(F, u') = h(u') \cdot \sum_{F_i \in MS(F), F_i \in u'} Sim(F, F_i)$; and (iv) according to the recommendation score, select a set of most recommendable users, not yet recommended to the target folder F.

Note that the more a folder $F_i \in u'$ is similar to the target folder $F \in u$, the more related, in terms of interests, are the users u' and u. Additionally, the more similar folders there are belonging to user u', the more this u''s interests overlap

[3] The recommendation of communities and collections are quite similar.

Table 1. (a) The data item matrix. (b) The folder profile matrix. (c) The folder-data item rating matrix.

	t_1	$\ldots$	t_k	$\ldots$	t_m
d_1	w_{11}	$\ldots$	w_{1k}	$\ldots$	w_{1m}
d_2	w_{21}	$\ldots$	w_{2k}	$\ldots$	w_{2m}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
d_j	w_{j1}	$\ldots$	w_{jk}	$\ldots$	w_{jm}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
d_n	r_{n1}	$\ldots$	w_{nk}	$\ldots$	w_{nm}

(a)

	t_1	$\ldots$	t_k	$\ldots$	t_m
f_1	w_{11}	$\ldots$	w_{1k}	$\ldots$	w_{1m}
f_2	w_{21}	$\ldots$	w_{2k}	$\ldots$	w_{2m}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
f_i	w_{i1}	$\ldots$	w_{ik}	$\ldots$	w_{im}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
f_v	w_{v1}	$\ldots$	w_{vk}	$\ldots$	w_{vm}

(b)

	d_1	$\ldots$	d_j	$\ldots$	d_n
F_1	r_{11}	$\ldots$	r_{1j}	$\ldots$	r_{1n}
F_2	r_{21}	$\ldots$	r_{2j}	$\ldots$	r_{2n}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
F_i	r_{i1}	$\ldots$	r_{ij}	$\ldots$	r_{in}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
F_v	r_{v1}	$\ldots$	r_{vj}	$\ldots$	r_{vn}

(c)

with those of user u, which explains the computation of the recommendation score.

Recommendation of data items. The only difference with the above user recommendation concerns the computation of the recommendable data items and their recommendation score. Indeed, we will exploit the fact that data items are pieces of text and that there might be ratings associated: (i) the *pool of possible recommendable data items* is the set of data items belonging to the folders $F_i \in MS(F)$, *i.e.* $P_D = \{d \in \mathcal{D} \colon \exists F_i . F_i \in MS(F), d \in F_i\} \setminus \{d \in \mathcal{D} \colon \exists F' \in u, d \in F'\}$ (we do not recommend data items already known to the user); (ii) the recommendation score for $d_j \in P_D$ w.r.t. F is computed as a linear combination of the *content-based* and the *rating-based recommendation scores*. The content-based recommendation score of $d_j \in P_D$ w.r.t. the target folder F is the content similarity between d_j and the folder profile of F. The ratings-based recommendation score of d_j w.r.t. F is the weighted sum $s^R(F, d_j) = \overline{r} + \frac{\sum_{F_i \in MS(F)} (r_{ij} - \overline{r}_i) \cdot RSim(f, f_i)}{\sum_{F_i \in MS(F)} \cdot RSim(f, f_i)}$, where $\overline{r}$ $(\overline{r}_i)$ is the mean of the ratings in the target folder F.

5 Conclusions

We envisage a Digital Library not only as an information resource where users may submit queries to satisfy their information needs, but also as a collaborative working and meeting space. Indeed, users looking within an information resource for relevant data may have overlapping interests, which may turn out to be of mutual benefit - users may well profit from each other's knowledge by sharing opinions and experiences. To this end, we have formalised a personalized collaborative Digital Library environment in which the user functionality may be organised into four categories: users may (i) search for information; (ii) organise the information space (according to the "folder paradigm"); (iii) collaborate with other users with similar interests; and (iv) get recommendations. We also described the CYCLADES system, which is an ongoing implementation of such an environment. We are aware that many concepts and techniques presented in this paper require further investigation, which we will carry out in the future.

Acknowledgements. This work is funded by the European Community in the context of the CYCLADES project IST-2000-25456, under the Information Societies Technology programme.

References

1. G. Amato and U. Straccia. User profile and applications to digital libraries. In *Proc. 3rd European Conf. on Research and Advanced Technology for Digital Libraries (ECDL-99)*, LNCS 1696, pages 184–197, Paris, France, 1999. Springer-Verlag.
2. N. J. Belkin and B. W. Croft. Information filtering and information retrieval: Two sides of the same coin? *Comm. of the ACM*, 35(12):29–38, 1992.
3. D. Billsus and M. J. Pazzani. Learning collaborative information filters. In *Proc. 15th International Conf. on Machine Learning*, pages 46–54. Morgan Kaufmann, San Francisco, CA, 1998.
4. K. Bollacker, S. Lawrence, and C. L. Giles. A system for automatic personalized tracking of scientific literature on the web. In *DL 99 - The 4th ACM Conf. on Digital Libraries*, pages 105–113, New York, 1999. ACM Press.
5. J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. 14th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-98)*, pages 43–52, Madison, Wisconsin, USA, 1998.
6. J. Callan. Learning while filtering documents. In *Proc of the 21th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval (ACM SIGIR-98)*, pages 224–231, Melbourne, Australia, 1998.
7. C. Carpineto, R. De Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.
8. M. Di Giacomo, D. Mahoney, J. Bollen, A. Monroy-Hernandez, and C. M. Rouiz Meraz. Mylibrary, a personalization service for digital library environments, 2001.
9. DublinCore. Dublin core metadata element set. `purl.org/metadata/dublin_core`, WWW.
10. S. Dumais and H. Chen. Hierarchical classification of web content. In *Proc. 23rd Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-00)*, pages 256–263, Athens, Greece, 2000.
11. D. Faensen, L. Faulstich, H. Schweppe, A. Hinze, and A. Steidinger. Hermes: a notification service for digital libraries. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 373–380, 2001.
12. C. Faloutsos and D. W. Oard. A survey of information retrieval and filtering methods. University of Maryland Technical Report CS-TR-3514, 1995.
13. L. Fernandez, J. A. Sanchez, and A. Garcia. Mibiblio: personal spaces in a digital library universe. In *ACM DL*, pages 232–233, 2000.
14. P. W. Foltz and S. T. Dumais. Personalized information delivery: An analysis of information filtering methods. *Comm. of the ACM*, 35(12):51–60, 1992.
15. D. J. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave information tapestry. *Comm. of the ACM*, 35(12):61–70, 1992.

16. J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proc. 22th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-99)*, pages 230–237, Berkeley, CA USA, 1999.
17. *Information Filtering Resources.* `www.enee.umd.edu/medlab/filter/filter.html`.
18. J. Mostafa, S. Mukhopadhyay, W. Lam, and M. Palakal. A multilevel approach to intelligent information filtering: Model, system, and evaluation. *ACM Transactions on Information Systems*, 15(4):368–399, 1997.
19. A. Moukas. *Amalthaea*: Information discovery and filtering using a multiagent evolving ecosystem. In *Proc. Practical Applications of Agents and Multiagent Technology*, London, GB, 1996.
20. P. Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proc. ACM 1994 Conf. on Computer Supported Cooperative Work*, pages 175–186, Chapel Hill, North Carolina, 1994. ACM.
21. L. M. Rocha. Talkmine and the adaptive recommendation project. In *ACM DL*, pages 242–243, 1999.
22. G. Salton and J. M. McGill. *Introduction to Modern Information Retrieval.* Addison Wesley Publ. Co., Reading, Massachussetts, 1983.

Virtual Tutor: A System for Deploying Digital Libraries in Classrooms

Joemon M. Jose, Hywel Braddick, Innes Martin, Brian Robertson,
Craig Walker, and Greg MacPherson

Department of Computing Science
University of Glasgow, Glasgow G12 8QQ, Scotland, UK.
{jj,braddihd,martini,robertsb,walkercp,macphegf}@dcs.gla.ac.uk

Abstract. In this paper, we introduce a collaborative learning environment called Virtual Tutor for exploiting valuable digital libraries for teaching and learning. Virtual Tutor facilitates the building of online lectures and their provision. It also supports collaborative and explorative learning. We describe the architecture of the system and its functionality. A summative evaluation is conducted on the system. The paper concludes by comparing the Virtual Tutor functionality with similar approaches and highlighting the strengths of the Virtual Tutor design.

1 Introduction

The proliferation of powerful computing systems and network technologies have greatly increased the ease with which digital resources are created and shared in both commercial and educational organisations. This increase, introduces the possibility of open-ended teaching and learning activities. Realising this potential, a number of educational digital resources have been created under the banner of digital library initiatives (e.g., JISC (www.jisc.ac.uk), NSF (www.nsf.gov) digital libraries activities). In addition, the Internet hosts valuable digital information that can also be used for teaching and explorative learning. However, our developing prowess in casting data in electronic form is not matched by any compatible facility for effective usage.

Online resources together with emerging network technologies provide new opportunities that were not feasible in traditional libraries for learning activities. In addition to the obvious way of delivering teaching materials online, it opens up the possibility of collaborative and explorative learning. Teachers can use digital resources to find relevant materials on the topic of discussion. This is faster than traditional paper and book based mechanisms and will lead to more timely and up-to-date information. The selected digital documents can then be provided to students as supporting material. Students will be able to use these resources in the context of discussion, and in addition, will also be able to use the digital resources for enhancing their knowledge. There exists a potential for explorative learning and online collaboration. However, these possibilities are under utilised due to the lack of proper tools.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 275–286, 2002.

In this paper, we introduce a collaborative learning environment called Virtual Tutor. This paper is organised as follows. In Section 2, we discuss issues related to online learning and derives a set of criteria for our development. Section 3 introduces the Virtual Tutor design and the rationale behind this design. In Section 4, we discuss a summative evaluation. We argue the power of our design in Section 5 and conclude in Section 6.

2 Background

2.1 Online Digital Resources

Solvberg [7] defines a digital library as one consisting of digital collections, a working environment and associated technologies and services. A collection contains information objects put together according to some rules or ideas. In order to support users in the information seeking process, catalogues, indexes and retrieval services are provided. For example, JISC in UK supported the creation of a number of digital libraries for online learning (e.g., AXIS (http://www.axisartists.org.uk), EEVL (http://www.eevl.ac.uk)). In addition, there exists repository like the World Wide Web where information is deposited without adhering to any specific rules.

These digital resources are valuable knowledge sources for teaching and learning purposes. Many academics place the latest essays and research articles on the Web, making it an important place for up-to-date information. A given domain may be supported by a number of specialised digital libraries. In order to use these resources for learning & teaching, one has to develop techniques to find and use the information in teaching. In addition, these materials need to be delivered online for easy and flexible access. Finding relevant information from a number of sources using search engines with different technologies is a difficult problem. Once relevant sets are retrieved, identifying and using a set of useful documents is another problem. In this work, we address these issues by building a workspace in which different search engines can be represented. It also provides facilities to view and compare documents for selection purposes [4].

2.2 Online Teaching and Learning

Online delivery of course materials can be extremely flexible. This flexibility comes from two angles. Firstly, students can explore the material at their own pace and time, and is a viable substitute for face-to-face learning. Secondly, it provides an opportunity to teaching staff to constantly rectify and revise their materials based on student feedback. A re-examination of the objectives and the organisation of course materials can lead to improvements in the way that topic is taught. This makes the materials of superior quality and up-to-date.

However, in order to make online learning successful, it should engage the student in the learning material. Milligan [6] argues that merely using electronic delivery as a means of enhancing the presentation of learning material is not enough. This is because, Milligan continues, the face-to-face learner gets far more from a lecture than merely the information that is written on the blackboard or spoken by the lecturer. In addition to this formal learning, they get the chance to meet and discuss issues with their peers, they have the opportunity to interrupt their lecturer when they fail to understand, they get an indication as to how quickly they should be progressing through the materials etc. This type of interaction helps to reinforce learning, and catch misconceptions early. In order to make online delivery effective, all these 'extras' have to be delivered alongside any online learning material.

Our goal is to build a system that supports active exploration, and learning. This system should keep the flexibility of online delivery while providing the benefits of classroom teaching. Basically, an ideal online learning material should extend beyond being a virtual course book, to being a virtual classroom.

2.3 Objectives

In this work, we aim to develop a collaborative environment for explorative learning. Our idea is to support collaboration through peer-to-peer consultation and to support exploration through the use of multiple digital resources. People learn more effectively by talking to others and swapping information and ideas. The basis of our approach is the belief that good online learning environment must provide not just the knowledge or information, but also the opportunity for communication and reinforcement of learning through reflection and collaboration. In the following we describe the desiderata for our work.

Desiderata

- Support Lecture preparation: While preparing course notes, Lecturers always look for additional material (on multiple digital resources) to verify a fact and/or to provide additional relevant material to students.
- Support explorative learning and collaboration: It is important to facilitate explorative learning. That is, students should be able to probe related materials (on multiple sources) and share their ideas with peers thus supporting collaboration.
- Support Flexibility: A learning environment should ideally offer the same flexibility as a piece of paper. Students take notes when they listen to a lecture and hence the system should provide a comparable facility.

3 Our Approach

To meet the criteria described above, we designed and implemented a system called Virtual Tutor. A sketch of the interface is shown in figure 1. Virtual Tutor can be used to create lectures. Lecture materials can be presented to the students with associated and supporting information. Teachers will be able to create a series of slides with links to relevant resources. Students will then be able to view the prepared lectures, adding their own comments, questions and links if they so desire. Since students can see what other students have taken from the lecture, they can use this information to make the best use of their time whilst using the facility. This will also provide feedback for lecturers on the material provided and highlight any areas of difficulty.

3.1 Interface

The layout of the main interface is critical to the success of creating a useful learning environment. The interface consists of two main parts, an overview window and the workspace. The workspace is designed to be a flexible area used to aid problem solving and organisation. The overview window will contain several different views of the overall structure of the current lecture (see figure 2). Both horizontal and vertical scrollbars will be used in order to give the impression of infinite size although, in reality, this size will be finite. The size of each component will be inversely proportional to the other with the use of a dynamic horizontal split pane, thus enabling the user to customise the main display depending on the task in hand. The following description gives the rationale behind the design decisions regarding each component included within this interface.

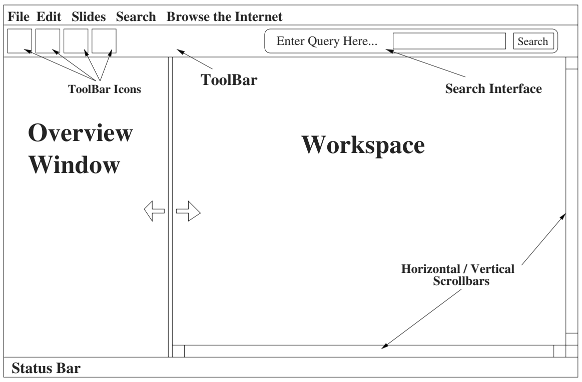


Fig. 1. Virtual Tutor Interface Design

3.2 Rationale behind This Design

The overview window will provide a hierarchical organisation of the slides and other associated information such as links to supporting materials, questions presented by the students and their answers (figure 2). It basically provides a structured view of the material. Comments, links, and questions are represented with distinguishable icons for easy understanding. Lecturers can use the 'miniSlides View' to create slides. The idea behind the workspace is to provide flexibility equivalent to a piece of paper.

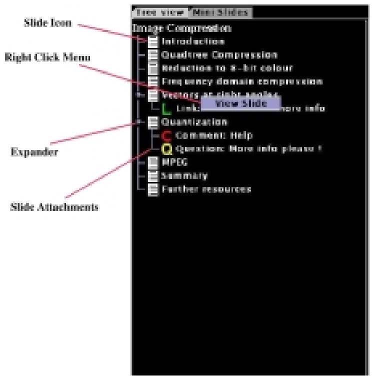


Fig. 2. Slide Viewer

People always need additional resources when teaching and learning. It is important to help students/staff to explore additional materials to clarify an issue or aspect. Lecturers always look for supporting material for emphasising a point. Partly due to the growth of the Internet, both in popularity and quantity of online data, it is increasingly common for academics to use the Internet as their first, and in many cases, only reference. In specialised environments, a number of digital resources may be available for a given domain. Examples are the digital collections from the JISC. Search functionality needs to support finding information from these resources. This will allow lecturers to look for information to construct a lecture, look for examples and counter examples. In addition, students may use these clues and explore this collection to look for additional materials (for example to write an essay). In the Virtual Tutor one can represent multiple search engines. Currently, we provide search facility for the Internet. A user can issue a search to a single search engine or more. The results will be displayed along with query-biased document summaries [8,9]. The workspace will allow selecting and viewing these retrieved set of documents. Users can drag and drop selected items to the workspace and link to the slides if needed. In addition, it allows them to group materials and store them for later use.

An important issue is to develop a medium for students to share their interpretations of a piece of information with each other and to explore the connections between this item and outside materials, including imagery, sounds, texts, and personal experience. This enables students to record their work, share it with others, and view and comment upon the work of others. The workspace acts as a medium for recording learning activities similar to taking notes on paper. Levy & Marshall [5] have shown that individuals “generally do not take notes by writing their observations down on a separate sheet of paper or in a text editor... Instead, they mark on the documents themselves”. For example, students often add annotations during lectures by marking the original printed lecture notes. These annotations (i.e., questions, links, comments, answers) would certainly be useful to the annotators themselves, but the sharing of annotations is an important way in which we use documents to collaborate. For example, students would have access to the annotations of fellow students, which would introduce all the benefits of collaborative learning as discussed earlier.

The workspace facilitates this activity by allowing students to note down questions and comments, and linking them to slides or slide groupings. In fact, this feature leaves traces of user actions behind, thus providing an opportunity to follow-it-up later. A system has been implemented in Java programming language and uses XML for data interchange between components. A snapshot of the interface is shown in figure 3.

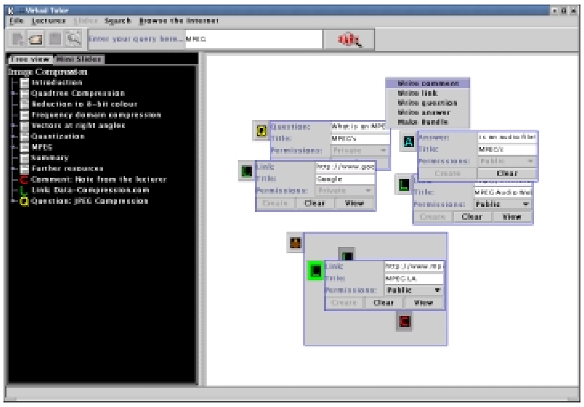


Image Compression
- Introduction
- Quadtree Compression
- Reduction to 8-bit colour
- Frequency domain compression
- Vectors at right angles
- Quantization
- MPEG
- Summary
- Further resources
- C Comment: Note from the lecturer
- L Link: Data-Compression.com
- Q Question: JPEG Compression

MPEG

Question: What is an MPE
Title: MPEG's
Permissions: Private

Link: http://www.goo
Title: Google
Permissions: Private

Answer: is an audio file!
Title: MPEG's
Permissions: Public

Title: MPEG Audio file
Permissions: Public

Link: http://www.mp
Title: MPEG LA
Permissions: Public

Fig. 3. Virtual Tutor Interface

3.3 Superimposition of Information

In the above section, we described the rationale behind our design. The Virtual Tutor can represent material such as a link to a Web resource, questions from students on a slide or a section, answers from lecturers, and comments. These, in a way, are a super imposition of information [1] on the basic lecture material.

In order to provide a flexible interface, we identified a hierarchy of superimposed information as described below:

- **link -** A link represents a shortcut to some resource. In general, resources may be anything such as video clips or databases but for the purpose of our prototype, the links will only provide shortcuts to web pages.
- **Comment -** If a user has something to add or say about a lecture, such as recommending a link or adding some extra information to expand on a slide then they can create a comment object and attach this to the lecture where appropriate.
- **Question -** If the user has a specific question about a slide, group of slides or the entire lecture, they can create a question object and attach it to the relevant part of the lecture.
- **Answer -** When a user sees a question which they are able to answer, they can create an answer and attach it to that question. Questions will not be limited to only having one answer so that multiple perspectives / viewpoints for each question can be seen.
- **Bundle** [2] **-** It will often be the case that users will want to add several of the above components to the lecture which are all related. Instead of adding these objects individually the user will be able to create a single bundle object with any number of objects inside. The bundle can then be added to the lecture where appropriate.

The class hierarchy shown in figure 4 describe various information units and their relationships. Similar objects will be made to inherit from the same superclass and hence share common properties, so can be treated in a generic manner where suitable. The main class will be the Lecture class which will store references to all components in the lecture as well as reference tables which will store the relationships between the components. The class diagram in figure 4 highlights the inheritance and dependency relationships between the classes.

3.4 System Features

When users create superimpositions they will be able to decide whether others will be able to view that component. This allows users to have private information in their view of the lecture whilst still enabling collaboration as the user can make components viewable to everyone by declaring them as public. The only user who will have permission to see all superimpositions, whether public or private, is the author of the lecture. This will allow the lecturer to see what problems people are having and to answer any questions which students may have made private.

A user can have his own work-space on which he can create and manipulate objects either to be included in the lecture or for general problem solving. This provides the user with an area of temporary storage where he can leave objects which may only be used for that session and not attached to the lecture. The

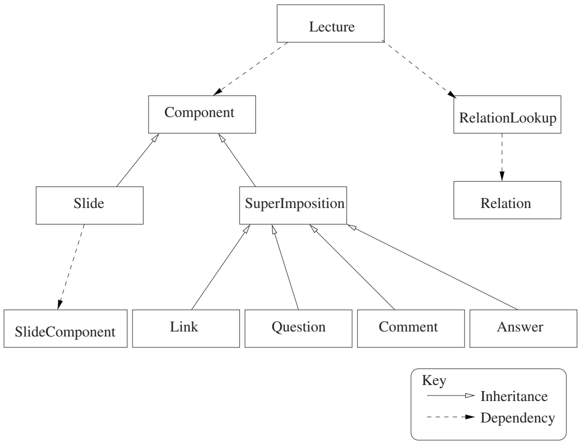


Fig. 4. Data Model class diagram

user will be able to add several objects to the workspace and then select these objects to be collectively added to a bundle. The workspace will also allow users to arrange their own personal information in a way that best suits them, thus giving a high degree of customisation and flexibility.

There are two main views for the lecture. One of these views is to be used to display an overview of all the slides, also indicating if there are any other components attached to each one. The other view is to be a structured hierarchical view showing all the components of the lecture and the relationships between them (see figure 2). If a user wishes to view an individual slide, he can select it from either of the viewers to be displayed in an individual slide viewer. The user will be able to have several slides open in his own viewers at the same time, thus allowing him to view and compare several different slides simultaneously.

When the user wishes to carry out a search, he will be able to select the domains he wishes to submit the query to. There will be a separate set of results returned for each of the domains selected (figure 5). Once the results have been returned, the user will then be able to view the results. If the user wishes, he will then be able to drag the link from the results to the workspace or directly on to the lecture. In order to help the user to judge the relevancy of a retrieved document, we have incorporated query-biased summaries as well [8,9].

4 Evaluation

We conducted a summative evaluation to investigate the strengths and weaknesses of the Virtual Tutor design. We used the following instruments for con-

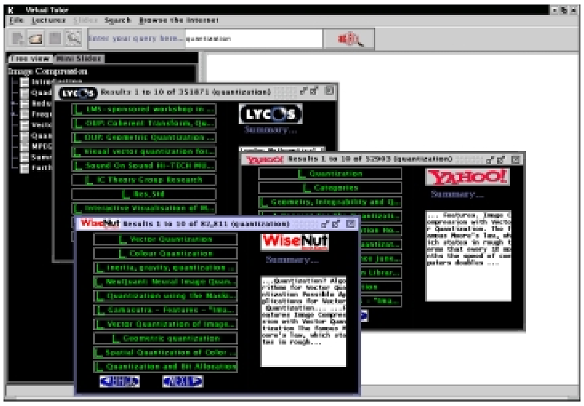

Virtual Tutor
File Lectures Slides Search Browse the Internet
Enter your query here... quantization
Tree view Mini Slides
Image Compression
Results 1 to 10 of 351871 (quantization)
LYCOS
Summary...
Results 1 to 10 of 52903 (quantization)
YAHOO!
Summary...
Quantization
Categories
Results 1 to 10 of 82,811 (quantization)
WiseNut
Summary...
Vector Quantization
Colour Quantization
Geometric quantization

Fig. 5. Search facility in virtual Tutor

ducting the evaluation: Think aloud observation; Interview; and Questionnaire. Subjects were given a quick reference guide and basic introductory session. They were expected to be able to use the basic functions of the program after having read this guide. The subjects were then given a number of tasks to complete, and were expected to try and solve these tasks without any further help from the evaluator. It was hoped not only that the techniques would allow specific problems in the design to be identified, but also to assess the system's functionality as well as usability. This includes the efficiency, effectiveness and satisfaction.

We used twelve undergraduate students. Four of them were chosen to go through the think aloud observation. The other eight users were either interviewed or filled out a questionnaire. From the questionnaire, it was gathered that most subjects were complimentary about the layout and system functionality. Think-aloud was definitely the most informative of the three evaluation techniques used, as is often the case when evaluating a graphical user interface. It gave valuable insight into how users interact with the system, and reveal their strategies for overcoming any problems encountered. Subtle user behaviour such as expecting a link to pop up from the tree view after a double-click was also discovered, as was the need for clearer feedback from the system after user actions. From the interviews conducted, it appeared that although most of the users were happy with the system, some minor interactions such as viewing super impositions (comments, links etc) could be improved. Users felt that the layout was adequate, the bundling technique was fine and the lecture material was presented clearly. From the three evaluation techniques carried out and through our own internal evaluation, it is clear that a solution has been developed that can be seen as a success. The attempted solution to the organisation of large amounts of heterogeneous data through our bundling technique has proved to be quite successful.

5 Discussion

Despite evident usefulness, collaboration with a view to learning is one area in which the paper medium still seems to be preferred to the digital. How does Virtual Tutor make progress on this issue? What are the strengths and weaknesses of the Virtual Tutor, and how it does contribute to online learning? In this section, we try to answer these questions and compare Virtual Tutor against other online learning environments.

Most of the research to-date addresses the issue of providing material online and supporting this process by developing administrative tools [6]. These systems are generally known as MLEs (Managed Learning Environments), VLEs (Virtual Learning Environments) etc. The main issues addressed are administrative and those of basic technology. These approaches overlook the ability of networks to bring people together in new configurations to support open-ended learning activities. We believe the Virtual Tutor design is a quantum leap forward compared to these systems.

Virtual Tutor facilitates collaboration by allowing students to ask questions and record them to the particular point of contention. The workspace facility allows them to record each action in such a away that others can get benefit out of them. We believe, as Hendry [4] argued, that leaving traces of their previous actions will facilitate collaboration. Possibly the most important concept that introduced in the development of this prototype has been the organisation of heterogeneous data into easily understandable, higher level groupings, where the user can group slides along with comments, links and questions to be added to lectures or left on the shared workspace for others to view. In addition to these, the workspace allows the representation of many different search engines and is extensible. Indeed, we have represented search facility from many different web search engines (figure 5). The results are presented with a query-biased summary of each retrieved document, thus facilitating easy relevance judgement [8,9]. This is one of the important features that makes the Virtual Tutor design different in comparison to other tools such as CSILE (http://csile.oise.utoronto.ca/intro.html).

Wilensky [10] defines spontaneous collaboration as "the ability of users to collaborate without first undertaking a substantial mutual administration commitment". Wilensky continues that this concept is likely to have a profound effect on the process by which lecturers/students teach and learn with the possibility of superseding traditional paper-based lecture notes or even the entire concept of lectures as we know them today. The workspace in the Virtual Tutor is designed with a view to enhance collaboration. On the workspace, one can take notes, make annotations and mark questions as one does with paper-based lecture notes. In addition, one can search for additional materials relevant to a topic and incorporate them into the workspace. This in a way leaves traces of one's activity and is helpful to that individual as well as others in understanding a topic of discussion.

In such a collaborative environment, people can work together across time and space outside normal class time. We believe that Virtual Tutor enables learners to accomplish tasks and develop understanding beyond what they can achieve individually, by making use of distributed expertise and multiple perspectives. In addition, as argued in [3], the communication that is required to support collaboration forces learners to articulate their understanding in ways that help them to organise their knowledge and acknowledge gaps in their understanding. We believe the Virtual Tutor makes an important contribution towards this end.

Recent educational research focuses on the value of collaboration and of open-ended activity for learning. The Collaboratory Notebook Project at Northwestern University investigates issues related to collaboration and open-ended learning [3]. The Notebook was designed specifically to support communication and collaboration among students, instructors, and mentors, and to allow people to work together across time and space outside the normal class time. However, the notebook is rigidly structured and quite over-determined which limits the flexibility of the application. The similarity between the Notebook and the Virtual Tutor is that of slide overview. The Notebook organises the system similar to that of slide overview (figure 2). Compared to this, the Virtual Tutor interface provides much more flexibility by the introduction of a workspace. In addition, Virtual Tutor provides extensible search facility because one of our aims is to make use of digital resources for explorative learning, whereas Notebook concentrates on the online delivery of teaching materials.

6 Conclusion & Future Work

In this paper, we have introduced a collaborative learning environment called Virtual Tutor for exploiting digital libraries for explorative learning. Virtual Tutor can be used to create lecture materials. For this, a lecturer can search multiple collections and select relevant information which can be presented to students within the context of individual slides. Lecturers will be allowed to build up a series of slides with links to relevant materials. Students will then be able to view the prepared lectures, adding their own comments, questions and links if they so desire. Since students can see what others have taken from the lecture, they can use this information to make the best use of their time whilst using the facility. This will also provide feedback for lecturers on the material provided, and highlight any areas of difficulty which can then be addressed.

Our aim is to use digital libraries for explorative learning. We have been successful in building a prototype and the summative evaluation showed its usefulness. The system allows us to represent multiple search engines and provides techniques to select and use the retrieved set of results. In future work, we would like to investigate how to use ever increasingly available video and image materials in the system. In addition, we are also planning an evaluation in an operational setting.

Acknowledgements. Our thanks to Ryen White for providing the WebDocSum search system and assisting with its installation.

References

1. Bowers, S. A generic approach for representing model-based superimposed information. Research proficiency exam report, Oregon Graduate Institute, May 1, 2000.
2. Delcambre, L. M. L., Maier, D., Bowers, S., Weaver, M., Deng, L., Gorman, P., Ash, J., Lavelle, M., and Lyman, J. Bundles in captivity: An application of superimposed information. In *Proceedings of the 17th International Conference on Data Engineering (ICDE 2001)* (April 2-6 2001), IEEE Computer Society Press, pp. 111–120.
3. Edelson, Daniel, C., Pea, Roy, D., and Gomez, Louis, M. The collaboratory notebook. *Communications of the ACM 39*, 4 (1996), 32–34.
4. Hendry, D. G. *Extensible Information-Seeking Environments.* PhD thesis, The Robert Gordon University, Aberdeen, September 1996.
5. Levy, D. M., and Marshall, C. C. Going digital: A look at assumptions underlying digital libraries. *Communications of the ACM 38*, 4 (April 1995), 77–84.
6. Milligan, C. Virtual learning environments in the online delivery of staff development. JTAP report, Heriot-Watt University, URL:
http://jisc.ac.uk/jtap/htm/jtap-044.html, 1999.
7. Solvberg, I. T. Digital libraries and information retrieval. In *LNCS 1980 (Proceedings of ESSIR 2000)* (2000), M. Agosti, F. Crestani, and G. Pasi, Eds., Springer-Verlag, Berlin, Heidelberg, pp. 139–156.
8. White, R., Jose, J. M., and Ruthven, I. Query-biased web page summarisation: A task-oriented evaluation. In *Proceedings of the Twenty Fourth Annual International SIGIR Conference on Research and Development in Information Retrieval* (September 2001), ACM Press, pp. 412–413.
9. White, R., Ruthven, I., and Jose, J. M. Web document summarisation: a task-oriented evaluation. In *Proceedings of the First International Workshop on Digital Libraries DLib2001* (September 2001), IEE Press, pp. 951–955.
10. Wilesky, R. Digital library resources as a basis for collaborative work. *Journal of the Americal society for Information science 51*, 3 (2000), 223–245.

Resource Annotation Framework in a Georeferenced and Geospatial Digital Library[*]

Zehua Liu[1], Ee-Peng Lim[1], and Dion Hoe-Lian Goh[2]

[1] Centre for Advanced Information Systems, School of Computer Engineering
Nanyang Technological University, Singapore, 639798, SINGAPORE
{aszhliu, aseplim}@ntu.edu.sg
[2] School of Communication and Information
Nanyang Technological University, Singapore, 639798, SINGAPORE
ashlgoh@ntu.edu.sg

Abstract. G-Portal is a georeferenced and geospatial digital library that aims to identify, classify and organize geospatial and georeferenced resources on the web and to provide digital library services for these resources. Annotation service is supported in G-Portal to enable users to contribute content to the digital library. In this paper, we present a resource annotation framework for georeferenced and geospatial digital libraries and discuss its application in G-Portal. The framework is flexible for managing annotations of heterogeneous web resources. It allows users to contribute not only the annotation content but also the schema of the annotations. Meanwhile, other digital library services, such as visualization and classification, can be provided over the annotations since they are treated as first class objects. This paper mainly focuses on the design of the resource annotation framework.

1 Introduction

There is a large amount of geospatial resources available on the World Wide Web. While these are valuable for educational and research purposes, there have not been many georeferenced and geospatial digital libraries (DLs) developed for such public domain resources. In the DLESE project, metadata of earth related web sites/pages are contributed by user communities to establish a digital library consisting of quality education and resource materials [9].

To create a better learning experience and to allow sharing of users' knowledge, it is often desirable to allow users to annotate the resources, store these annotations in the DLs and make them available to other users. Nevertheless, in most existing georeferenced and geospatial DLs [8,2,9], users cannot easily place comments on geospatial and georeferenced resources or share their knowledge about these resources due to the lack of direct support for annotation services in these systems. User contribution of content, including annotations, is usually not permitted or only permitted under certain pre-defined format(s) in most DLs. This greatly hampers knowledge sharing.

[*] This work is partially supported by the SingAREN 21 research grant M48020004.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 287–298, 2002.

Supporting annotations on web-based geospatial and georeferenced resources poses several challenges. As web resources are distributed and heterogeneous, it is unrealistic to impose a single format or even a set of pre-defined formats for all annotations. Users should be allowed to create annotations on different types of resources as well as on types of resources that were not known when the annotation framework was designed.

Besides supporting annotations on resources, digital library services must be provided to enable meaningful use of contributed annotations. For digital libraries of geospatial resources, annotation support should also be able to take geospatial attributes into consideration when offering services such as query and visualization.

In this paper, we propose a resource annotation framework for enabling contribution and sharing of user knowledge on geospatial and georeferenced resources. This framework is developed based on G-Portal [5], a georeferenced and geospatial digital library. In G-Portal, resource annotation is provided as one of the digital library services. The contributed annotations are treated as first class objects, i.e. they are considered as resources in the DL as well. Other digital library services can be provided over these annotations, by treating them as normal resources.

In G-Portal, each annotation is "typed" by associating an annotation schema with it. Annotation schemas are extensions of the basic resource schema. Apart from allowing users to add annotations to the DL, the framework also allows users to define customized annotation schemas and create annotations based on these new schemas. This gives the annotation framework the flexibility required to work in heterogeneous web resources.

1.1 The G-Portal Project

G-Portal [5] is an ongoing digital library project at the Centre for Advanced Information Systems in Nanyang Technological University. The aims of the project include the identification, classification and organization of geospatial and georeferenced content on the Web, and the provision of digital library services such as classification and visualization. In addition, authorized users may also contribute metadata resources to G-Portal, making it a common environment for knowledge sharing. In G-Portal, metadata resources are structured descriptions about Web content and annotations.

G-Portal also provides a platform for building applications that use geospatial and georeferenced content. This is achieved through **projects** that represent collections of metadata resources gathered for specific purposes or applications. Resources within projects are further organized into **layers** which allow finer grained logical organization of metadata resources.

Metadata resources within a project are visualized using either a map-based interface or a classification interface. The map-based interface displays resources with spatial attributes. For example, metadata resources of countries, rivers and mountains and their associated content (such as identifying a particular climatic region on a map) can be shown in the map-based interface for easy viewing.

Navigation tools such as zoom and pan are provided for users to browse the map. In addition, layers in a project may be shown or hidden, changing the visibility of the associated resources.

Resources with or without spatial attributes can also be displayed within the classification interface that categorizes and presents resources using classification schemas. Examples of such resources include general information about climate ("Why are land and sea breezes a feature of many coastal regions") and population ("What problems does the growth of squatter settlements create for large urban areas?")

The map and classification interfaces are synchronized so that when a resource on one interface is accessed, related resources on the other interface are highlighted as well. For example, if a user selects a resource in a region on the map interface, the classification interface will highlight related resources, which may appear under different categories. In both the map and classification interfaces, the full content of any selected metadata resources can also be viewed.

1.2 Paper Outline

The remaining sections of this paper are organized as follows. In Section 2, we describe the design of resource annotation. In Section 3, we present the annotation framework in the context of G-Portal. In Section 4, we compare our work with some of the related work. Finally, some concluding remarks are given in Section 5.

2 Design of Resource Annotation

In this section, we propose a new resource annotation framework consisting of a schema approach to represent annotations and their relationships with the annotated resources. A detailed description of the annotation and its schema, with examples, is presented.

2.1 Representation of Annotations

In G-Portal, we distinguish the resources maintained by the DL system from those pre-existing web resources located at public domain web sites. The former are referred to as the *metadata resources*, while the latter are known as the *raw resources*. Metadata resources are the more structured versions of their raw counterparts and are contributed by users knowledgeable about the associated web sites or pages. To allow a more flexible use of metadata resources, we also allow metadata resources to be created without associating them with web resources. For the rest of our discussion, we shall use the terms "resource" and "metadata resource" interchangeably unless otherwise stated.

EXtensible Markup Language (XML) [3] was chosen to represent resources in G-Portal. This facilitates sharing and publication of resources for other systems to use. The XML representation also supports easy transformation of resources

```
<!-- BasicResource.xsd -->
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
    <xsd:element name="Resource" type="ResourceType"/>
    <xsd:complexType name="ResourceType">
        <xsd:sequence>
            <xsd:element name="ID" type="xsd:string"/>
            <xsd:element name="ResourceName" type="ResourceNameType"/>
            <xsd:element name="Location" type="LocationType"/>
            <xsd:element name="Creator" type="CreatorType"/>
            <xsd:element name="Source" type="SourceType"/>
            <xsd:element name="Content" type="ContentType"/>
        </xsd:sequence>
    </xsd:complexType>
    ... ...
    <xsd:complexType name="ContentType">
    </xsd:complexType>
</xsd:schema>
```

Fig. 1. The Base Schema for Resources

from one format to another. Each resource must be created using some *resource schema* defined using XML Schema [7]. All resource schemas are derived from a *base resource schema* as shown in Figure 1. The base resource schema includes the common attributes of a resource such as identifier, location, and access control. Each derived resource schema can include other attributes relevant to the kind of resources covered by the schema. Each resource is assigned a unique identifier within the G-Portal system. The location attribute registers the geospatial properties (i.e., shape and location) of the resource. The access control attribute keeps the ownership and security information about the resource.

In order to treat annotations as first-class objects, we can either treat them as a new class of objects or as a special type of resource. In our proposed framework, the latter is adopted. In other words, we also introduce a schema for describing a set of annotations. This will be elaborated in Section 2.2. Other than the attributes of basic resource, a new attribute (known as the `AnnotatedResources` element) to identify the annotated resources is added to the annotation resource. Making annotation a subtype of resource allows us to treat annotations as resources. This also makes it possible for annotations to be annotated, just like ordinary resources. In this way, the DL services designed for normal resources can be applicable to annotation resources. From the implementation point of view, the DL system will not be unnecessarily complex because of the introduction of such a resource annotation framework, since there is still only one type of first-class objects to deal with.

Another side effect of an annotation being a resource is that annotations may have geospatial attributes as other resources do. The geospatial attributes will usually be derived from the resources being annotated, if they have geospatial attributes. An annotation may also have its geospatial attribute value directly assigned by the annotator.

Unlike most existing annotation frameworks where annotations can be created only for individual resources, our framework allows *annotating multiple re-*

```
<!-- BasicAnnotation.xsd -->
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
    <xsd:redefine schemaLocation="Resource.xsd"/>
    <xsd:element name="Resource" type="AnnotationType"/>
    <xsd:complexType name="AnnotationType">
        <xsd:complexContent>
            <xsd:extension base="ResourceType">
                <xsd:sequence>
                    <xsd:element name="AnnotatedResources" type="AnnotatedResourcesType"/>
                </xsd:sequence>
            </xsd:extension>
        </xsd:complexContent>
    </xsd:complexType>
    <xsd:complexType name="AnnotatedResourcesType">
        <xsd:sequence>
            <xsd:element name="Resource" type="xsd:string" maxOccurs="unbounded"/>
        </xsd:sequence>
    </xsd:complexType>
</xsd:schema>
```

Fig. 2. The Base Schema for Annotations

```
<?xml version="1.0" encoding="UTF-8"?>
<Resource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:noNamespaceSchemaLocation="Resource.xsd">
    <ID>Country_65</ID>
    <ResourceName><Name>Singapore</Name></ResourceName>
    <Location Type="Geometry"> <Geometry> ... ... </Geometry> </Location>
    <Creator> ... ... </Creator>
    <Source><Link Type="External">http://www.sg/</Link></Source>
    <Content> ... ... </Content>
</Resource>
```

Fig. 3. A Simple Resource about Singapore

sources with a single annotation. For example, in a map that displays South-East Asian countries, a comment about the historical relationship between Malaysia and Singapore can be made by annotating the resources representing the two countries. In traditional annotation systems, this can be achieved only by annotating one country and mentioning the name of the other country within the annotation. For instance, an annotation is created for the resource corresponding to Malaysia and this annotation carries a reference to the Singapore resource as its attribute. While this approach may achieve the same effect as annotating both resources, it is not intuitive because it requires one of the annotated resources to be used as the anchor. To manipulate this annotation, one has to take the extra step to find out all other annotated resources from this anchor. Hence, our proposed framework has taken the option to allow annotation based on multiple resources and provides a standard way of representing the annotated resources as a list under the `AnnotatedResources` element.

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xsd:redefine schemaLocation="BasicResource.xsd">
    <xsd:complexType name="ContentType">
        <xsd:sequence>  ... ...  </xsd:sequence>
    </xsd:complexType>
  </xsd:redefine>
</xsd:schema>
```

Fig. 4. A Resource Schema for Country

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified">
  <xsd:redefine schemaLocation="BasicAnnotation.xsd">
    <xsd:complexType name="ContentType">
        <xsd:sequence>
            <xsd:element name="Comment" type="xsd:string"/>
        </xsd:sequence>
    </xsd:complexType>
  </xsd:redefine>
</xsd:schema>
```

Fig. 5. An Country Annotation Schema

2.2 Annotation Schema

As shown in Figure 2, the basic annotation schema extends the basic resource schema by appending a new element (`AnnotatedResources`) to keep the resource ids of the annotated resources. Other important attributes of a resource, such as id and geospatial attribute, are inherited from the basic resource schema.

The definition of the content of the basic annotation schema is left to users, just as in the definition of resources. New types of annotations can be created by altering the format of the content of the annotation.

```
<?xml version="1.0" encoding="UTF-8"?>
<Resource xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
          xsi:noNamespaceSchemaLocation="Annotation.xsd">
    <ID>Annotation_01</ID>
    <ResourceName>
        <Name>Annotation about Singapore and Malaysia</Name>
    </ResourceName>
    <Location Type="Geometry">  <Geometry> ... ... </Geometry>  </Location>
    <Creator> ... ... </Creator>
    <Source> ... ... </Source>
    <Content>
        <Comment>Singapore is located beside Malaysia</Comment>
    </Content>
    <AnnotatedResources>
        <Resource>Country_65</Resource>
        <Resource>Country_60</Resource>
    </AnnotatedResources>
</Resource>
```

Fig. 6. A Simple Annotation about Singapore and Malaysia

2.3 Annotation Example

This section gives an example of an annotation represented in XML. An example of a resource about Singapore is shown in Figure 3, using a country resource schema given in Figure 4. Note that basic attributes of a resource include id, name, location, creator, source, and the actual content. Details of some attributes are omitted because they are not relevant to the annotation aspect. Resources about other countries, such as China and Malaysia, are represented similarly.

Before one can annotate countries, a country annotation schema is first derived from the basic annotation schema shown in Figure 2. The derived annotation schema is shown in Figure 5. The new annotation schema can be derived by altering the definition of the "Content" node in the basic annotation schema.

An annotation about both Singapore and Malaysia is shown Figure 6. Detailed definition of the individual elements is omitted due to space constraints. Comparing the format to that of the resource, one additional element "`AnnotatedResources`" is inserted, which is used to indicate the ids of resources to be annotated. In this example, we are annotating both the Singapore (resource id "Country_65") and Malaysia (resource id "Country_60") country resources. The "Content" element specifies the details of the annotation. (Here, only a simple comment is given.) The structure of the "Content" element depends on the annotation schema and is not fixed.

3 Annotations in G-Portal

This section describes how the resource annotation framework is applied in the G-Portal system. We will focus on the creating of annotations, browsing of annotations and classification of annotations.

3.1 Creating an Annotation

Registration of Annotation Schemas. As annotations must be instantiated by some annotation schemas, the annotator must register his or her annotation schemas with G-Portal. The annotation schemas can be defined beforehand and made available as XML Schema files on the web. The registration of an annotation schema simply involves providing the URIs of these schema files to G-Portal.

Creating Annotation Layers in a Project. Recall from Section 1.1 that a *project* contains a collection of resources and that all these resources are organized into *layers*. Annotations, as a type of resource, should also be grouped into some layer in the project in order for G-Portal to visualize them, since G-Portal only visualizes resources in the current project.

Each project has an owner. Before he can annotate resources within a project, the project owner must configure the project layers and their resources to be readable by the public. The owner should also grant users permission to create new layers within the project. Such permission, applicable to new annotation layers only, will not affect the original layers constructed by the owner.

When the necessary access rights have been granted, an annotator can first create a new layer under the current project. The new layer will have the annotator as the owner. Nevertheless, since the annotation layer belongs to the project, its existence will also be dependent on the project.

Bookmarking in G-Portal. The next step to creating an annotation is to indicate the resources to be annotated. This is accomplished by the *bookmarking* facility provided by G-Portal. The bookmarking facility was initially developed to allow easy exploration of resources using both the *map-based interface* and the *classification interface* in G-Portal, where users can bookmark target resources in one interface and continue exploring among these resources in the other interface.

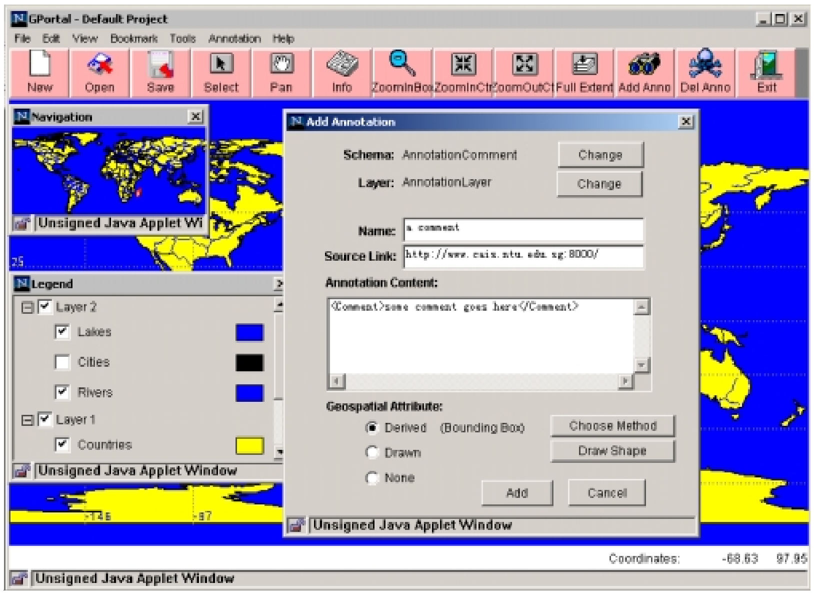
Bookmark resources are created incrementally by repeatedly selecting resources and invoking the command "Add to Bookmark". Users can also ask G-Portal to select all bookmarked resources or remove some resources from the bookmark. Bookmarked resources are highlighted differently from those not bookmarked and the normal selected resources, and are not affected by the normal selection operation.

Bookmarks in the two interfaces are synchronized. Resources bookmarked in one interface will also be highlighted in the other, so that users can easily switch between the two interfaces. For example, the user can draw a rectangle to select resources in the map-based interface and bookmark them, and then go to the classification interface to explore the bookmarked resources one by one.

To facilitate annotation, bookmarking is used to select the resources to be annotated. This is especially useful when users want to annotate multiple resources, located in different places on the map.

Construction of an Annotation. Once the resources to be annotated are bookmarked, a new annotation record can be constructed within the annotation layer by selecting the appropriate annotation schema to be used. This annotation schema can be selected from a list of annotation schemas registered by the annotator.

After the annotation schema is selected, G-Portal provides an interface, as shown in Figure 7, for the user to enter the content of the new annotation. Since the content of annotations can be defined freely by users, there is no fixed interface for users to enter the content. Moreover, the XML hierarchical format of annotation content makes it more difficult to provide a meaningful interface for content creation. In the current implementation of G-Portal, we have decided to provide a simple text editor for users to enter XML content. The user is allowed to enter any XML text conforming to the annotation schema. He also provides values for other attributes such as Name and Link. The rest of the attributes, including ids of annotated resources and geospatial attributes, will be automatically generated and combined with the XML content entered to form the



Schema: AnnotationComment

Layer: AnnotationLayer

Name: a comment

Source Link: http://www.cais.ntu.edu.sg:8000/

Annotation Content:

```
<Comment>some comment goes here</Comment>
```

Geospatial Attribute:

● Derived (Bounding Box)
○ Drawn
○ None

☑ Layer 2		
☑ Lakes	🟦	
□ Cities	■	
☑ Rivers	🟦	
☑ Layer 1		
☑ Countries	🟨	

Coordinates: -68.63 97.95

Fig. 7. The Add Annotation Dialog

complete annotation. Note that the ids of the annotated resources are automatically identified and inserted into the annotation as the "AnnotatedResources" attribute.

Geospatial Attribute of Annotations. In G-Portal, annotations, just like resources, may contain geospatial attributes that describe their geographical representation. The annotator may directly provide G-Portal with the geospatial information, if the annotation itself has such information. He provides the shape information by choosing the "Drawn" option in Figure 7 and drawing the shape on the map. For example, when commenting on the origin of Singapore's Chinese population, the user may want to use the southern part of China as the geospatial property.

Alternatively, the user may ask G-Portal to automatically derive this geospatial information. This is done by selecting the "Derived" option in Figure 7 and clicking on the "Choose Method" button to select an appropriate derivation method. The selected method will be shown within brackets. One such method is to use the geospatial attributes of all the annotated resources. Another possible operation is to use the smallest rectangle that covers all resources, i.e. the bounding box. This is especially useful when annotating a group of resources located near each other, such as the ASEAN (Association of South-East Asian Nations) countries.

3.2 Browsing Annotations

As mentioned earlier, G-Portal provides two interfaces - the map-based interface and the classification interface, for presenting resources. Users can browse through annotations using these two interfaces, just as with the other resources.

The map-based interface is more intuitive but limited to annotations with geospatial attributes. Users can use the navigation tools provided by G-Portal to navigate the map and to zoom in and out till the target information is located. The classification interface shows all annotations, including those without geospatial attributes. Annotations are classified into categories according to some classification schema [6]. The categorized annotations are presented in the classification interface in the tree structure determined by the category hierarchy. Users browse through the classification interface in a way similar to looking for a file in Windows Explorer.

To further assist the navigation of resource annotations, when an annotation is selected by the user, the resources annotated by that annotation will be automatically highlighted as well. In this way, the relationship between annotations and annotated resources becomes more obvious and can be identified by users more easily.

Resources that have been annotated will be presented in a different highlighting style. Users can ask to see all the annotations for a particular resource. The result of the request is a list of annotations, which the user can go through one by one.

Users can view the content of an annotation by double-clicking on the annotation or selecting the Info Tool and single-clicking on the annotation. The current implementation of G-Portal displays content of resources, including annotations, in plain XML format, with indentation formatting.

3.3 Classification of Annotations

Classification of resources, including annotations, is based on rules that rely on values of certain attributes of the resources. The actual classification is performed by the classification engine on the server. The classification produces a hierarchy of categories with annotations assigned to the categories. The result is presented in the classification interface when requested.

Apart from defining the classification rule based on standard attributes of resources, special attributes, especially the ids of annotated resources, can be used to provide useful classification of annotations. For example, based on the ids of the annotated resources, annotations may be grouped by countries and by regions. This kind of classification would not be possible for normal resources without introducing extra attributes into the resource to indicate the names or ids of the annotated countries.

4 Related Work

Annotation has always been seen as an important digital library function that solicits user contributed knowledge to be shared among users [11].

In the context of web annotation, the Annotea web-based annotation system supports RDF-based annotations of web pages or other web objects with URIs Universal Resource Identifiers) [4]. Each annotation is represented as a set of metadata and an annotation body. Multiple annotation classes (similar to our annotation schemas) can be created by Annotea to instantiate different annotations for different purposes (Advice, Comment, Example, etc..). G-Portal is also similar to Annotea in providing some database servers to store the created annotations. On the other hand, G-Portal and Annotea have a few notable differences. First, G-Portal's annotations are created over the metadata resources of the web objects instead of the web objects directly. Second, G-Portal provides an additional annotation element to accommodate the geospatial property of annotations. Third, the content of annotations in Annotea is not structured (free text or HTML document) and the annotations can be associated with a single document and some portion of it. Finally, G-Portal provides both map and classification views of the annotations.

In the DLESE and ADEPT digital library projects, a common metadata framework known as the ADN Framework has been proposed to standardize the representation of metadata resources including annotations [9,8,1]. Annotations are mainly defined for educational purposes. The metadata format of annotations consists of mainly the contributor, creation date and description components. The description component of an annotation is essentially a free text comment on the annotated web object. G-Portal, on the other hand, has adopted a more flexible annotation schema structure which allows the basic annotation schema to be extended to include different elements that cover a wide variety of annotations. The ADN Framework is currently under some major changes to adopt the XML Schema for describing metadata.

Wilensky, in the UC Berkeley Digital Library project, proposed a new *multivalent document* model that supports documents with behaviors that allow annotations to be added to documents and be manipulated with a variety of operations, e.g. copyediting [10]. This document model has been implemented in a GIS Viewer browser that supports the addition of annotations to a document containing geospatial and multimedia information. Nevertheless, the concept of annotation schemas and collection-level visualization and classification of annotations were not included.

5 Conclusion

In this paper, we proposed a resource annotation framework in the context of a digital library for geospatial and georeferenced web resources. The framework is flexible enough to handle the heterogeneity of web resources and different annotation needs. Annotations created in the framework are treated as first-class objects so that existing digital library services, such as classification and visualization, can be used for them. The framework has been built into the G-Portal system.

References

1. ADEPT/DLESE. ADN joint metadata content model. http://www.dlese.org/metadata/.
2. H. Chen, B.R. Schatz, T.D. Ng, J.P. Martinez, A.J. Kirchhoff, and C. Lin. A parallel computing approach to creating engineering concept spaces for semantic retrieval: The Illinois Digital Library Initiative Project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):771–782, August 1996.
3. eXtensible Markup Language. http://www.w3c.org/xml/.
4. Jose Kahan and Marja-Ritta Koivunen. Annotea: an open RDF infrastructure for shared web annotations. In *Proceedings of the Tenth International World Wide Web Conference (WWW 10)*, pages 623–632, Hong Kong, China, May 1-5 2001.
5. Ee-Peng Lim, Dion Hoe-Lian Goh, Zehua Liu, Wee-Keong Ng, Christopher Soo-Guan Khoo, and Susan Ellen Higgins. G-portal: A map-based digital library for distributed geospatial and georeferenced resources. In *Proceedings of the Second ACM+IEEE Joint Conference on Digital Libraries (JCDL 2002)*, Portland, Oregon, USA, July 14-18 2002.
6. Ee-Peng Lim, Zehua Liu, and Dion Hoe-Lian Goh. A flexible classification scheme for metadata resources. In *Proceedings of Digtial Library – IT Opportunites and Challenges in the New Millennium (DLOC 2002)*, Beijing, China, July 8-12 2002.
7. XML Schema. http://www.w3c.org/xmlschema/.
8. T. Smith, G. Janee, J. Frew, and A. Coleman. The Alexandria Digital Earth ProtoType system. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries (JCDL 2001)*, pages 118–119, Roanoke, VA, USA, June 2001.
9. T. Sumner and M. Dawe. Looking at digital library usability from a reuse perspective. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries (JCDL 2001)*, pages 416–425, Roanoke, VA, USA, June 2001.
10. Robert Wilensky. Digital library resources as a basis for collaborative work. *Journal of the American Society of Information Science*, 51(3):228–245, 2000.
11. Catherine Marshall Xerox. The future of annotation in a digital (paper) world. In Harum and Twidale, editors, *Successes and Failures of Digital Libraries*, pages 97–117. Urbana-Champaign: University of Illinois, 2000.

Building Policy, Building Community: An Example from the US National Science, Technology, Engineering, and Mathematics Education Library (NSDL)

Ellen Hoffman[1] and Edward A. Fox[2]

[1] Teacher Education, Eastern Michigan University
and Vice Chair, NSDL Policy Committee, Ypsilanti, MI 48197 USA
ehoffman@online.emich.edu

[2] Department of Computer Science, Virginia Tech
and Chair, NSDL Policy Committee, Blacksburg, VA 24061-0106 USA
fox@vt.edu

Abstract. Developing a major national resource such as the US National Science, Technology, Engineering, and Mathematics education Digital Library (NSDL) requires policy and technical expertise. Policy issues are as much a challenge as infrastructure in the construction of such a major digital library. This paper focuses on the policy making structures that have developed since 2000 when the US National Science Foundation began its funding for NSDL. The authors are the Vice Chair and Chair of the NSDL Policy Committee and report here on issues of voice, authority, and central issues for policy construction as these have emerged in the NSDL context.

The US National Science, Technology, Engineering, and Mathematics (STEM) Education Digital Library (NSDL), which officially opened its virtual doors in early December 2002, is still an emerging resource. As envisioned in a 2001 white paper by the NSDL community, the library "will be a gateway to diverse digital collections of quality [science, mathematics, engineering and technology] SMET educational content and services developed by a rich array of SMET educators" [1]. The initial basis for the library is the result of multiple funded projects representing collections and services developed through formal and informal partnerships among universities, K-12 schools, professional organizations, government agencies, non-profit organizations, and corporations. Additional collections funded through non-NSF government programs and private foundations also have been provided by other initial contributors. The library is tied together through an infrastructure developed by a consortium funded through a Core Integration grant with broad responsibilities for outreach, technical systems, and collection and service coherence. When the NSDL opened its virtual doors in December 2002, it was the culmination of over a decade of research and collaboration among librarians, technologists, and educators in the public and private sectors.

From the beginning, library developers recognized the challenges in devising a major national resource and the need for community building processes that would

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 299-300, 2002.

engage many talented individuals into contributing to the library's growth. While the library funding flows to a limited group of partners, the goals for the library emphasize more extended user involvement and input in NSDL growth and development. As a result, the community (currently defined as the leaders of the NSF funded projects) adopted a program for governance to help define relationships, provide forums for participation by many different individuals, and build a structure through which policy decisions could be made. As noted in the NSDL Interim Governance Document [2], the governance model is designed to:

- Create a public sphere of influence that promotes partnering and shared vision
- Establish a framework for accountability that also accommodates course corrections
- Foster communications among stakeholders, including end-user feedback
- Balance differing interests, with sensitivity to minority perspectives
- Provide checks and balances in formulating policy and decision making
- Reflect the special roles and commitments of NSF awardees."

The governance process, as it is currently implemented, is summarized in Table 1. There is a focus on community building.

Table 1. Summary of NSDL Governance Structures

Governance Structure	**Membership**
NSDL Assembly	One representative from each NSF funded project; additional organizations can be added as NDSL grows
NSDL Policy Committee	Elected by Assembly as policy setting body and representatives to work with Core Integration Project
Standing Committees (five: Technology, Content, Community Services, Educational Impact and Evaluation, Sustainability)	Open membership to any interested individual; policy development and recommendations to the Policy Committee; work done primarily by email and conference telephone calls
Subcommittees, Task Forces	Work groups to focus on specific issues and needs for NSDL, created by Standing Committees as needed (one to date: K-12)
Special Interest Groups	Email groups around topics defined by special interest (may be disciplinary, library management issues, design groups, etc.)

As the new governance structure stabilizes and matures, many challenging issues remain to be resolved. The success of the new policy structure is believed to be only as strong as the community that NSDL represents.

References

1. "Pathways to Progress: Vision and Plans for Developing the NSDL."
http://www.smete.org/nsdl/index.html June
http://www.dlib.org/dlib/january02/arms/01arms.html.
2. "Interim NSDL Governance Document, Adopted December 2001."
http://www.nsdl.org/policy

Building a Digital Library from the Ground Up: An Examination of Emergent Information Resources in the Machine Learning Community

Sally Jo Cunningham

Department of Computer Science, University of Waikato,
Private Bag 3105, Hamilton, New Zealand
sallyjo@cs.waikato.ac.nz

Abstract. A "bottom up" approach to digital library development begins with an investigation of a community's information needs and available documents, and then designs a library to organize those documents so as to fulfill the community's needs. The 'home grown', informal information resources developed by and for the machine learning community are examined as a case study.

1 Introduction

The difficulty with providing a novel digital library interface or novel contents lies in ensuring that the resulting digital library will be useful to, and usable by, its intended user community. The problem is to discover what documents the community finds useful, what natural organizations of those documents exist, and what vocabulary is used. An ethnographic approach seems appropriate for discovering how to tailor generic digital library architectures to a particular community—that is, to examine the above issues from the target community's point of view, in the community's own words. This paper takes such an approach by examining the WWW-based information resources created by members of the machine learning (ML) research community, for use by that community.

2 Information Resources Analyzed in This Case Study

The following information resources were selected for analysis in this paper:

- *Online machine Learning Resources*: http:/www.ai.univie.ac.at/oefai/ml/ml-resources.html
- *KDnuggets*: http://www.kdnuggets.com
- *David Aha's Machine Learning Page*: http://www.aic.nrl.navy.mil/~aha/research/machine-learning.html
- *Mlnet*: http://www.mlnet.org
- *The Data Mine: http://www.cs.bham.ac.uk/~anp/TheDataMine.html*

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 301-302, 2002.

3 Document Types and the Information Needs They Support

The information resources listed in Section 2 contain a number of different document types, which provide answers to these questions:

- *Who is working in this field?* Collections of homepages of machine learning researchers provide an excellent introduction to members of that community, and homepages of research groups give the 'big picture' of how different research agendas fit together.
- *What events are occurring?* Events include conferences, workshops, funding opportunities, and data mining competitions. Conference websites provide a sense of context; groupings of the papers at a conference provide valuable snapshots of the state of various sub-fields at that particular time.
- *Where can scientific 'results' be found?* These resources contain links to formally refereed articles.
- *Where can 'raw material' for research be located?* Several collections of test datasets exist, as well as implementations of ML algorithms.
- *How do sub-fields, and the field as a whole, define themselves?* Machine learning is an emerging field, its boundaries in flux. What is the relationship among machine learning, data mining, and knowledge discovery in databases—are they the same? Each website implicitly defines its own boundaries of machine learning through the documents and links that it includes.

4 Organization of Documents for Browsing

Browsing, rather than searching, is intended as the primary interaction technique for the websites analyzed in this paper. The subject topics indicate that this community recognizes and uses a far richer and finer-grained categorization of this field than the formal computing resources. Further, the topics are not hierarchically organized. Little or no attempt is made to define the relationships between topics, and there is no attempt to construct a comprehensive ontology for the discipline as a whole.

The machine learning websites analyzed cover a far broader range of document types than digital libraries designed for computer scientists. The additional types of documents and information provide richer support for some tasks that a conventional collection also supports—e.g. browsing by subjects/topics and exploring relationships among researchers.

Subscription Clubs for E-journals: Indian Initiatives

I.R.N. Goudar and Poornima Narayana

Information Centre for Aerospace Science and Technology
National Aerospace Laboratories, Bangalore – 560 017, India
{goudar,poornima}@css.nal.res.in

1 E-journals and Consortia

E-publishing has brought about a revolution in journal publication, subscription, access and delivery. While dwindling library budgets and the growing number of journals have forced libraries to form consortia for the purpose of accessing e-journals, other role players (primary publishers, vendors, etc.) have their own reasons to encourage the cause. Consortia can have its own structure of governance and can act as a corporate body on behalf of its members, with set goals and benefits: e.g. an increased user and access base, optimal use of funds, infrastructure development and adoption of IT, and an enhanced image of library services. Other by products include union catalogues, shared expertise and library systems, access to other electronic resources, archiving, development of standards and above all ILL. Consortia may be centralised or decentralised; participant-oriented or purpose-oriented or client-oriented.

2 Consortia Values: Libraries vs. Publishers

Libraries	**Publishers**
• Lower price	• Pricing/Education
• Usefulness	• Usage Reporting
• Member driven	• Linking/Delivery
• Full text access	• Interface options
• Access to Internet resources	• Indexing/Filtering
• Combined purchasing power	• Gain credibility with libraries
• Simplify purchase procedures	• Increased marketing
• Distribute financial and other risks	• Reduced cost of production
• Increase participation of members	• Reduced surcharges like mailing
• No storage & documentation problem	• Less extra effort and expenditure for giving access to new customers
• Instant Access	
• Quality of services	• Get consortium tool
• Free flow of information	○ Gather library information
• Sharing of ideas, information	○ Invoice libraries
• Contribution of time, resources	○ Product support

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 303-304, 2002.

3 E-journal Pricing Models and Consortia Issues

With no universally accepted pricing and licensing models, pricing can vary from publisher to publisher, and even between libraries from the same publisher, depending upon the factors and issues listed below.

Influencing Factors	Publisher Issues
Quantum of business	Free titles on Internet
Number of consortia members	Free access against print subscription
Types of institutions	All titles of a publisher for fixed fee
Contract period	Surcharge on print subscription
Number of IP enabled nodes	Discounts for electronic journals
Number of campuses	Capped annual inflation
Value added services	Discounts on non-subscribed titles
Rights to archive	Access to subject clusters of the journals
Perpetual access	Protection of current revenue
Training facilities	Uncertainty of new subscription
Multi year agreement	Single point payment

Consortia management has to deal with three broadly defined categories of issues - strategic (mission, funding, geographical coverage, types of libraries); tactical (programs, services, technology); and practical (governance, staffing , payment).

4 Indian Consortia Initiatives

Even high budget libraries are unable to maintain a minimum core journal subscription level. This, together with increased awareness about e-sources, has led to some consortia being formed in India. Six Indian Institutes of Management have joined to access the e-journals and databases of Elsevier, Kluwer, etc. Forty national laboratories of the Council of Scientific and Industrial Research have access to ScienceDirect. The Ministry of Human Resources Development is finalizing a consortia deal for IISc, IITs, IIMs. INFLIBNET (INFormation and LIBrary NETwork) gives access to JCCC (Journals Customised Contents for Consortia), a new local product covering the acquisitions of all participating libraries. The ICICI Knowledge Park provides access to J-Gate, a TOC/database service of more than 8000 STM journals, with a linking facility to fulltext.

Constraints include a lack of awareness about consortia benefits; slow adoption of e-sources by users; rigid mindset of librarians; inadequate funding; insistence of publishers for single point payment; rigid financial and audit rules; asset definition against payment; lack of good infrastructure and telecommunications link; shortage of trained manpower; and big brother attitude of major member libraries.

Consortia are time consuming, hard to build and sustain. But they are a potent force as they reduce unit cost of e-information, increase resource and user bases, help libraries to achieve more together than they can alone. It is satisfying to note that some librarians in India have dared to form consortia to provide access to e-journals - to meet the information needs of Indian academic and R & D workers.

A Multilingual Multi-script Database of Indian Theses: Implementation of Unicode at Vidyanidhi

Shalini R. Urs, N.S. Harinarayana, and Mallinath Kumbar

Department of Library and Information Science
University of Mysore
Mysore 570 006, India

1 Introduction

The theses submitted to universities are a rich and unique source of information on a given topic in terms of breadth as well as depth of treatment and coverage. The insight into the theoretical and methodological foundations of a topic, the data sets, and the exhaustive bibliography-all these attributes make the doctoral thesis a rich resource. A thesis represents the outcome of a focused and extensive study involving intellectual labour of more than three years. Archiving and enhancing access to theses is the main mission of Electronic Theses and Dissertations (ETD) initiatives worldwide. ETDs gracefully overcome many of the problematic issues relating to the archiving and accessing of theses.

India with its enormous system of higher education spanning two hundred and eighty one universities (including deemed universities) is a reservoir of extensive doctoral research in the country [1]. Hard statistics regarding the doctoral research output is not available as there is no system/mechanism to deposit, document and archive the Indian theses. Estimates of doctoral research output can only be extrapolations of available statistics. Based on such extrapolations we estimate that annually 25,000 to 30,000 doctoral theses are produced in India. English is the predominant language of the Indian theses. However, India with its legacy of linguistic diversity, rich Indic literature and the increasing vernacularisation of higher education has spawned considerable research output in Indic languages/scripts as well. It is estimated that nearly 20 to 25 percent of the theses produced in India are in Indic languages and scripts. This trend is on the increase as a consequence of the increasing vernacularisation of higher education and the government policy of promoting regional languages. Documenting the research out put in all languages of India is one of the missions of *Vidyanidhi.*

2 Vidyanidhi

Indian Digital Library of Electronic Theses is an initiative currently underway at the Department of Library and Information Science, University of Mysore [2] [http://www.vidyanidhi.org.in]. The National Information System of Science and Technology (NISSAT), Department of Scientific and Industrial Research, Government of India has sponsored this project. Vidyanidhi looks at ETDs primarily as a means of –

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 305-314, 2002.

- Expanding access
- Enduring archiving
- Extending dissemination
- Enlarging audience
- Enhancing expressiveness

to/of Indian doctoral theses.

In line with its mission and objectives Vidyanidhi is conceived as having two main layers of the digital library- the metadata layer and the full text layer. We are developing these two layers as two separate products. The top layer is the metadata database and the other inner layer is the full text. This paper outlines the issues relating to the metadata layer-bibliographic database of Indian theses.

3 Vidyanidhi Multilingual Multi-script Database

Given the multilingual nature of the Indian theses literature, one of the major questions to be resolved was the language and script of the database. There are two possible approaches to resolving this question. One is to follow the transliteration approach i.e irrespective of the language of the thesis; the bibliographic records would be in Roman script. The other is to describe the theses (metadata) in the language of the theses. We at Vidyanidhi resolved to adopt the second approach that is appropriate in a multilingual content scenario. Having chosen to adopt this policy- of Vidyanidhi database is to be a multilingual and multi-script one, the major issue to confront and address was that of the challenges of multilingual and multi-script databases.

It was, thus our endevour to identify and implement a suitable database design and application taking into consideration the demands of Indic languages and scripts and our specifications for a truly multilingual-multi script database. Encoding and software platforms and tools were the two critical issues that needed to be resolved. The complexities and perplexities of encoding Indic scripts may be summarized as follows [3,4,5]-

- Indic scripts are syllable oriented-phonetic based with imprecise character sets
- The different scripts look different (different shapes) but have vastly similar yet subtly different alphabet base and script grammar
- The Indic characters consist of consonants, vowels, dependent vowels-called ‘matras’ or a combination of any or all of them called conjuncts.
- Collation (sorting) is a contentious issue as the script is phonetic based and not alphabet based

There are three possible approaches to managing the encoding issue of Indic scripts. They are-

- Transliteration approach-where in the Indic characters are encoded in either ASCII or any other proprietary encoding and use glyph technologies to display and print Indic scripts-currently the most popular approach for desktop publishing.
- Develop an encoding system for all the possible characters/combinations running into nearly 13,000 characters in each language-with a possibility of

a new combination leading to a new character- an approach developed and adopted by the IIT Madras development team[3]

- Adopt the ISCII/Unicode encoding

In view of the inherent limitations and the demands of a database, the first approach is highly unsuited for database applications. The second approach is (claimed by the developers) said to be very well suited for linguistic text processing.

Given the importance of ‘data’ and ‘data manipulation’ in a database and the versatility of the UNICODE, we at Vidyanidhi resolved to pursue the UNICODE approach. One major factor in favour of UNICODE is its ability to simultaneously handle more than one Indic Script- a feature not possible with ISCII. The factor that weighed in favour of UNICODE was the software support. Today most database products and applications support UNICODE

4 UNICODE and INDIC Scripts

Unicode is a standard that is designed to support all of the world’s scripts. Unicode provides a unique code for every character, regardless of platform, program, or language. The latest version of the Unicode standard is the 3.2.0 released on 27th March 2002. The standard version is the 3.0 [6].

The Table 1 below gives a list of scripts, Unicode range, languages which use these scripts.

Table 1. List of scripts

Script	Unicode Range	Major Languages
Devanagari	U+0900 to U+097F	Hindi, Marathi, Sanskrit
Bengali	U+0980 to U+09FF	Bengali, Assamese
Gurumukhi	U+0A00 to U+0A7F	Punjabi
Gujurati	U+0A80 to U+0AFF	Gujarati
Oriya	U+0B00 to U+0B7F	Oriya
Tamil	U+0B80 to U+0BFF	Tamil
Telugu	U+0C00 to U+0C7F	Telugu
Kannada	U+0C80 to U+0CFF	Kannada
Malayalam	U+0D00 to U+0D7F	Malayalam

One of the critical considerations for UNICODE implementation for Indic scripts was the support for UNICODE and the attendant support for Indic Scripts. Microsoft provides excellent support for UNICODE as well as Indic scripts in many of its products and tools[7]. MS SQL support for UNICODE and Windows XP support for most of the Indic Scripts, and the availability of UNISCRIBE, True Type fonts such as Tunga (for Kannada language) and Mangal font(for Devanagari) UNICODE Arial font and other utilities make Microsoft platform and environment an attractive and feasible solution for the implementation of UNICODE for a multilingual and multi-script database such as the Vidyanidhi

The Vidyanidhi implementation of UNICODE has been on the Microsoft platform. The database is the MS SQL on Windows XP and 2000 server platform. The scripting languages used are Javascript, ASP and HTML. Currently the Vidyanidhi database has upwards of 22,000 records of which 19,000 are for theses in English, 2,200 for theses in Hindi and 640 are for in Kannada.

This paper reports the Vidyanidhi experiences in the design, development and implementation of UNICODE for Hindi and Kannada scripts in a database application. The implementation has been fairly successful in meeting the requirements of a database as well the Hindi and Kannada scripts, especially in respect of data entry, display and collation of the records. The Vidyanidhi experience clearly demonstrates the suitability and the simplicity of implementing UNICODE for Indic languages. Despite the robustness, technical soundness and the practical viability of UNICODE, adoption of UNICODE for Indic scripts has not been widespread-almost non-existent. To the best of our knowledge, there has been no report of any attempt in this area. We believe that the reasons for the non-implementation of UNICODE are largely due to the misconceptions and misconstructions rather than the limitations of UNICODE. Though there are a few problem areas, which are not satisfactorily addressed, UNICODE is perhaps the best option available for a multi language multi-script environment such as the Indic scripts. The UNICODE and the Microsoft software tools have fairly satisfactorily handled the issues of data entry, display and collation.

5 Vidyanidhi Database: Design and Implementation

Designing the underlying structure and functioning of a multilingual database is a vexing matter. The Vidyanidhi database has been conceived and implemented on the following lines-

- An integrated Vidyanidhi database in Roman and Indic scripts. The integrated database is the complete database (Master database) with metadata records for theses in all languages in a single sequence. These records are in Roman script as well as in Indic scripts (Devanagari & Kannada), but encoded in Unicode UTF 8.
- Each script /language will have one table. Currently there are three separate tables for the three scripts- one each for Roman, Hindi (Devanagari), & Kannada
- The theses in Indic languages will have two records -one in the Roman script (transliterated) and the other in the vernacular. However the theses in English will have only one record (in English)
- The two records are linked by the ThesisID number-a unique id for the record
- The bibliographic description of Vidyanidhi follows the ThesisMS Dublin Core standard adopted by the NDLTD and OCLC [8]

6 Vidyanidhi Architecture

The Vidyanidhi digital library is built on Windows platform and Microsoft environment with the MS SQL as back end, front ends in ASP and the processing achieved through the Javascript programs (Fig. 1).

7 Accessing and Searching the Vidyanidhi Database

Currently the Vidyanidhi website and the search interfaces are in English (Indic Language web pages are in the pipeline). However searching is facilitated in English as well as Indic Languages. There are two possible ways of searching the Vidyanidhi database. One can search the integrated database that has all the records irrespective of language/script or the respective vernacular database having records of theses in that language only. The difference between the two approaches is –one affords search in the English language and the other in the vernacular. The first approach also provides for viewing records in Roman script for all theses-search output- that satisfy the conditions of the query and also an option for viewing records in vernacular script for theses in vernacular. The second approach- enables one to search only the vernacular database and thus is limited to records in that language. However, this approach enables the search to be in the vernacular language and script.

Access to and the search interaction in the Vidyanidhi Database is facilitated in the following manner. When one gets to the Metadata Database (Fig. 2 a) one can choose to search by language, either –

- All
- Hindi
- Kannada

While the 'All' option facilitates the searching of the integrated database, the other two- Hindi and Kannada limits the access and search to only the respective languages. We will illustrate the search process with the following example-

- Select all
- Select the field –
  - Author
  - Keyword
  - Title
  - University
  - Subject
  - Guide
  - Language
  - Year

Further narrowing of the search is also provided- where in one can combine any two fields. Once the database is queried using any of the above conditions, then the search results indicates the number of records output under – All, Hindi, Kannada. The searchers will have the option to click on their choice and view records in either Roman script only or Roman as well as respective scripts

One can also search the vernacular database only by selecting the appropriate language. However this approach would limit the search to records in a single language. As mentioned earlier the main feature of this approach is it facilitates the search process in the language of the database (Hindi/Kannada). However, in this approach one can view records only in the vernacular script and not in the Roman script.

8 Unicode and Indic Scripts

Notwithstanding the technical soundness, robustness and fairly well thought out arrangements of alphabets and the availability of adequate software support, Unicode

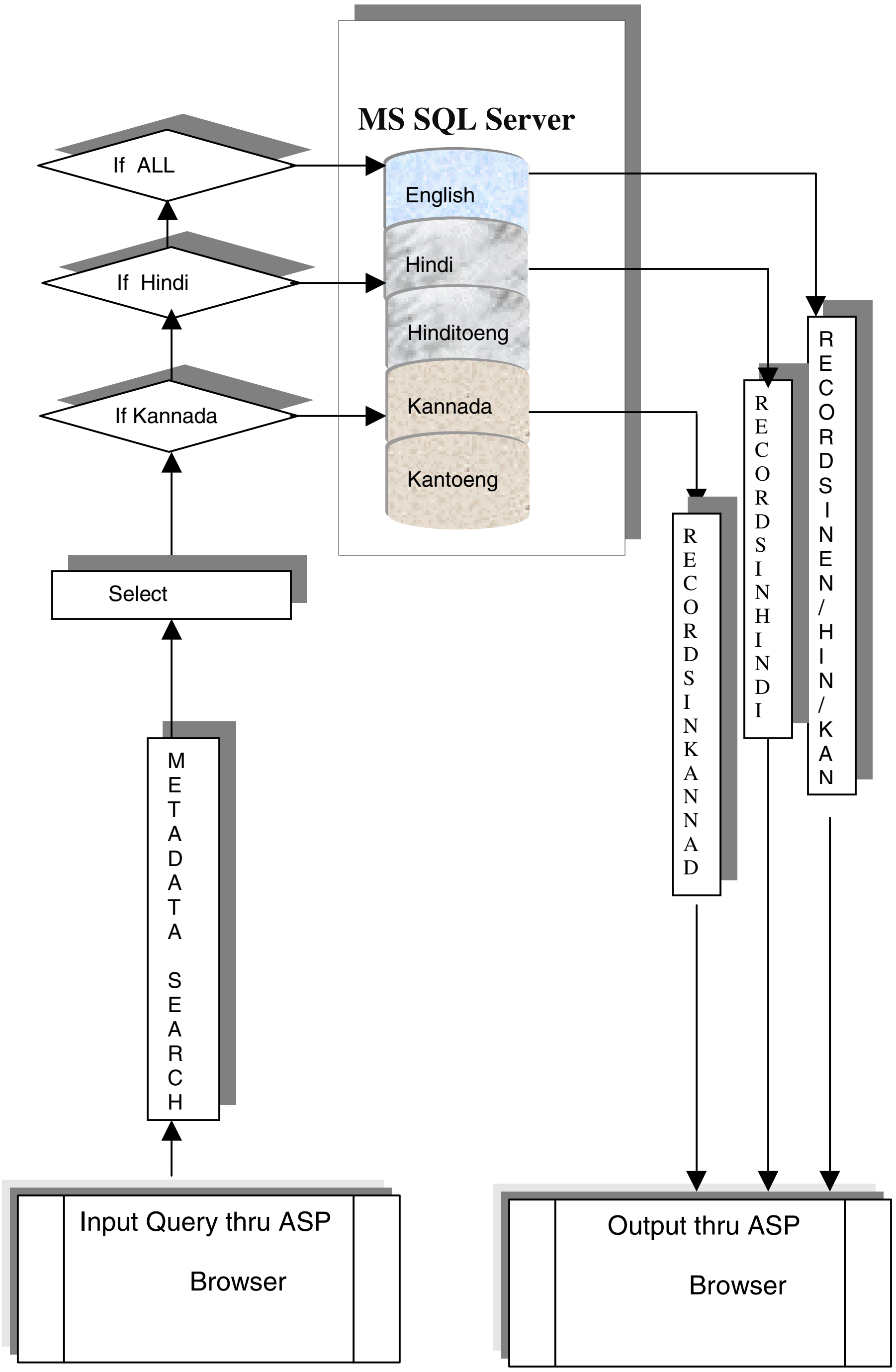
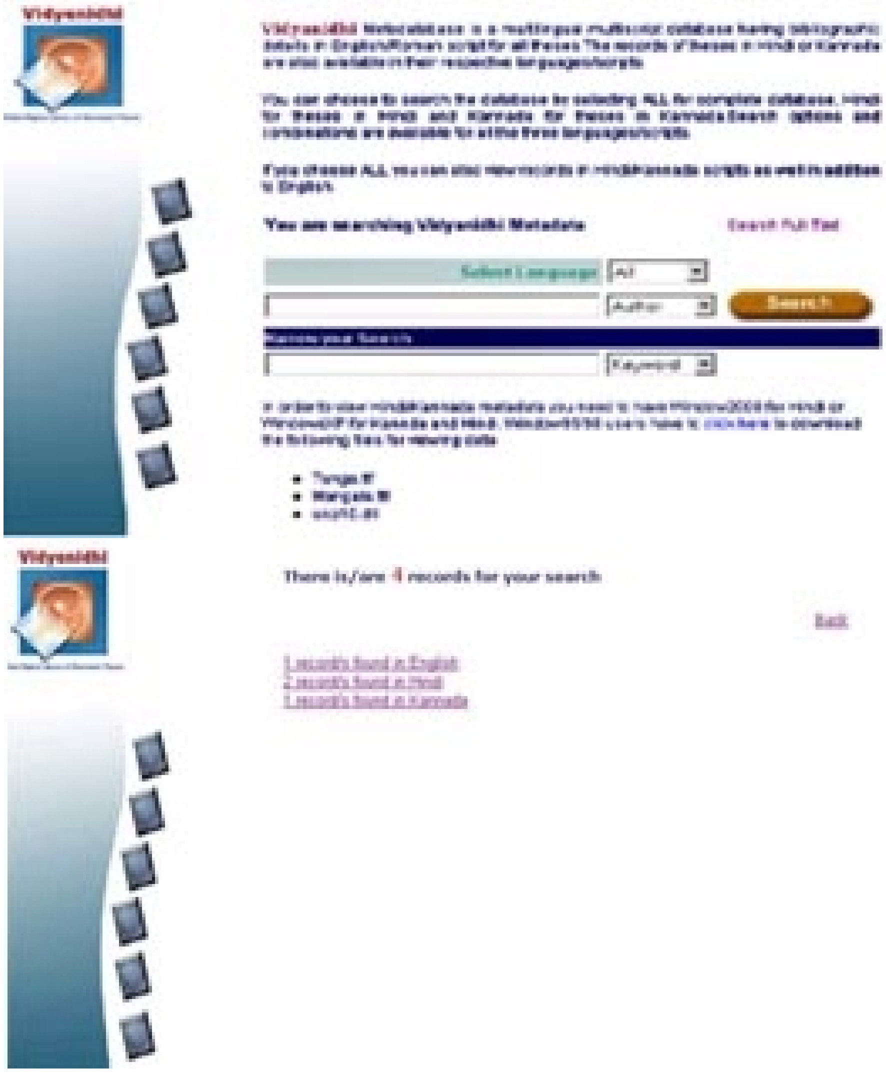


Fig. 1. Architecture of Vidyanidhi Metadata Database

implementation has been very rare. Despite our systematic search of the bibliographic databases such as Web of Science, INSPEC, and LISA we did not come across a single reference to any implementation of Unicode for Indic scripts. We believe that the reasons are to be found in the misconceptions and misconstructions that surround Unicode rather the 'reality' of the limitations of Unicode. Most of the misconceptions centre around 'supposed problems/inability ' of Unicode. The problem areas are -

- Data Input
- Display and printing
- Collation



Vidyanidhi Metadatabase is a multilingual multiscript database having bibliographic details in English/Roman script for all theses. The records of theses in Hindi or Kannada are also available in their respective languages/scripts.

You can choose to search the database by selecting ALL for complete database, Hindi for theses in Hindi and Kannada for theses in Kannada.Search options and combinations are available for all the three languages/scripts.

If you choose ALL, you can also view records in Hindi/Kannada scripts as well in addition to English.

You are searching Vidyanidhi Metadata

Search Full Text

Select Language: All

Author

Search

Narrow your Search

Keyword

In order to view Hindi/Kannada metadata you need to have Windows2000 for Hindi or WindowsXP for Kannada and Hindi. Windows95/98 users have to click here to download the following files for viewing data

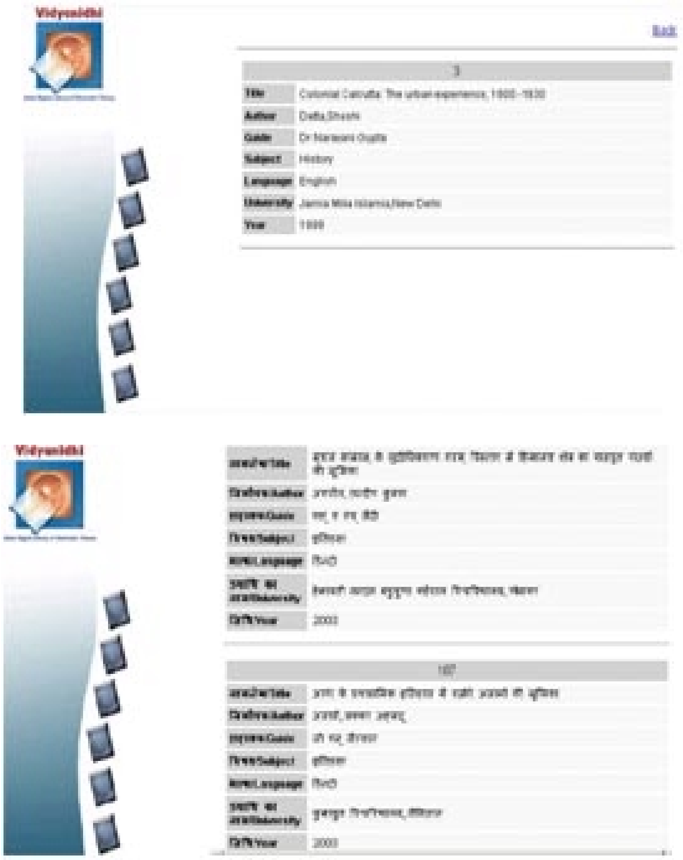
- Tunga.ttf
- Mangal.ttf
- [illegible]

There is/are 4 records for your search

Back

1 record/s found in English
2 record/s found in Hindi
1 record/s found in Kannada

Fig. 2a. All language search

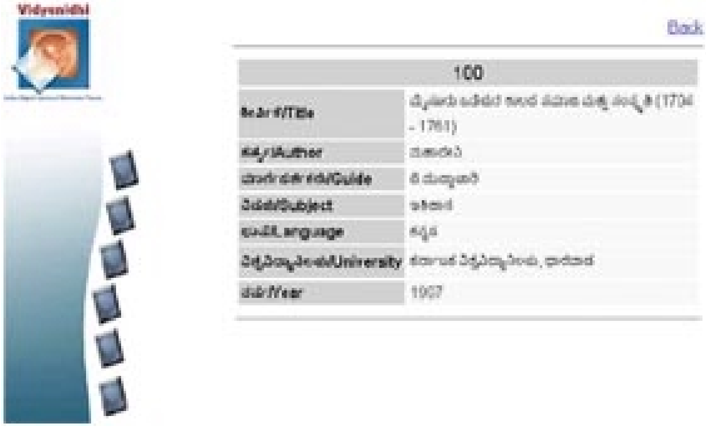


3	
Title	Colonial Calcutta: The urban experience, 1800 -1930
Author	Datta,Shashi
Guide	Dr Narayani Gupta
Subject	History
Language	English
University	Jamia Millia Islamia,New Delhi
Year	1999

शीर्षक/Title	[illegible] की भूमिका
शोधार्थी/Author	[illegible]
मार्गदर्शक/Guide	[illegible]
विषय/Subject	इतिहास
भाषा/Language	हिन्दी
उपाधि का विश्वविद्यालय/University	[illegible] विश्वविद्यालय, [illegible]
तिथि/Year	2003

107	
शीर्षक/Title	[illegible] की भूमिका
शोधार्थी/Author	[illegible]
मार्गदर्शक/Guide	[illegible]
विषय/Subject	इतिहास
भाषा/Language	हिन्दी
उपाधि का विश्वविद्यालय/University	[illegible] विश्वविद्यालय, [illegible]
तिथि/Year	2003

Fig. 2b. Display vernacular records (Hindi)



100	
ಶೀರ್ಷಿಕೆ/Title	[illegible] (1704 - 1761)
ಲೇಖಕರು/Author	[illegible]
ಮಾರ್ಗದರ್ಶಕರು/Guide	[illegible]
ವಿಷಯ/Subject	ಇತಿಹಾಸ
ಭಾಷೆ/Language	ಕನ್ನಡ
ವಿಶ್ವವಿದ್ಯಾನಿಲಯ/University	ಕರ್ನಾಟಕ ವಿಶ್ವವಿದ್ಯಾಲಯ, ಧಾರವಾಡ
ವರ್ಷ/Year	1997

Fig. 2c. Display of vernacular records(Kannada)

8.1 Unicode and Data Input for Indic Scripts

One of the misconceptions regarding Unicode centers on the unconventional Keyboard layout for Indic scripts. Many believe that as compared to the more common and familiar QWERTY Key board layout and transliteration to Roman alphabet, the atypical keyboard layout is a deterrent the data input. Based on our experimentation and implementation, we find that this apprehension is unsubstantiated [9]. Our testing and comparison with other methods/approaches to Kannada Data input clearly shows that Unicode is in no way slower, inconvenient or 'more' difficult than the others. In certain respects it is slightly better than the others. We were pleasantly surprised at the ease and speed with which our data entry operator could input little over 600 records in Kannada and 2200 records in Hindi. But for minor irritants such as certain punctuation marks (which requires toggling between English and Kannada), in all most all cases data input is no hassle at all. Strictly speaking the punctuation marks are not part of the Kannada repertoire as period is the only punctuation mark conventionally used in the Kannada script. However, all the other punctuation marks have come into usage of late and hence used in Kannada.

8.2 Display and Printing of Kannada /Hindi Script

Barring a few problem areas, the glyphs, shaping engine and rendering of the Kannada and script is reasonably satisfactory. The problem areas may be summarized as follows-

- Formation and rendering of certain consonant and Matra combinations of characters (Ex- ಮೂ , ಯೂ)
- Formation and rendering of a few conjuncts (Consonant combinations with 'matra') (Ex- ತಾOF , ವಾOF)
- Formation and rendering of 'Repha' symbol along with Post fix symbols- (Ex- ರ್ಯ , ರೖ)

Excepting the above cases, handling of the Kannada script in Unicode in the Microsoft platform and environment is more than satisfactory. As a matter of fact, the handling of conjunct clusters is very satisfactory. The maximum number of consonant clusters in Indic script is supposed to be four and Unicode is able to handle the same.

8.3 Collation Issues

Collation is a debatable issue in any language, which is a cross between a syllabic and phonemic writing system. Consensus regarding culturally expected and linguistically correct order is hard to come by in such languages. The above scenario not withstanding, in the matter of sorting (collation) also, our experience with Unicode implementation is very satisfactory. Barring two instances Unicode accomplishes the same sorting as that of “ **Nudi**”- Kannada script software endorsed by the Department of IT, Government of Karnataka [10]. The two contentious instances of collation are –

- Sorting of the consonant Lla (ಳ) immediately after La (ಲ)
- The case of a pure consonants

(Ex- ರ್) being sorted at the end of the consonant and vowel combine ರ – ರಃ

Sample pages [Appendix 1] of the brief records ‘ author’ and ‘title’ only and their sorting demonstrates the suitability and simplicity of implementing Unicode for Indic scripts in a database application.

9 Conclusion

Designing and developing a database involving Indic Languages is a challenging task. Our implementation of Unicode for a multi language multi-script environment such as a database of Indian theses at Vidyanidhi showcases the capabilities and potential of Unicode in meeting this challenge. Considering the complexity and minutiae of a family of Indian languages and scripts with strong commonalities and faint distinctions among themselves, Unicode has been able to handle the situation very ably. Having been convinced of the capability and the facilities of Unicode for handling Indic languages and scripts, we propose to focus our future efforts in the direction of implementing Unicode for other Indic scripts and setting up a test bed of data for experimentation and comparative studies.

References

1. Association of Indian Universities. Universities handbook. 29th edition. New Delhi: Association of Indian Universities, 2002.
2. Urs, Shalini R and Raghavan, K.S. Vidhyanidhi: Indian Digital Library of Electronic Theses. Communications of the Association of Computing Machinery, May 2001.
3. IIT, Madras, India. Multilingual Systems: Representing Text in Indian languages http://acharya.iitm.ac.in/multi_sys/ele_pro.html
4. Srinath Sastry, C.V. Unicode for Kannada script. Report issued by Directorate of Information Technology, Government of Karnataka. Bangalore, India, 2002.
5. Wissink, Cathy. Issues in Indic Language. http://www.unicode.org/notes/tn1/Wissink-IndicCollation.pdf
6. Unicode Consortium. http://www.unicode.org
7. Kaplan, Michael. International Features of SQL Server 2000 http://www.microsoft.com/sql/techinfo/development/ 2000/intfeatures.asp
8. Networked Digital Library of Theses and Dissertations (NDLTD). ETD-ms: an Interoperability Metadata Standard for Electronic Theses and Dissertations. http://www.ndltd.org/standards/metadata/current.html. Updated on 2nd August 2001.
9. Urs, Shalini R, Harinarayana, N.S. and Kumbar, Mallinath. Unicode for encoding Indian Language databases: a case study of Kannada and Hindi scripts. 22nd International Unicode Conference (September 9 –13, 2002), San Jose, California.
10. Kannada Ganaka Parishad. Nudi: Kannada Script Software. http://www.bangaloreit.com/html/education/Nudi.html

A Workbench for Acquiring Semantic Information and Constructing Dictionary for Compound Noun Analysis

Kyung-Soon Lee[1], Do-Wan Kim[2], Kyo Kageura[1], and Key-Sun Choi[2]

[1] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
{kslee, kyo}@nii.ac.jp

[2] Division of Computer Science, KAIST/KORTERM
373-1 Kusung Yusong Daejeon, 305-701, Korea
dwkim@csone.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Abstract. This paper describes a workbench system for constructing a dictionary to interpret compound nouns, which integrates the acquisition of semantic information and interpretation of compound nouns. First, we extract semantic information from a machine readable dictionary and corpora using regular expressions. Then, the semantic relation of compound nouns are interpreted based on semantic relations, semantic features extracted automatically, and subcategorization information according to the characteristics of a head noun, i.e. attributive or predicative. Experimental results show that our method using hybrid knowledge depending on the characteristics of a head noun improves the accuracy rate by 40.30% and the coverage rate by 12.73% better than previous researches using semantic relations extracted from MRDs. As compound nouns are highly productive and their interpretation requires hybrid knowledge, we propose a workbench for compound noun interpretation in which necessary knowledge such as semantic patterns, semantic relations, and interpretation instances can be extended, rather than assuming a pre-defined lexical knowledge.

1 Introduction

The semantic interpretation of compound nouns (consisting of two nouns), or analyzing semantic relations between constituent nouns, is useful for various applications. For instance, it makes possible the regeneration or paraphrasing of natural language sentences. Also, it is useful for the syntagmatic query expansion in information retrieval and for the classification of answer type in question answering.

Much work has been done in the interpretation of compound nouns or noun sequences ([3], [16], [10], [14]). The methods used in interpretation fall into two categories: those based on semantic relations and those based on semantic features. The method based on semantic relation ([17], [16], [15]) interpret a compound noun using rules for lexical patterns and their semantic relations which are extracted from machine readable dictionaries (MRDs). The MRD data is the structured and incomplete information which has limited expressions to define terms. On the other hand, the

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 315-327, 2002.

corpora data is the unstructured but more complete information which includes various expressions between terms. What is necessary is a combination of the corpora and the MRD data, each of which is inadequate, but which, when combined, creates a rich source of semantic information. The method based on semantic features ([1], [6], [8]) considers all the possibilities of combination between semantic features for a modifier noun and a head noun. The feature-based method has difficulty in dealing with ambiguous cases where the same feature sequences can take different relations.

In Korean text, the appearance of compound nouns is a general phenomenon and highly productive. The enumeration of a modifier noun and a head noun can make up a compound noun and constitutes the majority of Korean compound nouns. In addition, predicative nouns as a head of compounds can take a wide variety of case relations with modifier nouns. But most predicative nouns are expressed with the same postpositions in MRDs and corpora. It makes it difficult to interpret semantic relations. Since they have selection restriction like a verb, we can interpret compound nouns for predicative nouns using subcategorization information. In machine translation, though in some cases word-to-word translation works, in many cases semantic interpretation of compound nouns is necessary. For example, '장애인[jang-ae-in](the handicapped) 학교[hak-gyo] (school)' should be translated to 'a school *for* the handicapped' through semantic interpretation of a compound noun, not 'the handicapped school' by word-to-word translation.

In this paper, we present a workbench system to integrate acquisition of semantic information and interpretation of compound noun for Korean semantic analysis. The semantic relation of compound nouns are interpreted based on semantic relations and semantic features extracted automatically from an MRD and corpora, and subcategorization information of predicative nouns according to the characteristics of a head noun. Because the compound nouns are highly productive and the information necessary for interpreting them is complex, we propose a workbench which integrates knowledge acquisition and compound noun interpretation with user's feedback. The system can keep logs for interpretation errors which are useful to analyze error patterns between a lexical pattern and semantic relations.

In the following, we first explain the method we propose for interpreting compound nouns in section 2. Then, in section 3, we explain the actual workbench system, which incorporates the mechanisms explained in section 2 and some additional features.

2 Interpretation of Korean Compound Nouns

Fig. 1 shows the overall architecture of our system for the interpretation of Korean compound nouns. The system acquires semantic information such as semantic relations and semantic features from an MRD and corpora, and constructs a semantic network of nouns. Using this semantic network, together with the subcategorization information of predicative nouns and interpretation rules, compound nouns are interpreted. Below, we explain the construction of the semantic network of nouns and the interpretation of compound nouns. Then we show the experimental results of extraction of semantic information and interpretation of compound nouns.

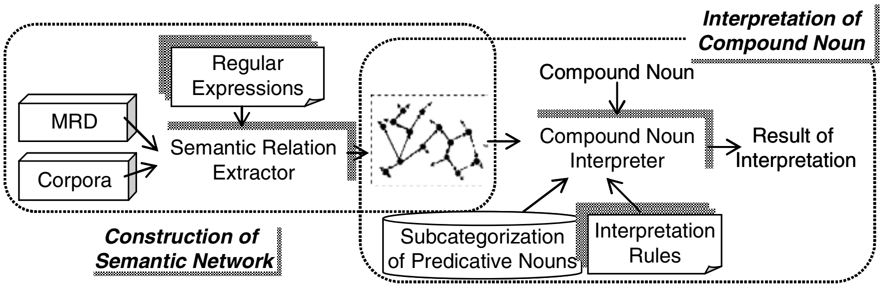


Fig. 1. System architecture for acquiring knowledge and interpreting compound nouns.

2.1 Semantic Information Automatically Extracted from MRDs and Corpora

As knowledge resources to interpret compound nouns, we extract semantic information such as semantic relations and semantic features from MRDs and corpora by defining regular patterns, respectively. What is necessary is a combination of the corpora and the MRD data, each of which is inadequate, but which, when combined, creates a rich source of semantic information.

The semantic relations extracted are as follows: <subject>, <object>, <location>, <time>, <possessive>, <whole-part>, <part-whole>, <instrument>, <purpose>, <material>, <cause>, <caused-by>, and <by-means-of>. The <hypernym> relation between nouns is also extracted. These classifications account for most of the compound noun classes studied previously in theoretical linguistics ([2], [4], [17]). The semantic features extracted are like these: <±abstract>, <±animal>, <±organization>, <±person>, <±location>, <±material> and <±time>. The '+' and '-' sign represent whether a noun have the feature. These features are used in defining interpretation rules unambiguously. Semantic information extracted form semantic network in which a link represents a semantic relation and a node represent a noun. Each node can have semantic features which are used for generalization.

For MRDs, semantic information is extracted from a head word and its definition sentence. The regular pattern consists of a word, part-of-speech tag and some symbols for matching. The symbol '|' means an option and '*' means any matching. We defined the expressions by analyzing a head word and its definition from some part of an MRD. The semantic feature is determined by hypernym of a head word. Corpora have richer terms and various semantic relations than MRDs although it has rare frequency of a regular pattern. The regular patterns for corpora have a simple sentence structure, which is different from that of MRDs since in MRDs, the subject is a head word implicitly (Table 1). Table 2 shows the example for compound noun, its semantic relation and its interpretation. Table 3 shows different semantic information extracted from an MRD and corpora for a noun '설명서(seol-myeong-seo)[manual]'. Using corpora, we can acquire <purpose> relation which is not extracted from MRDs.

From extracted semantic information, MRDs is useful to extract semantic features and hypernym relation between nouns, but difficult to extract various semantic relations. On the other hand, corpora are very useful to extract various semantic relations between nouns. Using various corpora is helpful to solve sparseness problem of semantic information.

Table 1. Regular patterns to extract <purpose> relation from MRDs and corpora.

MRD	regular pattern	A/ncn */jco 위하/pvg+어/ecs
	semantic relation	[head word] –<purpose>→ [A]
Corpora	regular pattern	B/ncn (은|는)/jxt C/ncn */jco 위하/pvg+ㄴ/etm D
	semantic relation	[B] –<purpose>→ [C]

Table 2. Korean compound nouns and their interpretation

Compound Noun		Semantic Relation	Interpretation
N1	N2		
사용자 (user)	설명서 (manual)	<purpose>	사용자를 위한 설명서 (manual for user)
가죽 (leather)	가방 (bag)	<material>	가죽으로 만든 가방 (bag made of leather)
자동차 (car)	바퀴 (wheel)	<whole-part>	자동차의 바퀴 (car’s wheel)

Table 3. Different semantic information acquired from an MRD and corpora

MRD	설명서[seol-myeong-seo](manual)–<hypernym>→글[geul](writings)
Corpus	설명서[seol-myeong-seo](manual)–<purpose>→사용자[sa-yong-ja](user)

2. 2 Interpretation of Korean Compound Noun Depending on Head Nouns

Compound nouns are interpreted using information on semantic network and subcategorization information of predicative nouns. The procedure of interpretation differs according to the type of head nouns, i.e. attributive or predicative. In the case of attributive nouns, the system interprets based on semantic relations and semantic features. In the case of predicative nouns, the system interprets based on semantic relation and subcategorization information.

In Korean, predicative nouns as a head of compounds can take a wide variety of case relations with modifier nouns. But most predicative nouns are expressed with the same postpositions such as ‘이[i]’ for subjective postposition and ‘를[leul]’ for objective postposition in MRDs and corpora. It makes difficult to interpret semantic relations. Since they have selection restriction like a verb, subcategorization information is useful to decide semantic relations in lexical patterns with ambiguity. Semantic features are used to supplement the interpretation based on semantic relation. The interpretation system gives weights to the results by interpretation rules according to their distance on semantic network and the types of semantic relations.

2.2.1 Interpreting Compound Nouns with Attributive Heads

When a head noun is attributive in a compound noun, the system interprets it based on semantic network and interpretation rules.

On the semantic network, some compound nouns are connected with direct link ('가죽[ga-juk](leather) 가방[ga-bang](bag)' and some compound nouns are connected indirectly through several links ('화장품[hwa-jang-pum](cosmetics) 가게[ga-ge](store)' in Fig. 2).

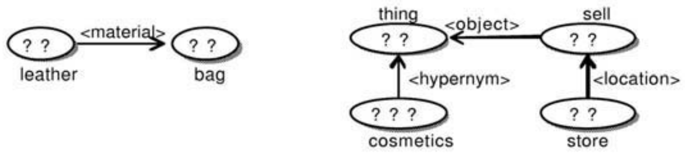


Fig. 2. Compound nouns connected with direct or indirect links on semantic network.

To interpret a compound noun connected by several links in semantic network, the system use interpretation rules for inference. Interpretation rules are constructed by using semantic relations and semantic features. The rules with semantic features are used when semantic information is lacking or insufficient for determining an interpretation. For <possessive> relation, the system interprets it by the rule based on semantic feature since the relation is not extracted from MRDs and corpora, which is represented with the same expression such as '의[eui]'. Table 4 shows interpretation rules for <material> relation.

Table 4. Interpretation rules for <material> relation.

Rule	Modifier noun	Head noun
Rule 15	<material>	semantic network
Rule 16	semantic network	<object>
Rule 17	<+material>	<-abstract>

Two nouns on the semantic network can be connected by direct and indirect links. Therefore, the system has to choose proper interpretation among those links. The system selects the best interpretation by weighting to the links according to the distance and the type of semantic relation. In other words, if two nouns are connected by a direct link, the interpretation of the link has the highest weight. According to the number of bridge node to connect two nouns increase, the weight of interpretation becomes lower. If two nouns are connected by several indirect links, we give higher priority according to the type of semantic relation as follows: Priority 1: <hypernym>, priority 2: <part>, <material>, priority 3: <object>, <subject>.

2.2.2 Interpreting Compound Nouns with Predicative Heads

When a head noun is predicative which represent state or action, the system interprets it based on semantic relations and subcategorization information.

Predicative noun and suffix ‘하다[ha-da](do)’ or ‘되다[doi-da](become)’ form a verb in Korean. For example, the combination of a predicative noun ‘거래[geo-lae](transaction)’ and suffix ‘하다’ make a verb ‘거래하다 [geo-lae ha-da](transact)’. Therefore predicative verb governs cases. Subcategorization is useful in case of subject and object relation., because they are difficult to be extracted from MRDs since most regular patterns are expressed with the ambiguous patterns such as ‘이[i]’ and ‘를[leul]’ which represent postpositions for subjective and objective, respectively. For example, a compound noun, ‘주식 [jusik] (stock) 거래 [geo-lae] (transaction)’, is interpreted as <object> relation based on subcategorization of ‘거래’ which is predicative and ‘증권[jeung-kwon]’ which has <hypernym> relation to ‘주식’ on semantic network. Matching is tried using nouns of N2 or concept such as <thing>, and their links on semantic network (in Figure 3).

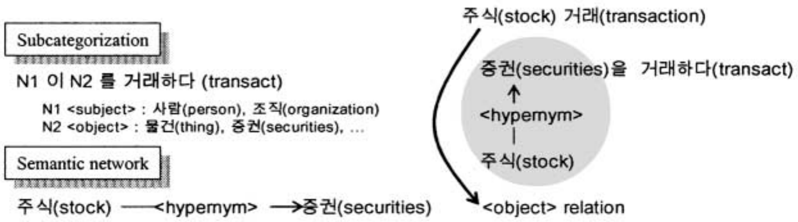

주식(stock) 거래(transaction)
Subcategorization
N1 이 N2 를 거래하다 (transact)
N1 <subject> : 사람(person), 조직(organization)
N2 <object> : 물건(thing), 증권(securities), ...
증권(securities)을 거래하다(transact)
<hypernym>
주식(stock)
Semantic network
주식(stock) <hypernym> 증권(securities)
<object> relation

Fig. 3. Example of interpreting a compound noun with a predicative head.

2.3 Experiments

We experimented for the extraction of semantic information and the interpretation of compound nouns. Table 5 shows resource statistics for an MRD and corpora to extract semantic information.

Table 5. The statistics of experimental data from an MRD and corpora.

Resources		The number of sentence	Ratio
MRD (definition)		8780 (for 5956 head words)	3.22%
Corpora	Fiction	251279	91.89%
	Essay	13385	4.85%
Total		273444	100%

2.3.1 Experiment 1: Extraction of Semantic Information

We extracted 18,262 semantic relations among 5,235 terms from MRD. In case of both MRD and corpus, we extracted 53,644 semantic relations among 10,255 terms by defining 128 regular expressions. 3,160 terms and 6,298 semantic relations are redundantly extracted in both MRD and corpus. The average number of semantic relations for one term is 5.23. We evaluated 500 randomly selected semantic relations. The precision are 80.6% and 82.6% for an MRD and corpora, respectively (Table 6). To extract 7 semantic features, we used only an MRD and decided it depending on hypernym of a head word. The precision is 97.32 % for 996 terms.

Table 6. The number of semantic relations and evaluation results (Corr: the number of correct answer, Samp: the number of sample randomly selected).

Semantic Relation	MRD		Corpus	
	Total	Corr /Samp	Total	Corr/Samp
BY-MEANS-OF	317	4 / 5	576	2 / 5
CAUSE	4		7	
CAUSE-BY	14		25	
HAS-OBJECT	5523	144 / 164	6798	43 / 49
OBJECT-OF	0		21379	283 / 335
HAS-PART	14	0 / 1	21	
PART-OF	50	2 / 2	1	
HAS-SUBJECT	2054	43 / 69	2365	6 / 11
SUBECT-OF	0		8877	76 / 93
HYPERNYM	7947	182 / 221	0	
INSTRUMENT-FOR	14	1 / 1	3	
LOCATED-AT	71	2 / 4	1157	2 / 4
LOCATION-OF	287	4 / 8	17	
MADE-INTO	3		16	
MADE-OF	35	1 / 1	45	
PURPOSE	309	8 / 8	182	1 / 1
TIME-OF	624	12 / 16	211	0 / 2
Total (Precision %)	18262	403 / 500 (**80.6%**)	41680	413 / 500 (**82.6%**)

2.3.2 Experiment 2: Interpretation of Compound Noun

Using subcategorization information ([7]) and semantic information automatically extracted, we interpreted compound nouns and evaluated the performance for 450 compound nouns randomly selected from compound nouns list constructed by Nam ([13]). Evaluator consists of five persons. We regard interpretation of compound noun as correct answer when three or more persons evaluate it as correct.

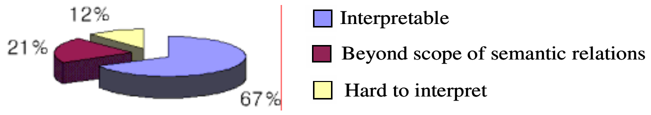


Fig. 4. The ratio of compound nouns according to interpretability.

Fig. 4 shows composition ratio of interpretability for compound nouns used in our experiment. 303 compound nouns are interpretable, 54 are hard to interpret, and 93 are beyond scope of semantic relations. The relations of beyond scope include <equal> (‘son daughter’) and <color> (‘black shoes’). It needs more semantic relations to interpret a compound noun.

We evaluated interpretable 303 compound nouns by precision and recall measures:

$$\text{Precision} = C / (C + W) * 100 \qquad \textbf{(1)}$$

$$\text{Recall} = (C + W) / (C + W + F) * 100 \qquad \textbf{(2)}$$

where *C* means the number of correct answers, *W* is for incorrect answers, and *F* is for failure of interpretation.

Table 7 shows the effect of interpretation according to resources. The performance using semantic information extracted from an MRD, corpora and subcategorization improves +40.30% in precision and +12.73% in coverage over the performance in previous researches using semantic relations extracted from only MRDs.

Table 7. The performance for interpretation of compound nouns according to resources used.

Resources	Correct	Incorrect	Failure	Precision	Recall
MRD	35	48	220	42.2 %	27.4%
MRD+Corpus	54	46	203	54.0%	33.0%
MRD+Subcategorization	55	41	207	57.3%	31.7%
MRD+Corpus+Subcategorization	73	38	192	65.8%	36.6%

Table 8. Error example for exceptional nouns applied to a regular pattern.

Head word	Definition in an MRD	Semantic information	
부대(sack)	종이(paper), 피륙(textile) 으로 만든 자루(bag)	부대 –<material>→종이	O
		부대 –<material>→피륙	O
바퀴(wheel)	둥근(round)모양(shape 으로 만든 물건 (thing)	바퀴–<material>→모양	X
조화(artificial flower)	인공(artificiality) 으로 만든 꽃(flower)	인공–<material>→ 조화	X

2.3.3 Error Analysis

Error types in extracting semantic information

(1) Error in part-of-speech tagging: Regular patterns were not applied or wrong semantic information was extracted by POS tagging error.

(2) Error in syntactic analysis for parallel structure: Since regular patterns are defined to extract simple parallel structure, we failed to acquire semantic information with complex parallel structure.

(3) Error from exceptional nouns for a regular pattern: The combination of special nouns and a regular pattern cause error. For example, when a regular pattern, '으로 만든 [eu-lo man-deun] (made {by,of,from)' for <material> relation, is applied to nouns such as '인공 [in-gong](artificiality)' and '모양[mo-yang](shape), the semantic information extracted is wrong.

Error types in interpreting compound nouns

(1) Lack of semantic information and error in semantic network: The error of semantic network give direct effect to the error of interpretation. By elaborating regular patterns and regulating semantic network, the error will lessen. And by expanding regular patterns and by using thesaurus and inference, we deal with lack problem of semantic information.

(2) Error in interpretation rule: In semantic network, we could not interpret by inference although two nouns are connected in case they are connected by <subject> or <object>. Interpretation rule using semantic features make errors. We interpreted <possessive> relation based on semantic features. There are many errors since it could not deal lexical level interpretation. It needs delicate interpretation rule.

To interpret a compound noun, two nouns should be connected on semantic network in any links. By constructing semantic network with semantic relations extracted from corpora as well as MRDs, semantic network could have much more information such as terms and links. Therefore, the system could obtain more correct interpretation. By interpreting compound nouns according to the characteristics of a head noun, i.e. attributive or predicative using hybrid knowledge, the coverage of interpretation could be extended.

3 A Workbench for Integration of Knowledge Acquisition and Interpretation

The interpretation method of compound nouns requires very large lexical and semantic information between nouns with high quality. But extracting large and faultless semantic information is very difficult job. By experiments and analysis, we observed that interpretation systems require more delicate regular patterns, more various semantic relations, and interpretation rules. We propose a workbench system which integrates knowledge acquisition and compound noun interpretation procedure for Korean semantic analysis with user's feedback. The system consists of semantic relation pattern extractor, knowledge indexer, compound noun extractor, and compound noun interpreter. The system keeps user's error correction log which make possible to specialize exceptional usages.

3.1 Semantic Relation Pattern Extractor

Semantic relation pattern extractor defines regular patterns and their semantic relations by searching lexical patterns for POS tagged corpus (Fig. 5).

1. Input POS tagged corpus.
2. Define lexical patterns to examine. The first and last part consists of nouns. For example, a lexical pattern to search is this:

   N1/n* */j* #4 N2/N* **(3)**

   where N1 means a modifier noun, N2 means a head noun, '/' symbol is for dividing lexical and POS tag, '*' symbol means any matching is possible, '#<num-

ber>' represents the maximum distance of N1 and N2, 'n*' is all nouns, and 'j*' is for all postposition at POS tag.

3. Results of all possible compound nouns (CN) list for the regular patterns. If a user define the regular pattern in (3), the results will include this:

버스/ncn+충돌/ncpa+에/jca 의하/pvg+ㄴ/etm 사고/ncn **(4)**

4. The system provides all possible regular patterns between two nouns. For (4), the system gives two results: '에/jca' and '에/jca 의하/pvg+ㄴ/etm'.
5. A user decides the regular pattern and the semantic relation for the compound noun, which predefined semantic relations are shown such as <by-means-of> and <purpose>. Also user can define new semantic relations.

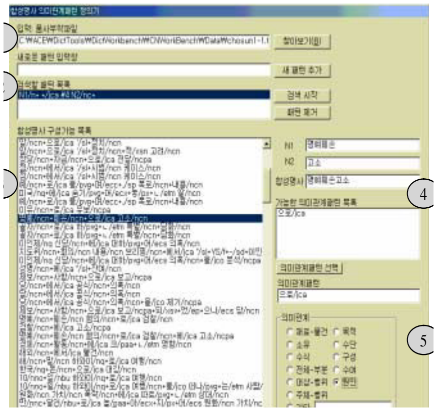


Fig. 5. Semantic relation pattern extractor.

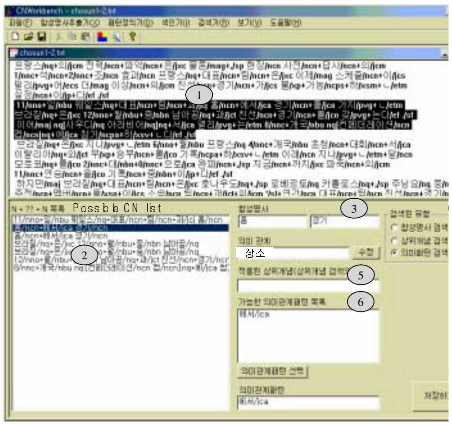


Fig. 6. Compound noun extractor.

3.2 Compound Nouns and Their Interpretations Extractor

Compound noun extractor extracts compound nouns and their possible interpretations based on semantic relations, inference, and on semantic network (Fig. 6).

1. A user selects the block to interpret extract compound nouns for POS tagged corpus and press the extraction button.
2. Then, system provides all possible compound nouns list and their semantic relations.
3. For selected passage including a compound noun, the system provides the compound nouns and the interpretation by knowledge of current system.
4. The result can be one by searching indexed knowledge for semantic relation patterns and existing semantic network.
5. Here a user can select the level of hypernym. It makes possible to generalize.
6. Regular patterns are presented. For the wrong interpretation of a compound noun, the system reserves error log.

3.3 Knowledge Indexer

Knowledge indexer constructs internal indexing structure of nouns and their semantic relations on semantic network. To interpretation a compound noun, the system use indexing information. Through indexing of incoming semantic information and compound noun extracted, we can extend knowledge database.

Indexing for semantic relation patterns: For each regular pattern (key part of indexing), the order of nouns and the possible semantic relations are indexed.

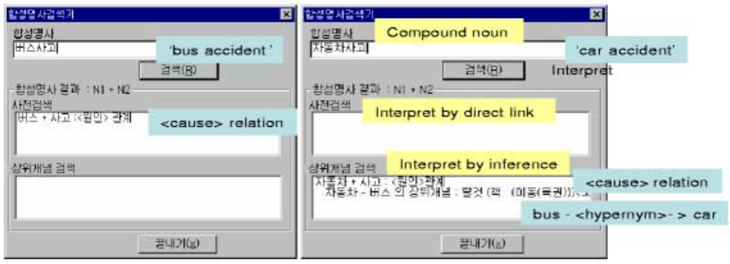
Key field (regular pattern)	Data field (<order of nouns, relation>)	
에/jca 의하/pvg+ㄴ/etm (by means of)	N1, N2	<cause>, <by-means-of>
으로/jca 만든pvg+ㄴ/etm (made from)	N1, N2	<material>

Indexing for compound nouns and their semantic relations: For each noun for a compound noun, the position and semantic relation are indexed. Key part of index is a noun, and data part is position, i.e. rear or front of noun and their semantic relation. For a compound noun, '가죽가방 (leather bag)', indexing structure is as follows:

Key field (noun)	Data field (<position, noun, relation>)		
가죽 (leather)	front	가방 (bag)	<material>
가방 (bag)	rear	가죽 (leather)	<material>

3.4 Compound Noun Interpreter

Compound noun interpreter interprets a compound noun by interpretation rules and inference. Figure 7 shows the interpretation of '버스사고 (bus accident)' by direct link and '자동차사고 (car accident)' by <hypernym> relation between '버스 (bus)' and '자동차 (car)' as <cause> relation.



버스 + 사고 :<원인> 관계

자동차 + 사고 : <원인>관계
자동차 - 버스 의 상위개념 : 탈것 (탁 (이동(목적))

Fig. 7. Examples for compound noun interpretation of 'bus accident' and 'car accident'

4 Conclusion

This paper has described the workbench system for constructing a dictionary to interpret Korean compound nouns, which integrates the acquisition of semantic information and the interpretation of compound nouns. To acquire knowledge for interpretation of compound nouns, we extracted semantic relations and semantic features from MRDs and corpora. The precision of semantic information is 80.6% and 82.6% from MRD and corpora, respectively. To interpret compound nouns, we used hybrid knowledge such as semantic relations, semantic features, and subcategorization information depending on the characteristics of a head noun. Experimental results show that our method improved the accuracy rate by 40.30% and the coverage rate by 12.73%, better than the rates obtained in previous work using semantic relations extracted from MRDs. By constructing a semantic network with semantic relations extracted from corpora as well as MRDs, the semantic network can have much more information such as terms and links. Therefore, the system can give more correct interpretation. By interpreting compound nouns based on hybrid knowledge depending on the characteristics of a head noun, the coverage of interpretation could be extended.

As compound nouns are highly productive and their interpretation requires complex knowledge, we built a workbench for compound noun interpretation in which necessary knowledge such as semantic patterns and interpretation instances of compound nouns can be extended, rather than assuming pre-defined lexical knowledge. The workbench integrated knowledge acquisition and interpretation.

References

1. Dolan, B., Vanderwende, L., Richardson, S.: Automatically deriving structured knowledge base from on-line dictionaries. In Pacific Association for Computational Linguistics. (1993).
2. Downing, P.: On the creation and use of English compound nouns. Language 53. (1977) pp.810-842.
3. Finin, T.W.: The semantic interpretation of compound nominals. U. of Illinois at Urbana-Champaign. University Microfilms International. (1980).
4. Jespersen, O.: A modern English grammar on historical principles., VI. George Allen & Unwin Ltd., London, 1909-49; reprinted 1954. (1954).
5. Kang, I.H.: Korean part-of-speech tagging based on maximum entropy model. MS thesis. KAIST (1999) (in Korean).
6. Kang, Y.H.: Noun semantic classification for Korean-to-English machine translation. MS thesis. Universty of Kyungpook (1989) (in Korean).
7. Kim, I.T.: Research on subcategorization of verb for Korean sentence analysis. Report of SERI. (1997).
8. Kim, S.N., Won, S.Y., Kwon, H.C. et al.: Korean compound noun analysis using semantic information. In KISS Fall. (1998) (in Korean).
9. Kurohashi, Sadao, Sakai, Yasuyuki.: Semantic analysis of Japanese noun phrases : A New Approach to Dictionary-Based Understanding. In ACL99. (1999).

10. Lehnert, W.: The analysis of nominal compounds. In U. Eco, M. Santambrogio, and P. Violi Ed., Meaning and Mental Representations, VS 44/45, Bompiani, Milan. (1988).
11. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 3. (1990).
12. Montemagni, S., Vanderwende, L.: Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries. In COLING92. (1992).
13. Nam, Ji-Sun and Choi, Key-Sun.: Korean electronic dictionary. Technical report. CAIR-TR-97-72. (1997) (in Korean).
14. Rosario, B., Hearst, M.: Classifying the semantic relation in noun compounds via a domain-specific lexical hierarchy. In EMNLP-2001. (2001).
15. Richardson, S., Dolan, W., Vanderwende, L.: MindNet: acquiring and structuring semantic information from text. In COLING-98. (1998).
16. Sparck Jones, K.: So what about parsing compound nouns? In K. Sparck Jones and Y. Wilks Ed., Automatic Natural Language Parsing, Ellis Horwood, Chichester, England. (1983) pp. 164-168.
17. Vanderwende, L.: The analysis of noun sequences using semantic information extracted from on-line dictionaries. Ph.D. thesis, Georgetown University. (1995).

Building Parallel Corpora by Automatic Title Alignment

Christopher C. Yang and Kar Wing Li

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
yang@se.cuhk.edu.hk

Abstract. Cross-lingual semantic interoperability has drawn significant research attention recently, as the number of digital libraries in non-English languages has grown exponentially. Cross-lingual information retrieval (CLIR) across different European languages, such as English, Spanish and French, has been widely explored, but CLIR across European and Oriental languages is still at the initial stages. To cross the language boundary, a corpus-based approach shows promise of overcoming the limitations of knowledge-based and controlled vocabulary approaches. However, collecting parallel corpora between European and Oriental languages is not an easy task. Length-based and text-based approaches are two major approaches to align parallel documents. In this paper, we investigate several techniques using these approaches, and compare their performance in aligning English and Chinese titles of parallel documents available on the Web.

1 Introduction

Many parallel text alignment techniques have been developed in the past. These techniques attempt to map various textual units to their translations and have proven useful for a wide range of applications and tools, e.g. cross-lingual information retrieval [12], bilingual lexicography, automatic translation verification and the automatic acquisition of knowledge about translation[16]. Translation alignment techniques have been used in automatic corpus construction to align two documents [8].

Given a text and its translation, an alignment is a segmentation of two texts such that the n^{th} segment of one text is the translation of the n^{th} segment of the other [17]. Empty segments are allowed, as they can correspond either to the translator's omissions or to additions. In other words, alignment is the process of finding relations between a pair of parallel documents. An alignment may also constitute the basis of deeper automatic analysis of translations. For example, it could be used to flag possible omissions in a translation, or to signal common translation mistakes, such as terminological inconsistencies.

There are three major structures of parallel documents on the World Wide Web : *parent page structure*, *sibling page structure*, and *monolingual sub-tree structure*. Resnik [14] noticed that if a web page has been written in many languages, the parent page of the Web page may contain links to different versions of the web page. For example, in a web page, there are two anchor texts A_1 and A_2. A_1 is linked to the

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 328-339, 2002.

Language 1 version and A_2 is linked to the Language 2 version. The sibling page structure refers to the case in which the page in one language contains a link directly to translated pages in the other language. The third structure contains a completely separate monolingual sub-tree for each language, with only the single top-level Web page pointing off to the root page of single-language versions of the site [14].

Parallel corpus can be generated using *overt translation* or *covert translation*. Overt translation [15] has a directional relationship between the pair of texts, which means texts in language A (source text) is translated into texts in language B (translated text)[24]. Covert translation [15] is non-directional. Parallel corpora generated by overt translation usually use the parent page structure and sibling page structure. To collect parallel corpora based on parent page and sibling structures, link analysis is sufficient. However, parallel corpora generated by covert translation use the monolingual sub-tree structure. Each sub-tree is generated independently. (The press releases of the Hong Kong SAR government is a typical example.) To collect parallel corpora based on the monolingual sub-tree structure, techniques that are more advanced than link analysis is required, since direct links or links through a parent page are not available between the pair of parallel documents. Length-based and text-based approaches are two typical approaches to align such parallel corpora.

Given a set of parallel texts, the alignment that maximizes the probability over all possible alignments is retrieved[6].

$$\arg\max_{A} \Pr(A \mid T_1, T_2) \approx \arg\max_{A} \prod_{(L_1 \Leftrightarrow L_2) \in A} \Pr(L_1 \Leftrightarrow L_2 \mid L_1, L_2) \qquad (1)$$

where **A** is an alignment, and T_1 and T_2 are the English and Chinese texts, respectively
L_1 and L_2 are the passages in two languages
$L_1 \Leftrightarrow L_2$ is an individual aligned pair
An alignment **A** is a set consisting of $L_1 \Leftrightarrow L_2$ pairs

1.1 Sentence Alignment

There are two major approaches to document alignment, namely length-based and text-based alignment. The length-based approaches make use of the total number of characters or words in a sentence, and the text-based approaches use linguistic information in the sentence alignment[4].

Length-based algorithms assume that the sentences, which are mutual translations in the parallel corpus, are similar in length[6]. The sentence alignment algorithm developed by Brown et al.[1] is based on the number of words in each sentence. Gale and Church [6] developed a similar algorithm except that alignment is based on the number of characters in sentences. These approaches based exclusively on sentence lengths work quite well with a clean input, such as the Canadian Hansards, and have been widely used by other researchers (e.g. Resnik [13], Chen et al.[2], and Wu [21]). However, for cases where sentence boundaries are not clearly marked, such as OCR input[3], or where the languages originate from different families of languages, such as Asian-European language pairs [5][21], these algorithms do not perform well.

Text-based algorithms use lexical information across the language boundary to align sentences. Warwick-Armstrong and Russell [20] used a bilingual dictionary to select word pairs in sentences from a parallel corpus and then aligned the sentences based on the word correspondence information. Another type of lexical information, which is helpful in the alignment of European language pairs, is called cognate [17]. Cognates are pairs of tokens of different languages, which share obvious phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations. The pair of *generation/génération* constitutes a typical example for English and French. Simard et al. [17] illustrated that cognates provide a reliable source of linguistic knowledge.

1.2 Title Alignment

According to the Collins Cobuild dictionary, if you align something, you "place it in a certain position in relation to something else, usually along a particular line or parallel to it." A textual alignment usually signifies a representation of two texts, which are mutual translations in such a way that the reader can easily see how certain segments in the two languages correspond [9].

Titles of two texts can be treated as representations of the two texts. Referring to He [7], the titles present "micro-summaries of texts" that contain "the most important focal information in the whole representation" and "the most concise statement of the content of a document". In other words, titles function as the condensed summaries of the information and content of the texts. The HKSAR government press releases have the titles of the Chinese articles and English articles listed on two separate Web pages. Aligning the parallel press releases requires alignment of the Chinese titles with the English titles.

In order to align titles effectively, the characteristics of title translation pattern should be first analyzed carefully. Similar to Gale and Church [6], three characteristics of translation pattern have been identified at the sentence (title) level:

1) One title in Language A translates into one title in Language B;
2) If title is not translated at all, document is available in one language only, e.g. English only or Chinese only articles;
3) Title in Language A has no equivalent title translation in Language B.

In the second and third cases, there is either no alignment or it is impossible to find an alignment based on the titles. For example, the English title, "From bak choy to baguette ...," appears in the English version of the Hong Kong government's press release web site. However, the corresponding Chinese title in the Chinese version of the web site is "盧偉聰情繫國際刑警心懸香港" (Lo Wai-chung loves Interpol and Hong Kong). Although the corresponding English and Chinese documents possess the relationship of covert translation, the Chinese title is not equivalent to the English title.

The characteristics of translation pattern at word level have been identified as follows:

1) One word in Language A is translated into one word in Language B;
2) Many words in Language A are translated into one word in Language B;
3) Some words are not translated;
4) A word in Language A is not always translated in the same way in Language B;

5) A word in Language A is translated into morphological or syntactic phenomena rather than a word in Language B.

In this paper, we investigate seven alignment techniques (three length-based approaches and four text-based approaches), and compared their performance in aligning the Chinese and English titles of Web documents. The seven techniques are:

1) Gale and Church's length-based approach
2) Wu's length-based approach with lexical cues
3) Sun et al.'s length-based approach with lexicon checks
4) Utsuro et al.'s dictionary-based approach
5) Melamed's geometric sentence alignment
6) Ma and Liberman's Bilingual Internet Text Search
7) Our proposed text-based approach using longest common subsequence

2 Length-Based Approaches

Length-based alignment methods[6] are developed based on the following approximation to Equation (1):

$$\Pr(L_1 \Leftrightarrow L_2 \mid L_1, L_2) \approx \Pr(L_1 \Leftrightarrow L_2 \mid \ell_1, \ell_2) \qquad (2)$$

where ℓ_1 = length(L_1) and ℓ_2 = length(L_2) measured in the number of characters

The length-based alignment model assumes that each character in L_1 is responsible for generating some number of characters in L_2. This leads to a further approximation that encapsulates the dependence to a single parameter δ. δ is function of ℓ_1 and ℓ_2.

$$\Pr(L_1 \Leftrightarrow L_2 \mid L_1, L_2) \approx \Pr(L_1 \Leftrightarrow L_2 \mid \delta(\ell_1, \ell_2)) \qquad (3)$$

Based on the Bayesian Rule,

$$\Pr(L_1 \Leftrightarrow L_2 \mid \delta) = \frac{\Pr(\delta \mid L_1 \Leftrightarrow L_2)\Pr(L_1 \Leftrightarrow L_2)}{\Pr(\delta)} \qquad (4)$$

Although it has been suggested that length-based methods are language-independent[6], they may in fact rely to some extent on length correlations arising from the historical relationships of the languages being aligned. If translated sentences share cognates, then the character lengths of those cognates are correlated. Grammatical similarities between related languages may also produce correlations in sentence lengths. However, Chinese and English have no history of common development.

An experiment has been conducted to test the correlation between the English and Chinese titles. Since Chinese texts do not contain obvious word boundaries but consists of a linear sequence of non-spaced or equally spaced ideographic characters [23], Wu's byte count approach [21] is used to count each Chinese character as a length of 2 and each English or punctuation character as a length of 1. Figure 1 shows the plot of the length of the Chinese titles against the English titles. The mean number of Chinese characters generated by each English character is c=E(ℓ_2/ℓ_1)=0.7316, with a standard deviation σ =0.19767. The correlation is 0.7033.

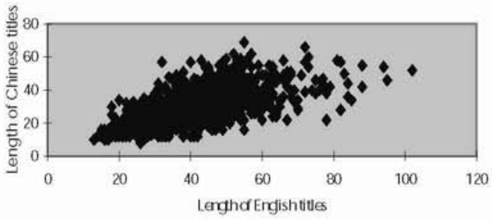


Fig. 1. A plot of the length of the Chinese titles against the length of the English titles for 150 aligned title pairs retrieved from the HKSAR press releases

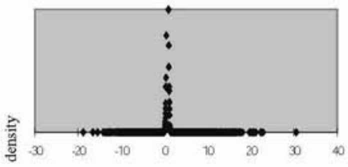


Fig. 2. Empirical density of δ for 150 aligned title pairs.

δ

Wu[21] assumed that ℓ_2 - ℓ_1c is normally distributed and it can be transformed into a new Gaussian variable of standard form (i.e. with the mean 0 and variance 1) by the appropriate normalization:

$$\delta(\ell_1, \ell_2) = \frac{\ell_2 - \ell_1 c}{\sqrt{\ell_1 \sigma^2}} \qquad (5)$$

Figure 2 plots the distribution of δ for 150 aligned Chinese and English titles.

According to Gale and Church [6], the prior of 6 classes of alignment are used to estimate $\Pr(L_1 \Leftrightarrow L_2)$. The six classes includes a sentence in one language matching exactly one sentence in the other language(1-1) and several additional possibilities (1-0, 0-1, 2-1, 1-2, 2-2). Table 1 shows the values of $\Pr(L_1 \Leftrightarrow L_2)$ for each of the six classes. For title alignment, only three classes, 1-1, 1-0 and 0-1, are utilized due to the characteristics of translation pattern at title level mentioned Section 1.2.

The dynamic programming algorithm is then applied to determine the minimum distance D(i,j) between the sentences s_1,…si and their translations t_1,…,t_j, under the maximum likelihood alignment. D(i,j) is computed by minimizing over the six classes.

$$D(i, j) = \min \begin{cases} D(i, j-1) + d(0, t_j; 0, 0) \\ D(i-1, j) + d(s_i, 0; 0, 0) \\ D(i-1, j-1) + d(s_i, t_j; 0, 0) \\ D(i-1, j-2) + d(s_i, t_j; 0, t_{j-1}) \\ D(i-2, j-1) + d(s_i, t_j; s_{i-1}, 0) \\ D(i-2, j-2) + d(s_i, t_j; s_{i-1}, t_{j-1}) \end{cases} \qquad (6)$$

Table 1. $\Pr(L_1 \Leftrightarrow L_2)$

Class	$\Pr(L_1 \Leftrightarrow L_2)$ used in [6]	$\Pr(L_1 \Leftrightarrow L_2)$ for title alignment
1-1	0.89	0.878
1-0 (or 0-1)	0.00099	0.122
2-1 (or1-2)	0.089	
2-2	0.011	

2.1 Wu's Length-Based Approach with Lexical Cues

To improve the purely length-based alignment, Wu [21] incorporated lexical criteria without giving up the statistical approach. Equation (3) is modified as follows:

$$\Pr(L_1 \Leftrightarrow L_2 \mid L_1, L_2) \approx \Pr(L_1 \Leftrightarrow L_2 \mid \ell_1, \ell_2, \nu_1, \omega_2, \ldots, \nu_n, \omega_n) \quad (7)$$

where v_i =no. of occurrences of English cue_i in L_1
w_i =no. of occurrences of Chinese cue_i in L_2

Consequently, Equation (4) is modified as follows:

$$\Pr(L_1 \Leftrightarrow L_2 \mid L_1, L_2) \approx \Pr(L_1 \Leftrightarrow L_2 \mid \delta_0(\ell_1, \ell_2), \delta_1(\nu_1, \omega_2), \ldots, \delta_n(\nu_n, \omega_n)) \quad (8)$$

2.2 Sun et al.'s Length-Based Approach with Lexicon Check

Sun et al.[18] utilized an English/Chinese lexicon to check the result of the alignments obtained from the length-based approach. A score S_A is computed for every aligned sentence pair. Each aligned sentence pair whose score is above a threshold, *t1*, is judged as correctly aligned. After removing the correct alignments, the rest are aligned by the length-based approach again. The second result is checked by lexicon again and the alignments whose score is above a threshold, *t2*, are considered as correct alignments.

$$S_A = \frac{No_{correct} \times 2}{No_{English} + No_{Chinese}} \quad (9)$$

where $No_{correct}$ corresponds to the number correct alignments of English and Chinese words identified by the lexicon
$No_{English}$ corresponds to the number of words in English sentence
$No_{Chinese}$ corresponds to the number of words in Chinese sentence

3 Text-Based Approaches

3.1 Utsuro et al.'s Dictionary-Based Approach

Utsuro et al.[19] proposed a sentence alignment approach based on word pairs available in the bilingual dictionary and the statistical information of word pairs. A dictionary was first used to align parallel sentences. Word pairs that are not available in the dictionary will then be evaluated based on their frequencies.

n sentences in language A and m sentences in language B are grouped into m×n pairs, P. (P = p1,p2,…,pk, where pk is a sentence pair). Words are first extracted from each sentence and their correspondences are identified using the dictionary.

Based on the word pairs that are identified, the score h(p) of the sentence pair p_k is computed as follows:

$$h(p) = \frac{n_{st}(p)}{n_s(a,x) + n_t(b,y)} \qquad (10)$$

where $n_{st}(p)$ the number of word pairs in the sentence pairs p
$n_s(a,x)$ is the number of words in the sequences of sentences $S_{a-x+1},\ldots,S_a$ in language S
$n_t(b,y)$ is the number of words in the sequences of sentences $T_{b-y+1},\ldots,T_b$ in language T

The score function follows the recursion equation below:

$$H(Pi)=H(Pi-1)+h(pi) \qquad (11)$$

where Pi is the sequence of sentence pairs from the beginning of the bilingual text to the pair pi .

The maximum score of H(Pi) will be the optimal solution to the alignment problem.

3.2 Melamed's Geometric Sentence Alignment (GSA)

Melamed [11] extended the Smooth Injective Map Recognizer (SIMR) to develop an algorithm called Geometric Sentence Alignment (GSA) for sentence alignment. The Smooth Injective Map Recognizer (SIMR) is based on a greedy algorithm for mapping bitext correspondence. A bitext comprises two versions of a text, such as a text in two different languages. Translators create a bitext each time they translate a text. Each bitext defines a rectangular bitext space. The lower left corner of the rectangle is the origin of the bitext space and it represents the beginning of two texts. The upper right corner is the terminus and it represents the end of two texts. The line between the origin and the terminus is the main diagonal. The width and height of the rectangle are the lengths of the two component texts, in characters.

Each bitext space contains a number of true points of correspondence (TPCs), other than the origin and the terminus. For example, if a token at position p on the x-axis and a token at position q on the y-axis are translations of each other, then the coordinate (p,q) in the bitext space is a TPC. Since distances in the bitext space are measured in characters, the position of a token is defined as the mean position of its characters. TPCs exist at the corresponding boundaries of text units such as sentences. Groups of TPCs with a roughly linear arrangement in the bitext space are called chains. For each bitext, the true bitext map (TPM) is shortest bitext map that runs through all the TPCs. SIMR considers only chains that are roughly parallel to the main diagonal. Since Chinese and English languages do not share an alphabet, the Chinese/English matching predicate deemed two tokens to match if they constituted an entry in the translation lexicon [10].

3.3 Ma and Liberman's Bilingual Internet Text Search (BITS)

Ma and Liberman [8] have developed a system called Bilingual Internet Text Search (BITS). To determine if two texts are mutual translations of each other, corresponding regions of one text and its translation will contain word token pairs that are mutual translations.

Given text A in language L1 and text B in language L2, text A and text B are tokenized. The similarity between A and B is computed as follows:

$$\text{sim}(A,B) = \text{Number of translation token pairs} / \text{Number of tokens in text A} \qquad (12)$$

If text B is most similar to text A, and sim(A,B) is greater than a threshold, t, then text A and text B are treated as a translation pairs. The following is their algorithm:

```
For each text A in language L1
    Tokenize A
    Max_sim=0
    For each text B in language L2
          Tokenize B
          s=sim(A,B)
          If s>max_sim Then
                max_sim=s
                most_sim=B
          Endif
    Endfor
    If max_sim>t Then
          Output(A,B)
    Endif
Endfor
```

3.4 Our Proposed Text-Based Approach

In our proposed text-based approach, the longest common subsequence is utilized to optimize the alignment of English and Chinese titles. The longest common subsequence (LCS) is commonly exploited to maximize the number of matches between characters of two sequences. Our alignment algorithm has three major steps: 1) alignment at word level and character level, 2) reducing redundancy, 3) score function.

Alignment at Word Level and Character Level. An English title, *E*, is formed by a sequence of English simple words, i.e., $E = e_1\, e_2\, e_3 \ldots\, e_i \ldots$, where e_i is the ith English word in *E*. A Chinese title, *C*, is formed by a sequence of Chinese characters, i.e., $C = char_1\, char_2\, char_3 \ldots\, char_q \ldots$, where $char_q$ is a Chinese character in C. An English word in E, e_i, can be translated to a set of possible Chinese translations, Translated(e_i), by dictionary lookup. $Translated(e_i) = \{ T_{e_i}^1, T_{e_i}^2, T_{e_i}^3, \ldots, T_{e_i}^j, \ldots \}$ where $T_{e_i}^j$ is the jth Chinese translation of e_i. Each Chinese translation is formed by a sequence of Chinese characters. The set of the longest-common-subsequence(*LCS*) of a Chinese translation $T_{e_i}^j$ and C is $LCS(T_{e_i}^j, C)$. $MatchList(e_i)$ is a set that holds all the unique longest common subsequences of $T_{e_i}^j$ and C for all Chinese translations of e_i.

$$MatchList(e_i) = \bigcup_j LCS(T_{e_i}^j, C) \qquad (13)$$

If there is no common subsequence of $T_{e_i}^j$ and C, $MatchList(e_i) = \varnothing$ and no reliable translation of e_i can be found in C. If there is at least one common subsequence of $T_{e_i}^j$ and C, we determine the most reliable translation based on the adjacency and length of Chinese translations found in C. Based on the hypothesis that if the characters of the Chinese translation of an English word appears adjacently in a Chinese sentence, such Chinese translation is more reliable than other translations that their characters do not appear adjacently in the Chinese sentence. For example, the English word “propose” can be translated as “建議” in Chinese. The translation “建議” can be aligning with “就建築條例的動議辯論” (on the "Construction Bill” motion debate) using *LCS*, which is not correct in this case. $Contiguous(e_i)$ is used to determine the most reliable translation based on adjacency.

$$Contiguous(e_i) = \{x \mid x \in MatchList(e_i) \textit{ and all the characters of } x \textit{ appear adjacently in } C\} \qquad (14)$$

The second criteria of the most reliable Chinese translations, is the length of the translations. $Reliable(e_i)$ is used to identify the longest sequence in $Contiguous(e_i)$.

$$\text{Reliable}(e_i) = \begin{cases} \arg\max_{x \in Contiguous(e_i)} |x| & \text{if } \text{Contiguous}(e_i) \neq \varnothing \\ \arg\max_{x \in MatchList(e_i)} |x| & \text{Otherwise} \end{cases} \qquad (15)$$

Resolving Redundancy. Due to redundancy, the translations of an English word may be repeated completely or partially in Chinese. For example, given E = *red color* and *C =赤紅色*, *Translated(red) = {紅, 紅色, 紅色的, 赤}*and *Translated(color)={色, 顏色, 色彩}*. *MatchList(red) = {紅, 紅色, 赤}* and *MatchList(color) = {色}*. *Reliable(red) = 紅色* and *Reliable(color) = 色*. To deal with redundancy, $Dele(x,y)$ is an edit operation to remove the $LCS(x,y)$ from x. *WaitList* is a list to save all the sequences obtained by removing the overlapping of the elements of $MatchList(e_i)$ and $Reliable(e_i)$. $MatchList(e_i)$ is initialized to $\varnothing$ and $Reliable(e_i)$ is initialized to ε.

$$WaitList = DELE(WaitList, Reliable(e_i)) \cup DELE(MatchList(e_i) \backslash Reliable(e_i), Reliable(e_i)) \qquad (16)$$

where

$$DELE(X, y) = \bigcup_{i=1}^{n} \text{Dele}(x_i, y)$$

x_i is the i^{th} element of X

Remain is a sequence that is initialized as *C*, and *Reliable*(e_i) are removed from *Remain* starting from the e_1 until the last English word. *WaitList* will also be updated for each e_i. When all *Reliable*(e_i) are removed from *Remain*, the elements in *WaitList* will also be removed from *Remain* in order to remove the redundancy.

Score Function. Given *E* and *C*, the ratio of matching is determined by the portion of *C* that matches with the reliable translations of English words in *E*.

$$Matching\_Ratio(E,C) \;=\; \frac{|C| - |Remain|}{|C|} \tag{17}$$

Given an English title, the Chinese title that has the highest *Matching_Ratio* among all the Chinese titles is considered as the counterpart of the English title. However, it is possible that more than one Chinese title have the highest *Matching_Ratio*. In such case, we shall also consider the ratio of matching determined by the portion of English title that is able to identify a reliable translation in the Chinese title.

$$Matching\_Ratio^{*}(E,C) \;=\; \frac{\sum_i R(e_i)}{|E|} \tag{18}$$

where $R(e_i) = \begin{cases} 0 & \text{if Reliable}(e_i) = \varnothing \\ 1 & \text{otherwise} \end{cases}$

If more than one Chinese title have the highest *Matching_Ratio* for the English title, *E*, the Chinese title with the lowest value of $|Matching\_Ratio(E,C)$ - $Matching\_Ratio^{*}(E,C)|$ is considered as the counterpart of *E*.

4 Experiments

An experiment was conducted to measure the precision and recall of the aligned parallel Chinese/English documents from the HKSAR government's press releases between 1998 and 2001, using the length-based and text-based approaches described in Sections 2 and 3. There were 31,567 Chinese articles, 30,810 English articles, and 23,701 pairs of English/Chinese parallel articles in the HKSAR government's press release corpus. Results are shown on Table 2.

Experimental results show that the text-based approaches out-perform the length-based approaches, and that our proposed text-based approach using LCS has the best performance. Chinese text contains fewer characters; character length is a less discriminating feature, varying over a range of fewer possible discrete values than English. As a result, the length-based approach is not as reliable as the text-based approach in title alignment. Lexical knowledge can effectively improve both precision and recall in title alignment. Since our proposed approach has adopted the longest common sequence to consider those Chinese translations that do not appear as adjacent characters in the Chinese sentence and the problem of redundancy, it produced the best performance.

Table 2. Experimental results

	Precision	Recall
Gale and Church's Length-based Approach	0.10	0.06
Wu's Length-based Approach with Lexical Cues	0.62	0.61
Sun et al.'s Length-based Approach with Lexicon Checks	0.76	0.05
Utsuro et al.'s Dictionary-based Approach	0.91	0.82
Melamed's GSA (Text-based Approach)	0.73	0.65
Ma and Liberman's BITS (Text-based Approach)	0.93	0.86
Our proposed Text-based Approach using LCS	1.00	0.87

5 Conclusion

Cross-lingual information retrieval has drawn significant attention recently. Parallel corpora are important linguistic resources that provide a statistical translation model to cross the language boundary. However, constructing English/Chinese parallel corpora is not an easy task due to significant differences between the two languages. In this paper, we investigated seven English/Chinese sentence (or title) alignment techniques. Three of them are length-based approaches and four are text-based approaches. Experimental results show that the text-based approaches out-performed the length-based approaches. In particular, our proposed text-based approach using LCS produced the best performance with 100% precision and 87% recall.

Acknowledgements. This project was supported by the RGC Direct Research Grant, 2050268, and the RGC Earmarked Grant for Research, 4335/02E.

References

1. Brown, P., Lai, J., and Mercer, R.: "Aligning sentences in parallel corpora". In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA(1991).
2. Chen, A., Kishida, K., Jiang, H., Liang, Q., Gey, F. :"Automatic Construction of a Japanese-English Lexicon and its Application in Cross-Language Information Retrieval". In Proceedings of the Multilingual Information Discovery And Access workshop of the ACM SIGIR'99 Conference, August 14(1999).
3. Church, K. W. : "Char_align: A Program for Aligning Parallel Texts at the Character Level". In Proceedings of ACL-93, Columbus OH (1993).
4. Fung, P. and McKeown, K. : " A technical word- and term-translation aid using noisy parallel corpora across language groups". In Machine Translation 12: 53-87(1997).
5. Fung, P.: "A Pattern Matching Method for Finding Noun and Proper Noun Translations from noisy Parallel Corpora". In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Boston, MA,(1995).
6. Gale, W. A., and Church, K.W.: "Identifying word correspondences in parallel texts". In Proceedings of the Fourth DARPA Workshop on Speech and Natural Language, Asilomar, California (1991).
7. He, S. : "Translingual Alteration of Conceptual Information in Medical Translation: A Cross-Language Analysis between English and Chinese". In Journal of the American Society for Information Science, Vol. 51, No. 11, pp.1047-1060(2000).

8. Ma X. and Liberman M.: “BITS: A Method for Bilingual Text Search over the Web”. In Machine Translation Summit VII, September 13th, 1999, Kent Ridge Digital Labs, National University of Singapore. (1999).
9. Macklovitch, E., Hannan, Marie-Louise: "Line'Em Up: Advances In Alignment Technology And Their Impact on Translation Support Tools". In Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96), Montréal, Québec. (1996).
10. Melamed, I. D. and Marcus M. P. :Automatic Construction of Chinese-English Translation Lexicons, IRCS Technical Report #98-28. (1998).
11. Melamed, I. D. : A Geometric Approach to Mapping Bitext Correspondence, In Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96), Philadelphia, PA. (1996).
12. Oard, D. W. : "Alternative approaches for cross-language text retrieval". In Hull D, Oard D,(Eds.) ,1997 AAAI Symposium in Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence, March(1997).
13. Resnik P. : "Parallel Strands: A Preliminary Investigation into Mining the Web for Bilingual Text". In Farwell D., Gerber L., and Hovy E. (eds.), Machine Translation and the Information Soup: Third Conference of the Association for Machine Translation in the Americas (AMTA-98), Langhorne, PA, Lecture Notes in Artificial Intelligence 1529, Springer, October (1998).
14. Resnik P. : "Mining the Web for Bilingual Text". In 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), College Park, Maryland, June(1999).
15. Rose, Marilyn Gaddis.(ed) :"Translation Types and Conventions". In Translation Spectrum: Essays in Theory and Practice, Marilyn Gaddis Rose, Ed., State University of New York Press, pp.31-33 (1981).
16. Simard, M.:"Text-translation Alignment: Three Languages Are Better Than Two". In Proceedings of EMNLP/VLC-99. College Park, MD (1999).
17. Simard, M., Foster, G., Isabelle P. :"Using Cognates to Align Sentences in Bilingual Corpora". In Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92), Montreal, Canada (1992).
18. Sun, L., Du, L., Sun Y. and Jin, Y..: "Sentence Alignment of English-Chinese Complex Bilingual Corpora". In Proceeding of the 5th Natural Language Processing Pacific Rim Symposium, Beijing, China (1999).
19. Utsuro T., Ikeda H., Yamane M., Matsumoto Y., and Nagao M.: "Bilingual Text Matching using Bilingual Dictionary and Statistics". In Proceeding of 15th International Conference on Computational Linguistics, Kyoto (1994).
20. Warwick-Armstrong, S. and Russell, G.: "Bilingual Concordancing and Bilingual Lexicography", Euralex (1990).
21. Wu, D. :"Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria". In 32nd Annual Conference of the Association for Computational Linguistics, Las Cruces, New Mexico, (1994) pp80-87.
22. Wu, D. and Fung, P.: "Improving Chinese Tokenization with Linguistic Filters on Statistical Lexical Acquisition". In 4th Conference on Applied Natural Language Processing,, Stuttgart, Germany, (1994) pp180-181.
23. Wu, Z. and Tseng G.:"Chinese text segmentation for text retrieval: Achievements and problems". In Journal of The American Society for Information Science, 44(9):532--542. (1993).
24. Zanettin, F.: "Bilingual comparable corpora and the training of translators," Laviosa, Sara. (ed.) META, 43:4, Special Issue. The corpus-based approach: a new paradigm in translation studies: 616-630 (1998).

Offline Isolated Handwritten Thai OCR Using Island-Based Projection with N-Gram Models and Hidden Markov Models

Thanaruk Theeramunkong[1], Chainat Wongtapan[2], and Sukree Sinthupinyo[2]

[1] Information Technology Program
Sirindhorn International Institute of Technology, Thammasat University,
PO. BOX. 22 Thammasat Rangsit Post Office
Pathumthani, Thailand 12121, Tel. +66-2-501-3505(-20) ext. 2004
thanaruk@siit.tu.ac.th

[2] Computer Science
Science and Technology Faculty, Thammasat University,
Pathumthani, Thailand 12121
Tel. +66-2-564-4444 ext. 2157
{chainat,sukree@hotmail.com}

Abstract. Many traditional works on offline Thai handwritten character recognition use a set of local features including circles, concavity, endpoints and lines to recognize hand-printed characters. However, in natural handwriting, these local features are often missed due to fast writing, resulting in dramatically reduced recognition accuracy. Instead of using such local features, this paper presents a method to extract features from handwritten characters using so-called multi-directional island-based projection. Two statistical recognition approaches using interpolated n-gram model (n-gram) and hidden Markov model (HMM) are also proposed. The performance of our feature extraction and recognition methods is investigated using nearly 23,400 hand-printed and natural-written characters, collected from 25 subjects. The results showed that, in situations where local features are hard to detect, both *n*-gram and HMM approaches achieved up to 96-99 % accuracy for close tests and 84-90 % for open tests.

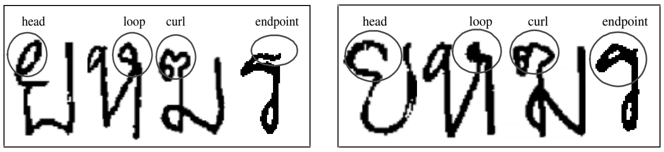
1 Introduction

Optical character recognition (OCR) contributes tremendously to improved the interaction between human and machine in many applications, including office automation, signature verification and many data entry applications in an organization [16]. In an OCR system, it is necessary to extract some features from a character and use them for recognition. There are two different possible types of features used in this task, i.e., global and local features. Some systems [2-4][10][11][20] used only global features, which are information extracted from the whole character image, to recognize the

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 340-351, 2002.

character. In [1][9][13][14][18], local features, information extracted from some parts of the character such as circles, concavity, endpoints and lines, are utilized for recognition. There are also some works [9][21] using a combination of local and global features.

In the early stage of Thai OCR, Kimpan [5] used template matching directly with 2-D image data of characters, i.e., global features, to recognize printed Thai characters. This technique is very simple but it needs highly intensive computation and is sensitive to noise and distorted characters. However, to achieve more performance and robustness, most previous Thai OCR methods extracted local features (or dominant features) of Thai characters, such as head and loop (or circle), concavity (or curl), endpoint, and line. Then a number of different techniques, e.g. structural technique [1][6] and neural networks [13][14][18] were applied. Although these local features are useful in improving recognition accuracy, in a real situation, they are often omitted, as shown in Fig. 1, resulting in lower recognition rate.



(a) (b)

Fig. 1. An example of Thai characters: (a) complete and (b) incomplete local features.

On the other hand, there were still a few works using global features in Thai character recognition [4][9][20]. Apirak [4] applied global features with HMM on online Thai OCR. Wongtapan et al. [20] proposed a feature extraction called island-based projection (IBP) together with interpolated *n*-gram model (*n*-gram) for recognizing offline Thai handwritten characters. In this work, the IBP represented a character image by transforming 2-D image data to 1-D data in vertical and horizontal directions. Although the method gained up to 88% accuracy, it was just a preliminary work with no comparison to other existing methods.

In this paper, we propose a method to extract features from handwritten characters using so-called *multi-directional island-based projection*. Two statistical recognition approaches using *interpolated n-gram model (n-gram)* and *Hidden Markov Model (HMM)* are also proposed and compared. The performance of the proposed feature extraction and recognition methods is investigated using nearly 23,400 hand-printed and natural-written characters, collected from 25 subjects. The rest of this paper is organized as follows. Section 2 presents our proposed feature extraction method. Section 3 illustrates the recognition models. The experimental results and analysis are given in section 4. Finally, conclusions and future work are described in section 5.

2 Multi-directional Island-Based Projection

If a statistical model is applied directly on 2-D character image grids, it is necessary to consider a large number of parameters and hence it requires large amount of training data. To solve this problem, there were some attempts [2][3][10][11][20][21] to transform a 2-D character image grid to 1-D sequential data before applying a statistical model. In this section, a new transformation method called multi-directional island-based projection is given. First, the character extracted from an image is normalized to a fixed size window. The normalized window can be scanned in various directions. The possible directions include vertical (V), horizontal (H), upper-left to lower-right diagonal (L), and upper-right to lower-left diagonal (R) directions (see Fig. 2 for horizontal and vertical directions and Fig. 3 for diagonal directions). Given a direction, a feature vector is constructed for each scanning line. Here, the number of feature vectors extracted for a direction equals to the number of scanning lines. In the example given in the figures, the window size is 36x36 pixels and the number of scanning lines is 36, i.e. one-pixel width per scanning. To encode a feature vector for each scanning line, our island-based projection is applied. Here, an island is defined as a group of *contiguous* active pixels. The method divides a scanning line into a number of even regions, e.g., six even regions (each of which holds six contiguous pixels) as shown in Fig. 2 and 3. The 1st element of a feature vector is the number of islands in the whole scanning line. The 2nd to the 7th elements are the numbers of islands in the six regions, respectively.

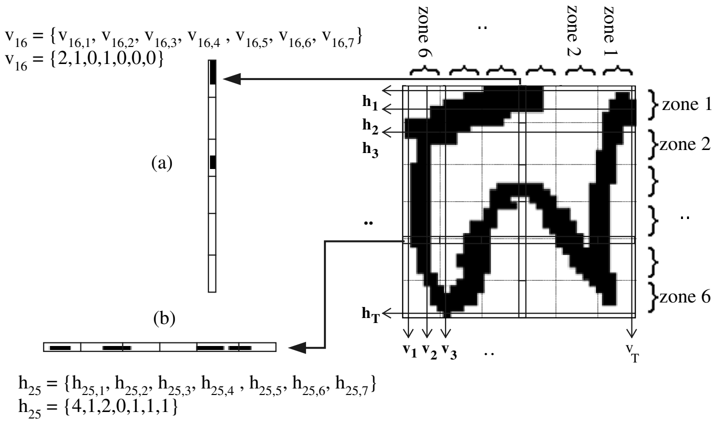

$v_{16} = \{v_{16,1}, v_{16,2}, v_{16,3}, v_{16,4}, v_{16,5}, v_{16,6}, v_{16,7}\}$
$v_{16} = \{2,1,0,1,0,0,0\}$
(a)
(b)
$h_{25} = \{h_{25,1}, h_{25,2}, h_{25,3}, h_{25,4}, h_{25,5}, h_{25,6}, h_{25,7}\}$
$h_{25} = \{4,1,2,0,1,1,1\}$
zone 6
..
zone 2
zone 1
zone 1
zone 2
..
zone 6
h_1
h_2
h_3
..
h_T
v_1 v_2 v_3 ..
v_T

Fig. 2. Extracting raw feature vectors of the vertical (V) and horizontal (H) directions.

Fig. 2 is an example of extracting raw feature vectors of a character "ฟ" in the V and H directions. It shows raw feature vectors on the 16th and 25th slices along the V and H directions, respectively. The vector *v16* is {2,1,0,1,0,0,0} while the vector *h25* is {4,1,2,0,1,1,1}. Finally, each vector is normalized by its size. For diagonal directions, scaling and rotation techniques are used (see Fig. 3). The vertical projections reflect the upper-right to lower-left diagonal (R) and the projections along the horizontal direction reflect the upper-left to lower-right diagonal (L) directions. Therefore, there are 36 feature vectors generated for each dimension. These feature vectors form a feature sequence vector. Due to 4-direction scanning, a character is represented by 4 sets of 36 feature vectors, i.e. 4 feature sequence vectors.

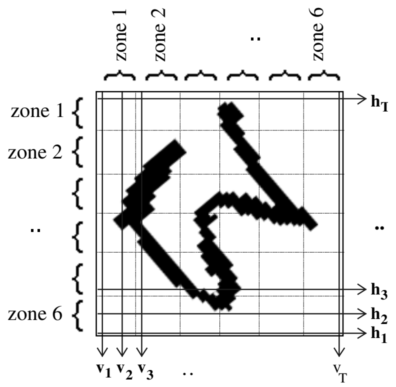


Fig. 3. Extracting raw feature vectors of the diagonal directions (R and L directions)

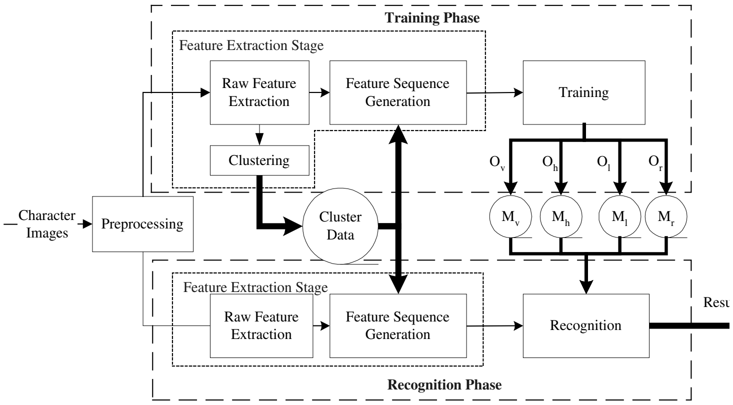


Fig. 4. The diagram of our recognition system

3 Recognition Models: *N*-Gram and HMM

3.1 The Recognition Scheme

We can divide our recognition system into four sub stages according to their tasks: (1) preprocessing, (2) feature extraction, (3) training, and (4) recognition (see Fig. 4). In the preprocessing stage, character images are modified by some operations to make the images suitable for the next stage. Second, the feature extraction process generates four feature sequence vectors that describe the writing pattern of a character image. Here, some tasks are utilized such as raw feature extraction, clustering, and feature sequence generation, in order. In the training process, feature sequence vectors are used to train a model of each character. The main idea of the recognition process is to find the character model with the highest probability given feature sequence vectors and then select it as the recognition result.

3.2 *N*-Gram & HMM

An *n*-gram, a statistical model, is widely used in speech processing and natural language processing. Our preliminary work, Wongtapan et al. [20] applied it to recognize handwritten Thai characters. In [20], it was shown that, even with simple and quick training, the model gains relatively high recognition rate. However, this work does not provide any comparison with other recognition models.

HMM is a doubly stochastic process with an underlying Markov process that is not directly observable (hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols [7] (see also [15]). HMM is a very powerful tool for modeling the problem of one-dimensional (1-D) data. However, it can be extended to deal with two-dimensional (2-D) data, such as an image; Pseudo 2-D HMM and 2-D HMM are such examples. They can directly model 2-D data but they require a lot of parameters and training samples, and also take much more time for learning than the 1-D HMM [11]. Besides this, the recognition rate of 1-D HMM and those of Pseudo 2-D HMM are not so different. Thus, multi-directional 1-D HMMs are used in this work.

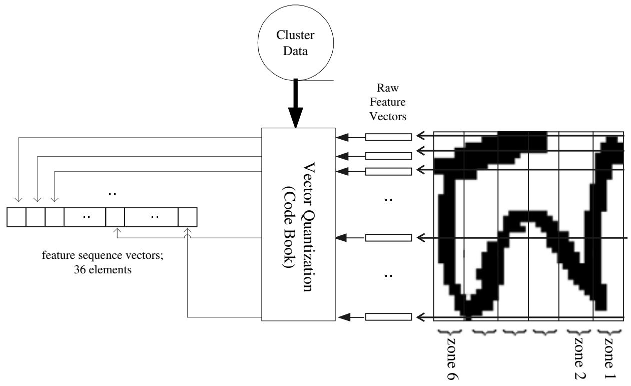

Cluster Data
Raw Feature Vectors
Vector Quantization (Code Book)
feature sequence vectors; 36 elements
zone 6
zone 2
zone 1

Fig. 5. An example of transforming the raw feature vectors based on the cluster data generates a feature sequence vector of the horizontal direction.

3.3 Training the Models

The character images are preprocessed by binarization (two-level thresholding), noise reduction, single-character extraction, size normalization and centering. After that, three tasks of the feature extraction stage are performed. First of all, raw features of each character in the training set are extracted by the MDIBP method according to V, H, L and R directions. Secondly, for each direction, the raw features of all characters in the training set are clustered by the *k*-mean algorithm, into *k* groups for reducing the variations of raw features, and results in accuracy improvement. Thus, we obtain *k* set of similar vectors, referred as *k* master clusters. Finally, the feature sequence generation transforms the raw features based on these master clusters, and then generates a feature sequence vector that describes a character in each direction. Each element of the feature sequence vector is an index of the vector in the master clusters, which the raw-feature vector is the closet, according to the Euclidean space (see Fig. 5). On the other hand, we can formulate a set of feature sequence vector O=$\{O^v,O^h,O^l,O^r\}$, where O^v,O^h,O^l and O^r are feature sequence vectors of V, H, L, and R directions, respectively. In the training stage, four n-gram's models of each character are trained by a counting function described in [20], while the parameters of four HMM's models are adjusted by the Baum-Welch re-estimation algorithm [15]. As a result, the models of each character are constructed. Since there are 4 directions, we will get *N* x 4 character models for *N* characters.

3.4 Recognition Process

A test character image is processed by the same operations as in the training phase. To generate feature sequence vectors O^v,O^h,O^l and O^r that describe the test character, the raw feature vectors of each direction are extracted and the feature sequence vectors are then generated. To recognize the character, the probabilities of four models of each character are independently estimated. These probabilities are then equally combined together before comparing with other models. The probability for each character model, i.e., the output of each HMM model, can be calculated by using Viterbi algorithm [15]. On the other hand, the n-gram models are calculated by the statistical equation shown in [20]. Finally, the character model with the highest probability is chosen as the recognition result as shown in equation (1).

$$i_{best} = \underset{i=1\ldots N}{\arg\max}\, P(O, M^i) \qquad \textbf{(1)}$$

Here, M^i is the model of the i^{th} character and *N* is the total number of distinct characters in the language. The character indexed by i_{best} will be the recognized character. In this work, we investigate the multi-directional IBP with two statistical models (*n*-gram and HMM) by comparing their recognition rates as shown in the next section.

4 Experimental Results & Analysis

4.1 Data Set & Model Parameters

In the experiments, 78 Thai characters were captured - 44 consonants, 24 vowels, and 10 Thai numbers. In the data set, the total number of hand-printed and natural-written characters is 23,400 characters, collected from 25 subjects. Here, each subject was requested to write 12 times for each character. Two types of experiments were examined: *Open-test* and *Close-test*. In the Open-test, we performed 3-fold-cross validation where 70 percent of the data set was used for training and 30 percent for testing. On the other hand, in the close-test, we used 100 percent for both training and testing. We also investigated the effect of the writing styles of individual subjects on recognition accuracy by evaluating two different environments, *writer dependent* and *writer independent*. In the writer dependent (WD) environment, each writer was separately evaluated. In contrast, the handwritten characters of two or more writers were performed together in the writer independent (WI) environment.

The N-gram's parameters: Coefficient constants of unigram, bigram and trigram are set as follows: α=0.05, β=0.95, and δ=0.05. These parameters are fixed for all dimensions. On the other hand, the coefficients for merging probabilities of 4 directions are set to be all 0.25 each. That is, the decision for each dimension has equal effect on the overall decision.

The HMM's parameters: A left-right model with the 4-states-jumped constraint is used. In this work, initial state transition values are specified as follows: A_{ij}=0.80 when i=j, A_{ij}=0.10 when j=i+1, A_{ij}=0.05 when j=i+2, A_{ij}=0.03 when j=i+3, and A_{ij}=0.02 when j=i+4. These values are adapted from Waleed and Nikola's work [12]. In contrast, the initial symbol probabilities are randomly assigned. The model always starts at the first state and the probabilities of the 4 directions are merged with an equal weight (Also see Section 4.3.1).

The writer-dependent and writer-independent results are given in Tables 1 and 2, respectively.

Table 1. Accuracy of the writer-dependent case.

Model	N-gram (%)		HMM (%)	
	Close-test	Open-test	Close-test	Open-test
Vertical	98.11	75.35	94.76	72.36
Horizontal	93.27	68.23	94.76	69.51
Left Diagonal	96.90	56.12	87.78	54.98
Right Diagonal	97.07	62.53	87.32	58.26
Combined	99.89	86.12	99.82	90.46

Table 2. Accuracy of the writer-independent case.

Model	N-gram (%)		HMM (%)	
	Close-test	Open-test	Close-test	Open-test
Vertical	89.28	69.76	86.21	68.32
Horizontal	78.62	63.51	78.70	61.27
Left Diagonal	84.07	58.49	78.79	59.03
Right Diagonal	84.90	63.40	79.01	60.95
Combined	96.30	84.08	97.24	85.10

4.2 Comparison with Other Feature Extraction Methods

Feature extraction is an important step for archieving the high performance of an OCR system. Thus, selecting the proper feature extraction method for the OCR system is a potential task. Due to idiosyncrasies in the characters of each language, a feature extraction method that proved successful in one language may turn out to be not very useful in other languages. In order to compare the performance of our feature extraction method with other existing methods, we explored some feature extraction methods that also extract global features: (1) Projection histogram [17], (2) Nafiz's method [10], and (3) Regional Projection Contour Transformation (RPCT) [2-3].

Projection histogram [17]: It is a well-known feature extraction method. It transforms 2-D data into 1-D data by a quantitative function. For example in the horizontal projection, $y(x_i)$ is the number of pixels with $x = x_i$.

Nafiz's method [10]: The authors did not name this feature exaction method, thus we refer to it as Nafiz's method. The method encodes the 2-D data and represents it as 1-D data as follows. First of all, the character image is separated into four directions. In each scanning line of the direction is then split transversely into four sub regions. Here, each region is coded by the power of 2. Next, in this scanning line, the medians of the black runs are identified. The code of regions where the median of runs are located, are summed up. Finally, by a clustering technique (vector quantization), the features of 2-D image are embedded into a sequence of 1-D codes, selected from a codebook. In the experiments, we ran tests on 16, 26 and 30-cluster data. However, the highest accuracy is obtained from the 30-cluster data.

Regional Projection Contour Transformation (RPCT) [2-3]: RPCT is a process of generating a feature sequence vector. Each feature sequence vector is a sequence of a direction of chain codes. To obtain a sequence of a chain code, the image is first projected into several directions such as vertical, horizontal, horizontal-vertical and diagonal-diagonal directions. A contour of each projected direction is then generated. By sampling points on the contour, we can find the sequence of a chain code for each direction of such contour and represent it as a feature sequence vector.

In the comparison of the four feature extraction methods, we performed the *open-test* in the *writer independent* environment. 9,320 characters, randomly picked from 10 subjects, were used. Half were hand-printed and half were natural written characters. Fig. 6 gives the recognition results obtained from the four feature extraction methods.

4.3 Result Analysis

There are several major factors in this recognition scheme that affect recognition accuracy: (1) combining techniques and their parameters, (2) the number of clusters and (3) the number of states of HMM. In order to archieve high accuracy, these factors are analyzed.

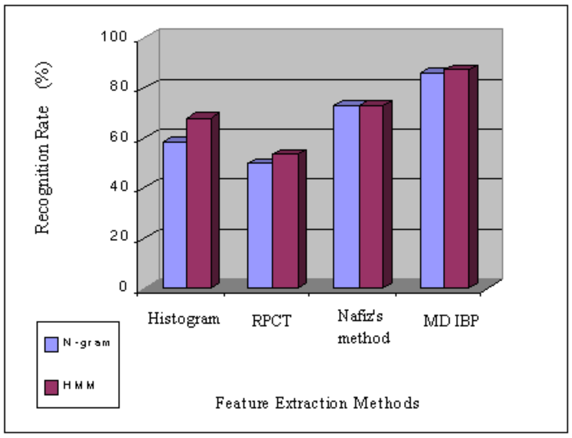


Fig. 6. Comparison of four feature extraction methods

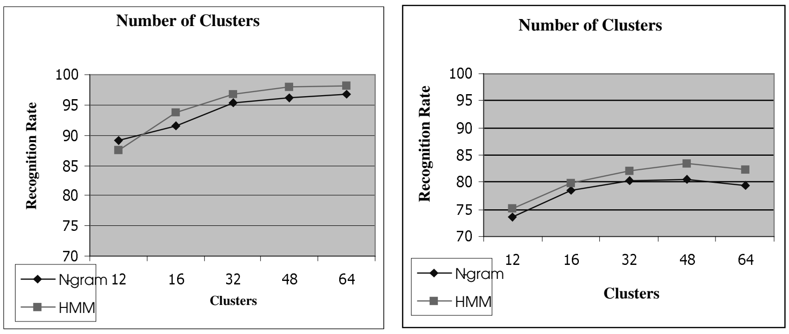
4.3.1 Combining Techniques & Parameters

Based on the recognition results in Tables 1-3, we found that a simple model such as pure vertical, pure horizontal and two pure diagonal models gave low accuracy. To achieve higher accuracy, combining multiple models is a possible solution. The combining technique is to merge the probabilities of multiple models using an equal weight. After this technique was applied to the recognition system, the recognition results improved.

The *n*-gram used in this work can be viewed as a double combining model since the model occupies two steps of a combining process. As the former step, the probabilities of unigram, bigram, and trigram are merged (interpolated) together by using some coefficient constants (i.e. α, β and δ). The latter combining step is to combine multiple models. In this process, we used an equal-weight combining method. In the former step, from a number of preliminary experiments, we found that the equal-weight method yielded unacceptable accuracy. Thus, unequal-weight coefficient constants were applied. We examined a number of coefficient constants and found that α=0.05, β=0.95, and δ=0.05 yielded a higher accuracy than others.

4.3.2 The Number of Clusters

In this experiment, we show that the recognition rate is affected by not only parameter adjustment in the recognition model and the model combining techniques, but also by the number of clusters. The reason why we require a clustering technique is to reduce the variations of raw feature vectors, which results in accuracy improvement, so this technique is crucial. However, the appropriate number of clusters is also significant. Fig. 7 shows that recognition accuracy depends upon the number of clusters. Both close-test and open-test are investigated. The 32-cluster gains relatively high accuracy, but it is a little bit lower than those of the 48 and 64 clusters. However, the computation time is drastically lower. Thus, the 32-cluster is applied in this work.



(a) (b)

Fig. 7. Numbers of clusters (k groups); (a) Close-test with WI, (b) Open-test with WI

4.3.3 The Number of States of HMM

Deciding the appropriate number of states of HMM is an important question. We investigated how many states should we have in the model in order to achieve high accuracy. Table 3 shows that the recognition rate depends upon the number of states. Notice that when we experiment on the full data set with a large number of states, it takes a long time to finish the recognition task. However, the recognition rate is relatively high.

Table 3. The number of clusters and states (Close-test, Writer-independent)

The number of Clusters	Hidden Markov Model (%)			
	4 states	8 states	16 states	32 states
12	61.61	69.55	85.61	87.45
16	71.15	77.69	88.90	93.72
32	80.15	85.78	92.93	96.70
48	84.35	90.68	94.75	98.02
64	87.56	92.38	96.43	98.26

5 Conclusion and Future Work

Local features are widely applied in several existing Thai OCR systems. In natural handwriting, these local features are often missed. Instead of using such local features, this paper presents a method to extract features from handwritten characters using so-called multi-directional island-based projection. Two statistical recognition approaches using interpolated n-gram model (n-gram) and hidden Markov model (HMM) are also proposed. The performance of our feature extraction and recognition methods is investigated using nearly 23,400 hand-printed and natural-written characters, collected from 25 subjects. The results showed that, in situations where local features are hard to detect, both n-gram and HMM approaches achieved up to 96-99 % accuracy for close tests and 84-90 % for open tests. In future, we plan to compare these methods with other techniques, such as the neural network, the time-delay neural network and the deformable template. These methods will be used for comparing the recognition rate and the computation time.

References

1. Airphaiboon, S., and Kondo, S.: Recognition of Handprinted Thai Character Using Loop Structures. IEICE Trans. INF.&SYST., Vol.E79-D, No.9, Sep (1996) 1296-1304.
2. Dehghani, S., and Nava.: Off-line Recognition of Isolated Persian Handwritten Characters Using Multiple Hidden Markov Models. International Conference on Information Technology: Coding and Computing (2001) 506–510.
3. HEE-SEON P., and SEONG-WHAN L.: Off-line Recognition of Large-set Handwritten Characters With Multiple Hidden Markov Models. Pattern Recognition, Vol. 29, No. 2 (1996) 231-244.
4. Jirayusakul, A., and Siriboon, K.: On-line Thai Handwritten Recognition using HMM. The Fifth Symposium On Natural Language Processing 5 (2002) 193–195.
5. Kimpan, C., Itoh, A., and Kawanishi K.: Recognition of Printed Thai character using a matching method. Proc. IEEE, vol. 130, Pt. E, No.6, Nov. (1983) 183-188.
6. Kimpan, C.: Printed Thai character recognition using topological properties method. INT. J. Electronics, vol.60, No.3 (1986) 303-329.
7. Kundo, A.: Handwritten Word Recognition Using Hidden Markov Model. Handbook of Character Recognition and Document Image Analysis. Edited by Bunke, H., and Wang, P., S., P. Singapore: Uto-Print (1997) 157-182.
8. Meknavin, S., Kijsirikul, B., Chotimongkol, A., and Nuttee, C.: Progress of Combining Trigram and Winnow in Thai OCR Error Correction. The 1998 IEEE Asia-Pacific Conference on Circuits and Systems (1998) 555-558.
9. Methasete, I., Jitapunkul, S., Kiratiratanaphrug, K., and Unsiam, W.: Fuzzy feature extraction for Thai handwritten character recognition. The Fourth Symposium On Natural Language Processing (2000) 136-143.
10. Nafiz, A., and Fatos., T. Y.: One Dimensional Representation Of Two Dimensional Information For HMM Based Handwritten Recognition. IEEE (1998) 948 –952.
11. Nishimura, H., Kobayashi, M., Maruyama, M., and Nakano, Y.: Off-line Character Recognition Using HMM by Multiple Directional Feature Extraction and Voting with Bagging Algorithm. Document Analysis and Recognition (ICDAR) (1999) 49-52.
12. Pessoa, and Lucio, F. C.: Multilayer Perceptrons versus Hidden Markov Models: Comparisons and Applications to Image Analysis and Visual Pattern Recognition. A Qualifying Examination Report: Doctoral. Electrical and Computer Engineering, Georgia Institute of Technology (1995).
13. Prokharatkul, P., and Kimpan. C.: Recognition of Handprinted Thai Characters Using the Cavity Features of Character Based on Neural Network. IEEE (1998) 149-152.
14. Prokharatkul, P., and Kimpan. C.: Handwritten Thai Character Recognition Using Fourier Descriptors and Genetic Neural Networks. The Fourth Symposium On Natural Language Processing (2000) 108-123.
15. Rabiner, L. R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceeding of The IEEE 77 (1989) 257-286.
16. Ren, T. I.: Pattern Recognition and Complex Systems. Doctoral dissertation, Antwerpen University (2000) 25-27.

17. Trier, O. D., Jain, A. K., and Taxt, T.: Feature Extraction Methods for Character Recognition—A Survey. Pattern Recognition 29 (1996) 1-24.
18. Veerathanabutr, P., and Homma, K.: The off-line Thai handwritten character recognition. The Fourth Symposium On Natural Language Processing (2000) 124-135.
19. Waleed, H. A., and Nikola K. K.: The Concepts of Hidden Markov Model in Speech Recognition. Technical Report TR99/09, Department of Knowledge Engineering Lab Information Science Department, University of Otago, New Zealand (1999).
20. Wongtapan, C., Theeramunkong, T., and Sinthupinyo, S.: Off-Line Isolated Handwritten Thai Character Recognition Using Interpolated N-gram Models. The Fifth Symposium On Natural Language Processing 5 (2002) 110–116.
21. Yang, H., Mou-Yen, C., and Amlan K.: Handwritten Word Recognition Using HMM with Adaptive Length Viterbi Algorithm. IEEE (1992) 153-156.

A Cache-Based Distributed Terabyte Text Retrieval System in CADAL

Jun Cheng, Wen Gao, Bin Liu, Tie-jun Huang, and Ling Zhang

Chinese Academy of Sciences, Beijing, 100080, P.R.China
{jcheng,wgao,bliu,tjhuang,lzhang}@jdl.ac.cn

Abstract. The *China-America Digital Academic Library* (CADAL) project aims to create a searchable collection of one million digital books freely available over the Internet. For this, a terabyte text retrieval system is required. This paper presents a cache-based, distributed terabyte text retrieval system, with fulltext retrieval, distributed computing and caching techniques. By distributing data by subject on different index servers, query searching is limited to specific index servers. With cache servers, response time is reduced. When queried, the system returns only highly relevant search results, to reduce the workload on the network. The prototype system shows the effectiveness of our design.

1 Introduction

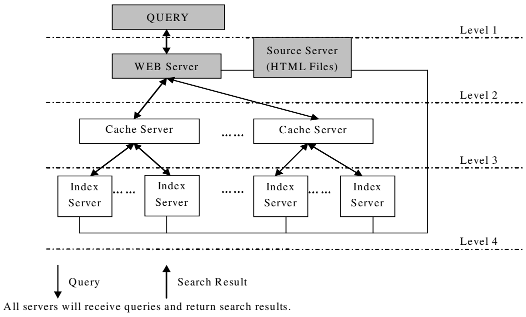
In the *China-America Digital Academic Library* (CADAL) project, the first goal is to have one million digital books on the Internet, in two or three years' time, put there by twenty libraries of high schools. These books will be freely available over the Internet. The second goal is to carry out research in areas such as data mining, intelligent information retrieval and ontology construction [1].

A terabyte text retrieval system is one of the sub-tasks in the project[2], due to the large number of books. To search this data quickly and efficiently, we need a fulltext retrieval system. Some web search engines use distributed systems to retrieve WebPages – e.g. Google and Alltheweb. There are differences between web search engines and the full text retrieval system in digital libraries. First, data on the web is dynamic while data in digital libraries is relatively static. Second, cataloguers have classified the digital books according to the taxonomy, generating semi-structured data. Because of these two factors, we can improve on the performance of existing search models.

2 System Design and Implementation

Existing commercial fulltext retrieval systems (TRS, ISearch, etc.) use both Chinese characters and words as basic index terms. They can hold about 6GB of text on one computer, and produce the search result in one second without the cache. If we want to work with terabyte text using these kinds of search engines as the core of our system, we have to adopt a distributed framework. The following figure shows the structure of the system.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 352-353, 2002.



- *WEB Server.* The web server has two functions. First, it receives the query that users input, and distributes it to cache servers according to the subject(s) of the query. The second function is to gather the search results of cache servers and merge them.
- *Cache Server.* Each cache server also has two functions. First, it receives the query statement from the web server. If the server has cached the same query and the search results (from a previous user), it will return the search results to the web server; otherwise it will decide which index server the query should be sent to, according to the query subject. Second, the server waits to receive the search results of each index server. When it gets all the search results, it merges them and returns a part of them to the web server, and puts all of them in cache. When the same query is subsequently received, the cache server can return these results to the web server without searching the index servers.
- *Index Server.* Index servers use commercial fulltext search engines as their core. Each index server also has two functions. First, it receives the query statement from the cache server. By searching its own index, it obtains the results and returns them to the cache server. Second, when we add new data to the source server, we can use index servers to index the new data. Current search engines return only the top thousand search results to a query. However, digital libraries need index servers to provide precise results from the text base. Another difference is that books are indexed according to their subjects as defined in the taxonomy, so the search range can be limited to specific index servers.

3 Conclusion and Future Work

We are trying to add text categorization to the system. By automatically categorizing user queries, users need not configure their search query. When the search results are returned, we will display them according to their subjects. Users can then select the subjects that they are most interested in. The system also needs more tests. Further, we do not have enough digital resources. There are only 4GB of data in one index server and only five index servers in all. The construction of the whole system structure has not been completed. In the future, we wish to explore some new retrieval models and text clustering approaches.

References

1. Bin Liu, Wen Gao, Ling Zhang, Tie-jun Hang, Xiao-ming Zhang, & Jun Cheng, Toward a Distributed Terabyte Text Retrieval System in China-US Million Book Digital Library. Joint Conference on Digital Libraries 2002, p. 7.
2. Zhihong Lu, Scalable Distributed Architectures for Information Retrieval. University of Massachusetts at Amherst, Ph.D. dissertation, 1999.

Rural Digital Library: Connecting Rural Communities in Nepal

Binod Vaidya and Jagan Nath Shrestha

Department of Electronics and Computer Engineering,
Institute of Engineering, Tribhuvan University, Kathmandu, Nepal
bvaidya@ioe.edu.np

Abstract. In Nepal, about 86% of the population still live in rural areas, and are dependent on subsistence agriculture. The information revolution poses particular challenges for knowledge-based services in Nepal. The context within which these challenges present themselves is that of globalization, in a world of increasing disparities between the rich and poor among and within nations. People in remote areas without electricity should not be deprived any longer of access to Information and Communication Technologies (ICTs), if one wants to avoid grave consequences arising from IT illiteracy. Keeping this in mind, His Majesty's Government, Nepal has announced some policies for ICT diffusion in Nepal. The paper presents suggestions for ICT deployment in rural Nepal and the concept of a Rural Digital Library, its technical aspects and cost-effective solutions for rural Nepal. It also identifies some knowledge-based services to deliver to rural communities.

1 Introduction

Nepal is a mountainous landlocked country with fragile and steep topography located on the southern slopes of the mid-Himalayas. The population of Nepal is estimated to be 23.4 million people, with at least 15 major racial groups [1] .

About 86 per cent of Nepal's population still live in rural areas, and are dependent on subsistence agriculture for their livelihood. Due to a lack of exposure, they conduct their activities in traditional ways with little or no modern technology. They lack access to basic services such as education and health care, and consequently migrate to urban areas, often in search of employment.

Having high growth potential, the information technology (IT) sector can provide developing countries with unprecedented opportunities, especially through globalization, which is a process to increase contacts between people across national boundaries – in economy, in technology and in culture.

Information and Communication Technologies (ICT) and related industries belong to the fastest growing segment of the global economy, and thus offer huge opportunities in investment and employment. In this context, the development of the IT sector is essential not only to participate in the global economy but also to fight poverty in developing countries.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 354-365, 2002.

Due to the lack of access to information and knowledge, living conditions in rural areas are not good. If access to information could be ensured for the poor, significant progress in poverty reduction could be achieved.

2 Role of ICT in Rural Development

The choice and use of a particular ICT for information access is dependent on its effectiveness in reaching the poor. Literacy and education will not only make people able to access information, but also empower them to broaden their choices and opportunities to improve their income levels and reduce social deprivation.

Thus, ICT is a powerful tool that can help economic growth as well as contribute towards progressive social change. In a developing country like Nepal, providing access and information on issues like education, health care, agriculture and human rights, is important as increased access to information goes hand in hand with socio-economic and overall human development.

In addition, recent technological advancements in the IT domain have opened up avenues of economic opportunities on which Nepal could capitalize.

ICTs can play a catalytic role in developing rural areas as follows:

- *Capability building and training* – Building human capabilities begins with basic education and literacy, and can help break the cycle of poverty and deprivation. We need well-trained teachers in IT to disseminate IT knowledge and skills training to a broader segment of the population.
- *Creating employment* – By creating employment opportunities in the rural areas with computer literacy training, software development and tele-centers, ICTs can help bridge the gap between urban and rural communities and reduce the rural-urban migration problem. Moreover, by providing training on computers, those trained may become small-scale entrepreneurs.
- *Targeting marginalized groups* – Most rural poor people lack the power to access information. ICTs can benefit all stakeholders, in particular youths and women. Other disadvantaged groups that can be targeted include the disabled and subsistence farmers.
- *Empowering rural communities* – ICTs can empower rural communities so that they can contribute to the development process. With ICTs, rural communities can acquire the capacity to improve their living conditions, and become motivated through training and dialogue with others to a level where they can make decisions for their own development. ICTs have the potential to penetrate under-serviced areas and enhance education through distance learning, facilitate development of relevant local content, and bring about faster delivery of information on technical assistance and basic human needs such as education, agriculture, and health.

3 Challenges in the Use of ICT

The information revolution poses particular challenges for knowledge-based services in Nepal. These challenges deal with participation in the information society : how

ICT impacts on access, cost effectiveness and quality of services. The context in which these challenges present themselves is that of globalization, in a world of increasing disparities between the rich and poor, among and within nations.

Internet access is the way to spread ICT. However, access is concentrated in high-income OECD countries, with only 14% of the world's population accounting for 79% of internet users [2].

The so-called "digital divide" is therefore an appropriate warning that ICTs are far from neutral. In Nepal, there is a serious information gap between those who have access to ICTs and those who do not. Knowledge and information are basic ingredients of the knowledge-based industries essential for facilitating rural development.

There are several challenges in the use and diffusion of ICTs for rural development in Nepal, which can be depicted as follows:

- *Policy considerations*: ICTs are dependent on national policy and regulations. In Nepal, the formulation and implementation of policies in the ICT sector is still rudimentary and calls for an integrated set of laws, regulations and guidelines that shape the generation and utilization of ICTs.
- *Infrastructure:* ICTs rely on physical infrastructures such as telecommunications and electricity. Even when such infrastructure is in place, difficulties arise when they are poorly maintained or too costly to use.
  - *Telecommunications:* Nepal Telecommunication Corporation (NTC) is a predominant operator for providing basic telephone and mobile services, and has basic monopoly in this sector. There is low tele-density in rural areas. Internet services are concentrated in the major cities of Nepal, so there are very limited points of presence (POP) of Internet Service Providers (ISPs). In 2000, Internet hosts are 0.1 per 1000 people whereas telephones (mainline and cellular) are 11 per 1000 people in 1999 [2]. These disparities impede internet access and hinder connection to the global network.
  - *Electricity*: In Nepal, the electricity supply is limited to about 18 per cent of the total population, while the rural population has just about 5 per cent. The best means of access to electricity in unconnected remote areas is the use of renewable energy sources such as solar PV, micro-hydro, etc. Although there is great potential for the use of such locally available resources, renewable energy technologies are not widely used in Nepal [3].
- *Local content and language barrier*: Information available through ICTs is mostly in English, which the majority of rural communities cannot understand. There is a marked shortage of relevant materials in local languages relevant to their needs. Moreover, Nepali fonts are yet to be standardized.
- *Literacy rate in rural areas*: Currently, the overall literacy rate for males and females in Nepal are 58% and 22.8% respectively. A large proportion of the rural population of Nepal is illiterate, so these individuals are disadvantaged and lack the basic skills required to harness the benefits of ICTs [2].
- *Availability of human resources*: Users of ICTs have to be trained in the use, application and maintenance of ICTs before they become confident and comfortable enough to use them. However, IT manpower is lacking, especially in the rural areas of Nepal.

- *Affordability to IT*: About 38% of the total population in Nepal live below the income poverty line. Working residents of the rural communities earn about one dollar per day and cannot afford the costs associated with extending telecommunications services to their rural communities.

The development of the ICT sector in Nepal is a tremendous task. IT facilities currently are grossly inadequate – limited to only a handful of the urban educated population. IT tools that facilitate overall development are yet to materialize. In general, the factors on which the development of the IT sector are founded: education, institutional environment and infrastructure, require improvement.

4 National Approaches to ICT

Considering the need to spread ICTs in the country, His Majesty's Government, Nepal (HMG/Nepal) in its Ninth Five Year Plan (1999-2002), has shown its determination to use ICTs for the overall national development.

Under the HMG/Nepal's policy, basic telephone services will be provided to all the Village Development Committees (VDCs) by 2003. Accordingly, NTC has plans to extend its telephone network to the far-flung VDCs in the remote areas of Nepal under the Special Rural Telecommunication Project, using Wireless Local-loop (WLL) technology. Under this project, the total capacity of the WLL networks will be about 32,000 telephone lines.

Furthermore, NTC has plans to install 1000 Very Small Aperture Terminal (VSAT) stations to provide basic telephone services in the remote northern parts of Nepal within the next 2-3 years. The ongoing Rural Telecom development program will provide at least 2 lines per VDC through the VSAT project, whereas WLL technologies will provide a minimum of 4 lines per VDC, with the capacity to increase to 16 lines [4].

Nepal Telecommunication Authority (NTA), which is under the Ministry of Information and Communication (MOIC) HMG/Nepal, began a policy to liberalize the provision of various telecommunication services. It has been determined that the private sector will be allowed to operate basic telecommunication services by 2004.

Recognizing the need to prioritize the development of the IT sector, the Ministry of Science and Technology (MOST) HMG/Nepal has made a concrete and long-term decision by approving the National Policy on IT, with a vision to "Place Nepal in the Global IT Map within five years". In "IT policy 2000", the objectives are to make IT accessible to the general public and increase employment, to build a knowledge-based society and to establish knowledge-based industries [5] [6].

MOST HMG/Nepal has announced the following strategies for its IT policy:

- Competent human resources will be developed by the public and private sectors working together;
- The government will promote and facilitate the rapid diffusion of the technology, and regulate its operation;
- Information technology will be applied for rural development programmes;
- Internet facilities will be provided to all VDCs of the country;
- Information technology network shall be extended to rural areas;

- Computer education will be included in the school curriculum;
- E-commerce will be promoted within an appropriate legal framework;
- IT related knowledge-based industries will be encouraged and promoted;
- IT service providers will operate in a competitive environment to ensure that services rendered will be of good quality and reasonably priced;
- IT will be used for all government activities and provide legal sanctions to them.

Below are some of the action plans to be adopted for the implementation of the national IT policies:

- *Infrastructure development*: Information network shall be established all over the country. Basic needs such as telecommunication and electricity services shall be provided. Internet connectivity shall be gradually extended to rural areas.
- *Human Resource Development*: Basic facilities, including Internet shall be provided to educational institutions for IT education. Computer education shall be effectively implemented in all the schools in Nepal by 2010. Thus IT shall be used to improve the quality of education.
- *Diffusion of Information Technology*: Educational institutions and hospitals, where telecommunication and electricity services are already available, shall be encouraged to use IT enabled services. Even in places where electricity service is not available, the diffusion of ICTs through solar power system shall be encouraged. Distance learning programs shall be realized through the Internet. Content and applications shall be prepared in the local language for rural people.
- *Promotion of knowledge-based services*: E-commerce, tele-medicine, tele-processing, distance learning shall be promoted by creating the necessary legal infrastructure and other arrangements.
- *Facilities*: An IT Development Fund shall be established to create public awareness about IT, assist rural networking, develop IT with market management, generate required manpower for this sector, and make social services easily accessible where such technology is used.

HMG/Nepal has identified a range of strategies to develop the IT sector, such as distance education, human resources development, promotion of IT industries. Policy measures identified for these strategies include the creation of an enabling environment in the form of increased private sector and non-government organizations' (NGOs) involvement, and access to IT services in rural areas.

Since 2002, MOST, HMG/Nepal has launched a Human Resource Development Programme to produce IT manpower of various levels - semi-skilled, skilled and highly skilled. In this programme, about 10,000 IT workers will be trained.

In line with its IT policy, MOST, HMG/Nepal has established the National Information Technology Center (NITC), which has as its main objective the carrying out of various tasks to facilitate the promotion and development of ICT sectors in Nepal. Accordingly, NITC has plans to assist government agencies in the computerization of their record-keeping operations and to facilitate marketing of software industries.

MOST, HMG/Nepal has initiated a programme called "ICT for Poverty alleviation/Development", with a target to establish 25 Rural Information Kiosks in various Districts of Nepal by 2004, in addition to 25 existing Information kiosks run under the Rural Urban Partnership Programme (RUPP)/ UNDP.

5 Rural Digital Library

In the Information Age, the "Digital Library" concept is emerging as a model for information retrieval and storage, knowledge acquisition and management, and communications for rural development [7] [8].

5.1 Rationale

The Proposed Rural Digital Library (RDL) shall be a partnership of Local Government authorities, NGOs, schools, and private organizations committed to creating a community-owned library system for rural development. Driven by ICT, the RDL can be viewed as providing the capability to transform individuals and communities, and to alleviate poverty to some extent.

In a world with an increasing number of social, political, and economic stress points, there is no alternative to the adoption of ICT by developing countries like Nepal. It is obvious that an information infrastructure is an essential social development tool as the world enters the 21st Century.

The RDL can be a response to the challenge and promise of ICT for rural development. It is a cornerstone of social development in the Information Age, with the mission to empower rural communities with the requisite technology infrastructure, financial capacity, and human resources to manage the transition to the information culture through a locally determined process.

The goal of the RDL is to develop a web-based resource network consisting of at least one electronic library with access to the Internet, in order to educate the isolated rural communities so that the global knowledge network can bring vast benefits to them. It is the vision of the RDL network to empower local communities through local self-determination and to help them benefit from access to global information.

5.2 Rural Digital Library: Structure

The Proposed RDL network design consists of one Main Library Server Center supporting 10 Local Server Centers in concentrated clusters of communities. The Main Library Server Center will be located in District Development Committee (DDC), whereas Local Server Centers will be located in ten neighboring VDCs. Each local server center can serve up to at least 20 User groups, as shown in Fig. 1.

The RDL network consists of four key components:

- *Systems Management Group (SMG):* A central Systems Management Group is meant for supporting knowledge management systems. It consists of representatives of local government authorities, NGOs, local schools and private organizations, and shall solicit funds, develop ICT architecture, provide project management and support library content development.
- *Main Library Server Center (MLSC):* A Main Library Server Center shall be located at a school in DDC, which will be used to support multiple local servers in a concentrated cluster of rural communities. MLSC will also provide Internet access to all the Local Server Centers through the NTC backbone network. In

- MLSC, a support division will be formed to provide direct support services to the ICT infrastructure in rural communities. The support division will develop technical support, service equipment, maintain systems operations and provide training. MLSC will have high-end computer servers, computer workstations, internetworking equipment, storage equipment, solar electric power equipment, and library reference materials.

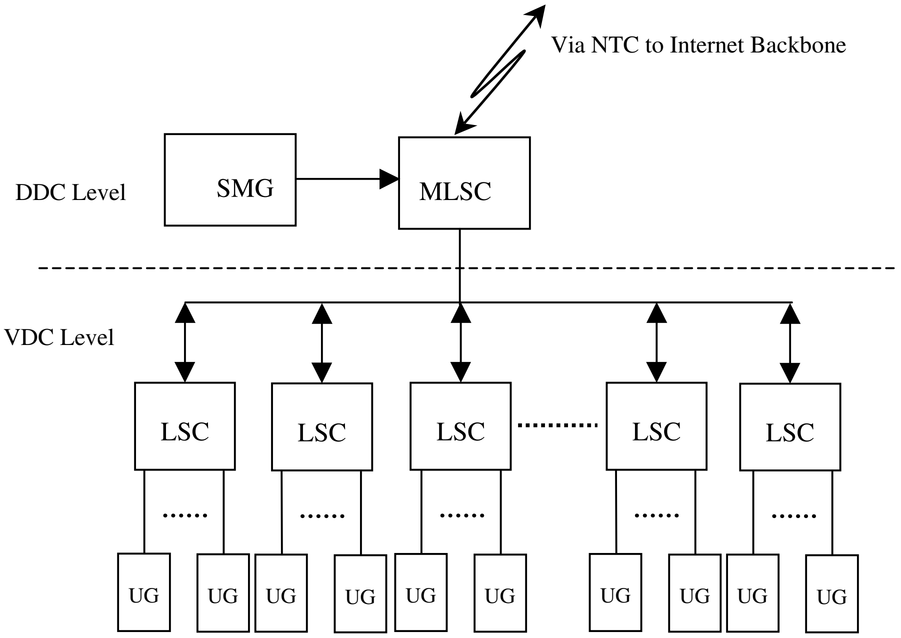


Fig. 1. Structure of Rural Digital Library

- *Local Server Center (LSC):* Local Server Centers shall be located at schools in different neighboring VDCs. LSC shall have staff trained at MLSC. The LSC will provide direct training and support services to its host community and networked User Groups within the Local Area Network (LAN) service areas. LSC shall consist of computer server, low-power computer workstations, internetworking equipment, and solar electric power equipment.
- *User Group (UG):* To provide ICT services at the user community level, the LSC will develop and maintain locally-owned and operated User Groups (UG). It shall be connected to the LSC either by line-of-sight, wireless technology or in LAN if closely located. UGs may be located at public access facilities such as schools, health centers, agricultural centers and government offices. An UG may consist of a computer network or stand-alone low-power computer, internetworking device and solar power equipment.

5.3 Technology Considerations

Wired networks are a logical and efficient choice for most networks wherever local loops and dedicated leased lines provided by telecommunication companies are affordable and easy to install. But wired networks are a restriction and a handicap for networks situated in remote areas where installation of dedicated leased lines is too costly and maintaining reliability of such dedicated links is not cost effective. Wireless networks provide a new level of flexibility to environments that are not well served by the traditional wired network.

Advantages of wireless networks are as follows:

- Fast setup and implementation- Wireless networks can be installed faster than traditional wired networks since there is no need for hard-wired cabling between host systems. They offer flexible installation and can be up and running with minimal setup and initiation.
- Disruption proof- Traditional wired networks are susceptible to disruption due to construction and repair operations with pole-mounted transmission cable, and subject to a high incidence of lightning and/or high winds. Wireless networks are less prone to disruption by external agents.
- Last mile communications- Compared to wired networks, wireless networks offer a far better range of up to 50 to 70 miles between units. This gives greater coverage through the network and provides "last mile" communication links to remote areas.
- Movable installation- One of the salient advantages of wireless networks is that the installation can be easily moved. If the situation so warrants, the wireless network can be easily relocated and made operational faster.

For basic telecom services in rural Nepal, wireless technologies such as WLL and VSAT are appropriate. In VSAT technology, C band antennas are larger than those of Ku band, and the power requirements of C band equipment are bigger than those of Ku band. In the deployment of VSAT in rural areas, Ku band has salient advantages over C band.

Depending on the power of the transmitters and the sensitivity of the receivers, wireless networks can be a cost effective option for local and remote area networks and for coupling a number of other wireless networks. The future of wireless networking is bright in meeting user needs and expectations, and the technology is well matured and will complement the wired world to enable seamless integrated networking.

However, issues such as facility obstructions, antenna locations, signal tuning, movement patterns, etc. must be evaluated and taken into account before committing to a wireless network installation.

The selection of computer equipment is an equally vital task. The factors to consider are as follows [9] :

- Power requirement – This is a potential hindrance, since most rural areas have no access to electricity.
- Cost – The high costs of computer hardware and software (operating system, application software) is also an obstacle. It should be possible to manufacture easily transportable, perfectly serviceable computers for basic computer

applications such as word processing, email and internet connectivity in rural areas.

- System performance - We emphasize simplicity and efficiency. Operating systems that consume very little resources and are fast enough to handle basic computing are preferred. Computers with simple stripped down versions of hardware and software will be instrumental in facilitating the entry of rural people into the information society.
- Support and maintenance - A key factor in the effective deployment of ICT in rural areas is support and maintenance services after the implementation. For sustainable utilization of computer systems, capable technical manpower should be trained to provide at least a basic level of support and maintenance.

Table 1. Factors influencing Pentium Computers

Factors	Pentium Desktop	Pentium Laptop
Power requirement	Computer without monitor – 150W CRT monitor – 120W LCD monitor – 45W	Laptop – 40W to 70W Notebook – 15W to 25W
Cost[1]	Locally assembled [2] Pentium III – NRs. 45,000 Pentium 4 – NRs. 60,000 Branded [2] Pentium III - NRs. 100,000 Pentium 4 – NRs. 120,000	Laptop Pentium III[2] – NRs. 200,000
System Performance	– Faster speed – Overall performance is good – Operating system (OS) and applications required more resources	– Faster speed – Overall performance is good – OS and applications required more resources
Support and Maintenance	After sales service is easier for the locally assembled one.	After sales service is quite difficult due to unavailability of components..

Table 2. Factors influencing ARM-based Computers

Factors	ARM Desktop	ARM Transportable
Power requirement	Entire Computer - 110W to 145W	Entire computer - 17W to 8.5W
Cost[3]	NRs. 120,000[4]	NRs. 80,000[5]
System Performance	– Medium speed – Overall performance is moderate – OS and applications required less resources	– Slow speed – Limited functionality – OS and applications required less resources
Support and Maintenance	Not available in Nepal yet	Not available in Nepal yet

[1] Price is as of May 12, 2002 (1 USD = NRs. 78.35)
[2] DOS Trading Pvt. Ltd., Kathmandu, Nepal
[3] Price is as of May 12, 2002 (1 USD = NRs. 78.35)
[4] http;//www.desktop.demon.co.uk/computers/riscpc/index.html
[5] http://www.explan.co.uk/hardware/solo

We are looking at both Pentium and Advanced RISC Machine (ARM) computer systems for ICT deployment in rural areas.

An x86 processor such as Intel's Pentium, Cyrix's MII, AMD's K6 are based on Complex Instruction Set Computing (CISC). Unlike x86 processors, ARM chips are based on Reduced Instruction Set Computing (RISC). ARM processors have exceptionally low power consumption when compared with the x86/Pentium series. ARM processors such as ARM-7500, StrongARM and ARM9 are generally used in Desktop ARM computer, whereas ARM-7500 or ARM9 is used in the Transportable ARM computer, so-called ExpLAN's Solo [10] [11].

Factors influencing the use of Pentium and ARM Computers are listed in Table 1 and Table 2 respectively.

5.4 Content and Knowledge-Based Services

Information should be delivered through an intranet (web or email access), as web-based information generation is easy and inexpensive. Uploading designed materials for rural communities in the Nepali language on the web will be essential for desired technology transfer. The information needs may vary from community to community. However, some surveys indicate that rural communities, being dependent on agriculture, are very interested in information which can improve agriculture. Therefore, information on horticulture, livestock, vegetables and cash crops as well as their markets, inputs, credit related information, will be of interest. The surveys also indicate that information related to health and education will be useful. In the process of generating information, the local content or indigenous knowledge should be posted at the Main Library Server Center. This will help in the conservation of traditional knowledge, serving both utilitarian and cultural purposes.

Design criteria are essential not only to access the global knowledge network, but also to empower rural communities to generate and distribute content in a standardized platform in order to collaborate with other affected communities to solve common problems.

The use of Internet technology requires an intermediary to operate and broker the information to the rural people, due to their low literacy level. The intermediary will need to be trained and will need good aptitude for this job. For effective technology transfer, the local community leaders should be trained along with the NGOs' staff. School children can be intermediaries between their parents and the Internet. The effectiveness of field staff of service delivery organizations in technology transfer to the rural poor is going to be enhanced by the availability of information through ICTs.

The following seven knowledge-based services have been identified to deliver content to rural communities:

- *Governance Information System*: Development of Governance Information System shall be a response of our constitution stating that every citizen has the right to information. It will bring transparency to local as well as regional government operations and the government-public interface, and improve the efficiency and effectiveness of the government's service delivery.
- *Agriculture Information System*: This system will contribute to expanding markets for agriculture products, increase in sale of agriculture products to the rural

population, provide increased market access for the rural population, eventually leading to increased incomes for the rural population.

- *Distance learning*: Distance learning programs will improve the quality of education and increase access to education. It will be focused especially on underprivileged groups such as the unemployed, women and older communities.
- *Tele-Medicine*: Tele-medicine will improve the quality of health-care services available to the rural population, and increase their access to health care facilities.
- *Productive Economic Activity*: Networking will vitalize local industries, leading to increased sales of products, increased market activities and increased income for the micro-entrepreneurs, thus helping to reduce rural poverty.
- *Energy and Environment*: Knowledge-based services in the areas of energy, the environment and natural resource management will contribute to expanding markets for environment-friendly products, lead to improvement in indoor air quality, increase awareness of alternative resources and environment conservation, and increase use of natural resources in a sustainable manner.
- *Natural Disaster*: ICT application for disaster management can provide rural communities with increased information about disaster prone areas and improve delivery of emergency services.

6 Conclusion

The development of ICT connectivity to the remote areas of Nepal is highly desirable Rural communities should be educated and provided with skills and knowledge that open up possibilities that lead to an improved quality of life. However, the state of IT enabled services in rural areas of Nepal in no way compares with that in urban areas. The striking disparities between urban and rural areas are also seen in IT education. This leads to a serious "digital divide" between those who have access to ICTs and those who do not. To close the information gap between urban and rural areas, low-power computer systems should be deployed in the rural areas.

Due to a sparsely distributed population in the remote areas of Nepal, it is economically viable to use wireless networks. Electricity in the unconnected remote areas would be provided by the use of locally available and appropriate renewable energy sources. Without basic infrastructures such as telecommunications and electricity, the fruitful realization of ICT applications is not possible.

It is evident that no wealthy developed country is information poor and no information rich country is poor and undeveloped. Thus, it is vital to focus on ICTs' diffusion for rural communities, so that rural people are exposed to information related to education, health care, agriculture, markets, etc. This will help to bring them out of poverty to a better life. It is, therefore, essential to ensure their accessibility to knowledge-based services, through the Rural Digital Library (RDL).

Not every individual will benefit directly from RDL. It is rather, through information disseminated by community leaders and key information providers such as healthcare workers, teachers, and community development specialists, that the impact of RDL information will benefit the entire community.

References

1. Results of Population on census 2001, HMG/Nepal National Planning Commission Secretariat, Central Bureau of Statistics, Kathmandu, Nepal, 2001.
2. Human Development Report 2001, UNDP, New York, 2001, 40 – 62.
3. J. N. Shrestha and B. Vaidya, "Role of Renewable Energy Technology for Information and Communication Applications in Rural Areas of Nepal", International Conference on Information Technology, Communication and Development (ITCD-2001), Kathmandu, Nepal, November 2001.
4. Nepal Telecommunication Corporation, Annual Report 2000/01, Nepal Telecommunication Corporation Kathmandu, June 2002, 175-182.
5. Formulation of Information and Telecommunication Policy and Strategy, Nepal, Pan Asia Networking (www.panasia.org.sg/grants/awards/99153.htm).
6. Information Technology Policy 2057 (2000), Information Technology for Development, IT Policy and Strategy Papers for Nepal, HMG/Nepal National Planning Commission Secretariat, Kathmandu, Nepal, 2001.
7. A Sharma and W Yurcik, “The Emergence of Rural Digital Libraries in India: The Gyandoot Digital Library Intranet”, ASIS Annual Conference (ASIS 2000), Chicago, USA, November 2000.
8. The Indonesian Digital Library Network (http://idln.lib.itb.ac.id).
9. J. N. Shrestha and B. Vaidya, "IT Education in rural Nepal using Renewable Energy Technology", Computer Association Nepal Conference 2002, Kathmandu, Nepal, January 2002.
10. ExpLAN Computers Ltd. (http://www.explan.co.uk).
11. Desktop Projects Ltd. (http://www.desktopp.demon.co.uk).

Collection Development for the Digital Age: The Case of Malaysia

Ahmad Bakeri, Abu Bakar, and Kaba Abdoulaye

Department of Library and Information Science,
Kulliyyah of Information and Communication Technology
International Islamic University Malaysia
Jalan Gombak, 53100 Kuala Lumpur, Malaysia

Abstract. Many libraries in Malaysia are currently experiencing a transition from a traditional print-based collection to one that holds a mix of print and electronic resources. This has led to the existence of so called hybrid libraries. It has also led to changes in collection development practices. The aim of this paper is to investigate the practices of and challenges confronting collection development in different library environments : traditional, hybrid and digital. Four libraries were selected to represent the three different environments : the Petaling Jaya Municipality Library, the Multimedia University Library, the International Islamic University Malaysia and the Hypermedia Library of Subang Jaya. Aspects of accessibility, including consortia access to electronic resources, governance, cost benefit issues, budgetary appropriations, selection criteria, and other collection development issues are discussed. It is hoped that raising these issues will contribute to strategic decisions of management and policy in the development of digital libraries in Malaysia.

1 Introduction

In recent years, collection development has witnessed dramatic changes in terms of materials handling and collection building. Now, it is technically feasible to have a library that does not need any library collection of its own at all. It needs only data communication resources and permission to legally access the materials its users need. Most libraries, however, still have a long way to go to reach this stage of development, especially those libraries operating in less developed countries. Some of these libraries are currently experiencing a transition from a traditional print-based collection to one that has a mixture of both print and electronic resources. This transition has resulted in changes to collection development practices and the role of librarians. This new breed of librarians has to find new ways of maximizing access to electronic resources, while attempting to fit emerging technologies into traditional library practices and functions.

In Malaysia, libraries have been slow to move towards an electronic environment. At the last count, there were only two libraries that could be considered as digital. This is in stark contrast with the tempo of development in information and

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 366-378, 2002.

communication technology (ICT) at the national level. Malaysia embarked on an ambitious plan to leapfrog into the Information Age by launching the Multimedia Super Corridor on 27 June 1998 – a world-first–to encourage invention, research and other ground-breaking multimedia and information technology developments. With the number of sophisticated investment, business, R & D and other incentives provided by the Government in the areas of ICT to accelerate Malaysia's entry into the Information Age, the library sector is supposed to be affected as well by this wind of change. But this change has not taken place, as reflected by the low count of existing digital libraries in the country.

So there must be underlying factors that work against the growth of digital libraries and the change to digital collections. There may be very good reasons for this underdevelopment in the library sector, but it is not the objective of this study to look into that. What we intend to address are the problems of changing from a traditional library collection to a digital collection, the solutions to some of the problems, and the prospects for the future development of digital libraries in Malaysia. It is hoped that such study can suggest a course of action that can be applied to hasten the development of digital libraries in Malaysia.

2 Digital Environment

It is difficult for anyone to associate a particular library to be of traditional, hybrid or digital model as there is no competent standard that may be used in our judgement of the status except for the indicator found in the digital resources of the library collection. All the three models will incorporate both print and digital resources in their collections and one may be different from another by virtue of the size of the digital resources in their collection. If size of print and electronic acquisitions is to be used as a measure of assessment, then judgement made will be tainted with element of biasness and subjectivity. Perhaps it is because of the digital content issue that the term "digital library" has been used to mean many different things over time. One way to overcome the classification problem is to scale the digital content to certain measurable dimension, such as any library that hold electronic materials less than 40% of the total collection is placed in the category of traditional, while those holding 40%-60% is categorized as hybrid and those above 60% belong to the digital library model.

The digital content of these libraries is shaped by the acquisition and selection process of the digital resources, cost and accessibility. Although in many ways selection criteria regarding electronic resources and print materials are similar, there are cases in which they differ. In her dissertation, Brody showed that in the case of electronic journal acquisition the reputation of the publisher and nature of the academic discipline are low indicators; issues of content are medium indicators; and the most important factors are recommendations from faculty members, cost, licensing issues and the influence of consortia [1]. Jewell, in his analysis for the Council of Library and Information Resources revealed that major research libraries used the following criteria for selecting their electronic resources : content (how it compares with the print), added value (value of search ability, currency), presentation and functionality (usability), technical considerations (software, hardware) licensing (access rights, costs), service impact (publicitiy) [2]

Cost is a vital ingredient in any acquisition work. Its position become more dominant when the acquisition is for electronic resources. As libraries develop collections that not only integrate but begin to replace print with electronic resources, libraries must contend not only with cost of the databases, electronic journals, and books and other electronic resources that are routinely purchased or licensed but also with the cost of hardware, software, technological innovations and the maintenance of these large files and systems over time. As pointed out by Miller, "to mount book and other materials into digital collections that are fully searchable, navigable and in compliance with emerging standards, not to maintain the maintenance of these large files and systems over time, the overhead can be staggering" [3]. Lesk attributed the slow growth of digital libraries to the question of cost including the large sum meted to the copyright owners. He argued that "in 1964, Arthur Samuel predicted that by 1984, paper libraries would disappear except at Museums. Why hasn't this happened? The biggest reason is that we cannot easily find $3 billion to fund the mechanical conversion of 100 million books to electronic form, plus the additional and probably larger sum to compensate the copyright owners for most of those books" [4].

Almost all libraries have to rely on the interlibrary loan services to cope with any deficiencies in their collection. However, users taste have changed during this transition towards electronic environment. They are beginning to expect electronic delivery that is speedy, full text and accessible at remote sites. This shift to on-demand delivery of materials from remote sites is a direct result of the recent proliferation of digital network. This delivery of materials from elsewhere "just in time" to answer a user's need has opened new opportunities for libraries. The libraries can now abandon the ideal to build a great comprehensive collection and instead spend their resources to provide an effective gateway to access the electronic resources. As Harloe and Budd argued "the historic quest for the great "comprehensive collection" has been superseded by the need to provide access to collective scholarly resources that no one library can afford" [5]. The other opportunity is for the libraries to reduce the cost by partaking with other libraries to form a consortium. It has been shown that negotiation for reduction in purchased price is possible if libraries lobby as a unified group. The Northeast Research Library Consortium, for example, calculated an annual savings of 25 percent on a $100,000 subscription [6].

3 Methodology

The study used a questionnaire for data collection. This instrument was preferable as most of the required information were mainly statistical data. Two academic libraries and two public libraries were selected for the study. These libraries are : International Islamic University Malaysia (IIUM), Multimedia University (MMU) in Kuala Lumpur, MPPJ Community library and MPSJ Hypermedia library. The questionnaires were personally sent/collected by the researchers to/from each library.

Only the MMU filled-in questionnaire was received through the fax. At least one week was given to each library for completing the questionnaire. The instrument of 22 questions, elicited information about the financial sources, annual budget for materials, and budget allocation according to types of materials. It also sought data on size of collection, cost of materials, channels for collection building, criterion for selecting library materials, and issues related to inter-library lending.

4 Data Analysis and Discussion

Participating Libraries

The investigated libraries were established between 1983 and 1999. The most recently established one is four years old while the oldest is 19. As illustrated in Table 1, there are 298 staff working at the participating libraries. Sixty one (21.4%) out of the 298 staff are professionals, and 237 (79.5%) are non professionals. Meanwhile, the staff population in each library varies. The IIUM reported the highest number of staff (169), followed by MMU (46), MPSJ (43) and MPPJ (40) respectively. Regarding professional staff, again, the IIUM reported the highest number of staff (34) followed by MPSJ (13), MMU (9), and MPPJ (5) respectively.

Table 1. Basic Information on the Libraries

Library	Year of establishment	Professionals	Non-professionals	Total
IIUM	1983	34	135	169
MMU	1997	9	37	46
MPPJ	1987	5	35	40
MPSJ	1999	13	30	43
Total		**61**	**237**	**298**

Financial Sources

Flourishing library activities depends heavily on sufficient financial support from the donors or the parent institutions. Many libraries are unable to carry out their excellent objectives and goals due to the financial incapabilities of providing necessary materials highly needed for the improvement of the library services. Table 2 shows financial sources for the participating libraries. State, local, or municipal government was found to be the main financial sources for the two public libraries, MPPJ and MPSJ, while academic libraries, IIUM and MMU, rely on federal government or corporations. None of these libraries used foundations, trustees or firms as sources for library budget.

Table 2. Financial Sources

Source	IIUM	MMU	MPPJ	MPSJ
Federal government	yes	no	no	no
State government	no	no	no	yes
Local/municipal government	no	no	yes	yes
Foundations	no	no	no	no
Trustees	no	no	no	no
Corporations	no	yes	no	no
Firms	no	no	no	no

Annual Budget for Materials

The annual budget of library materials within five years (1999-2001) varies among the participating libraries. Table 3 shows breakdown of the budget by year. The annual budget of IIUM has been increasing within the past four years. In 1999 the library reported 2,355,903[1] budget for library materials compared to 2,130,000 in the previous year. Another significant increase occurred in 2001, in which the library allocated more than seven million for the material compared to five million in the year before. Further analysis shows that, from 1997-1999, the MMU allocated two million on library materials, while the MPPJ allocated only 200,000. For the past two years, the MMU witnessed a decrease of annual budget from 12 million, in 1999, to nine million, in the year 2000/2001, while the IIUM budget increased from two million plus, in 1999, to five million, in the subsequent year, and seven million plus in the following year.

Table 3. Annual budget for Library Materials

Library	1997	1998	1999	2000	2001
IIUM	N/A	2,130,000	2,355,903	5,000,000	7,950,000
MMU	12m	12m	12m	9m	9m
MPPJ	200,000	200,000	200,000	300,000	500,000
MPSJ	N/A	N/A	N/A	N/A	550,000

Budget Allocation According to Types of Library Materials

The participating libraries were asked to indicate annual budget allocation for the different types of library materials. The materials were categorized as electronic (databases, CD, VCD and DVD) and printed materials. Findings show that, the responding libraries to this question allocated more than half of the annual budget on printed materials. In 1997, the MMU allocated the entire budget on printed materials. Most probably because the library was newly established and the need for the

[1] The entire amounts mentioned in this paper are in Malaysian currency, i.e. Malaysian Ringgit (RM).

electronic materials could be satisfied in other libraries. However, in the subsequent year, 1998, the university allocated more than one million for electronic materials, while the printed materials received more than ten million (Table 4). It should be noted that, other participating libraries also paid more attention to the printed materials than the electronic materials. For instance, from 1997-2001 the MPPJ library allocated between ten thousands to fifty thousand for electronic material, while an amount of twenty thousands to five hundreds thousands were allocated for printed materials in the same period. Similarly, for the year 2001, the MPSJ allocated twenty-six thousands for electronic materials, while the printed items were given five-hundreds fifty thousands.

Table 4. Budget Allocation According to Types of Library Materials

			Electronic Materials		
Library	**1997**	**1998**	**1999**	**2000**	**2001**
IIUM	N/A	N/A	N/A	N/A	N/A
MMU	N/A	1.5m	2m	2.5m	2.6m
MPPJ	10,000	10,000	10,000	10,000	50,000
MPSJ	N/A	N/A	N/A	N/A	26,000
			Printed Materials		
IIUM	N/A	N/A	N/A	N/A	N/A
MMU	12m	10.5m	10m	6.5m	6.5m
MPPJ	200,000	200,000	200,000	300,000	500.000
MPSJ	N/A	N/A	N/A	1m	550,000

Size of Collection

Table 5, illustrates collection size of the four investigated libraries. The libraries possess a considerable size of electronic and printed collection. However, the size or volume of printed material was much bigger than the electronic materials. Of the four participating libraries, the two academic libraries, IIUM and MMU, reported a bigger size of electronic and printed materials compared to the two public libraries. Of all, the IIUM occupied the first place, followed by MMU, MPPJ and MPSJ respectively.

Table 5. Collection at the End of 2001

Library	Materials	
	Electronic	**Printed**
IIUM	30,483	34, 4164
MMU	3,030	57,400
MPPJ	3,000	120,000
MPSJ	750	35,000

Unit Cost for Library Materials

Unit cost is defined as the cost expended to acquire one item of the respective material, i.e. total cost of item over number of item. The participating libraries were asked to indicate this cost for library materials. The analysis of data reveals mixed responses. Except for the IIUM, unit cost, either electronic or printed item, for all the participating libraries was found not less than RM30 but more than RM100. Furthermore, unit cost of the electronic or printed materials for MPPJ were between forty to hundred fifty, MMU between thirty to three-hundred for electronic materials and hundred fifty or above for printed items, and hundred for any item at MPSJ library (Table 6).

Table 6. Unit Cost for Library Materials

Library	**Materials**	
	Electronic	**Printed**
IIUM	N/A	N/A
MMU	30-300	150-above
MPPJ	40-150	40-150
MPSJ	100-above	100- above

Means for Building Collection

As illustrated in Table 7, all the participating libraries reported the utilization of "purchase" as a means for collection building. Academic libraries, IIUM and MMU, make use of all the four means listed in the table. On the other hand, public libraries, MPPJ and MPSJ, did not use "exchange" and "gift". Meanwhile, the MPPJ was only the participating library found not using "subscription" for collection building.

Table 7. Means for Collection Building

Library	**Means**			
	Subscription	**Purchase**	**Exchange**	**Gift**
IIUM	yes	yes	yes	yes
MMU	yes	yes	yes	yes
MPPJ	no	yes	no	no
MPSJ	yes	yes	no	no

Criteria for Selecting Electronic Materials

The participating libraries were also asked to indicate important criteria for selecting electronic materials. A checklist comprising 18 characteristics were listed for the purpose. As shown in Table 8, academic libraries, IIUM and MMU considered, almost, the entire listed criterion important. On the other hand, public libraries, MPPJ and MPSJ, do not see most of the criterion important for selecting electronic materials. The former voted 13 criteria, while the later chose only eight. Among the criteria considered not important to MPPJ and MPSJ were “uniqueness of content, capabilities of features”; “geographic parameters”; “hardware compatibility”; and “service implications”. Likewise, “license restrictions” was not important to MMU and MPSJ libraries. Further analysis shows that of the four participating libraries only the MPPJ did not consider **“quality”** as an important criterion for selecting electronic materials. Similarly, “number of access points/indexes”; “network capability”; “strength of retrieval”; “software compatibility”; and “remote accessibility” were not selected as important criteria for electronic material at MPSJ library.

Table 8. Criteria for Selecting Electronic Materials

Criteria	IIUM	MMU	MPPJ	MPSJ
Quality	yes	yes	no	yes
Subject matter	yes	yes	yes	yes
Currency, authority, completeness	yes	yes	yes	yes
Language	yes	yes	yes	yes
Uniqueness of content, capabilities of features	yes	yes	no	no
Number of access points/indexes	yes	yes	yes	no
Geographic parameters	yes	yes	no	no
Relevance of material for users	yes	yes	yes	yes
Relevance of material for reference	yes	yes	yes	yes
Cost	yes	yes	yes	yes
Network capability	yes	yes	yes	no
User friendly	yes	yes	yes	yes
Strength of retrieval	yes	yes	yes	no
Hardware compatibility	yes	yes	no	no
Software compatibility	yes	yes	yes	no
Service implications	yes	yes	no	no
Remote accessibility	yes	yes	yes	no
License restrictions	yes	no	yes	no

Tools for Selecting Electronic Items

The participating libraries reported mixed responses in choosing tools used for selecting electronic items. Of the 13 tools listed in the checklist (Table 9), IIUM uses 8 (61.5%), MMU 3 (23.0%), and the MPPJ and MPSJ 6 (46.1%). All these libraries were found utilizing “Barnesandnoble.com”, while none of them reported the use of “EBSCO CD-ROMs”, “Lexis/Nexis”, “Borders.com”, and “Tower records.com”. Three libraries, MMU, MPPJ and MPSJ, were found not using “LC” and “Global Books in print”. Furthermore, the “Online Catalog WWW”, “OCLC” and “Sites of other libraries” were not used by MMU. The MPPJ was the only library found not using “Amazon.com”, similarly, the MPSJ also did not use “EBSCO.Online” for selecting electronic items.

Table 9. Tools for Selecting Electronic Materials

Tool	IIUM	MMU	MPPJ	MPSJ
Online catalog WWW	yes	no	yes	yes
Listservs	no	no	yes	yes
Amazon.com	yes	yes	no	yes
EBSCO. Online	yes	yes	yes	no
EBSCO CD-ROMs	no	no	no	no
LC	yes	no	no	no
OCLC	yes	no	yes	yes
Global Books in print	yes	no	no	no
Lexis/Nexis	no	no	no	no
Sites of other libraries	yes	no	yes	yes
Barnesandnoble.com	yes	yes	yes	yes
Borders.com	no	no	no	no
Tower records.com	no	no	no	no

Table 10. Criteria for selecting Printed Materials

Criteria	IIUM	MMU	MPPJ	MPSJ
Quality	yes	yes	no	yes
Cost	yes	yes	no	yes
Copyright	yes	yes	no	no
Publisher	yes	yes	yes	no
Edition	yes	yes	yes	no
Subject matter	yes	yes	yes	yes
Currency, authority, completeness	yes	yes	yes	yes
Language	yes	yes	yes	yes
Relevance of material for users	yes	yes	yes	yes
Relevance of material for reference	yes	yes	yes	yes
Service implications	yes	yes	yes	yes
Remote accessibility	yes	yes	yes	yes
License restrictions	yes	yes	yes	yes

Criteria for Selecting Printed Materials

The participating libraries indicated criteria they believe important for selecting printed materials. A checklist comprising 13 characteristics (Table 10) were listed for the purpose. The entire criteria were measured important to academic libraries, IIUM and MMU. However, public libraries, MPPJ and MPSJ, did not regard some criteria important for the selection of printed materials. For instance, both libraries agreed that "copyright" was not important for selecting printed materials. Moreover, "quality", and "cost" were well thought-out unimportant to MPPJ library, while MPSJ did not consider "publisher" and "edition" among important criteria.

Tools for Selecting Printed Materials

Of the 16 tools listed in Table 11, the IIUM reported the use of 13 (81.2%) tools for selecting printed materials, the MPSJ 12 (75.0%), and MMU and MPPJ 8(50.0%) for each. All the participating libraries were found utilizing “Any article that mentions or describes a book”, “advertisements and flyers”, and “newsletters and bulletins”. None of them reported the use of “circulation statistics”. Three libraries, IIUM, MMU and MPPJ were found not using “CD-ROM in print”, “reviewing periodical”, and “reference questions”. Furthermore, “standard selection lists” and “Baker & Taylor, Brodart, Blackwell” were not used by public libraries, MPPJ and MPSJ. The MMU was the only library found not using the first four tools (i.e. “best sellers”, “award winners”, “newspaper book review sections” and “any journal that reviews books”).

Table 11. Tools Utilized for Selecting Printed Materials

Tool	IIUM	MMU	MPPJ	MPSJ
Best sellers	yes	no	yes	yes
Award winners	yes	no	yes	yes
Newspaper book review sections	yes	no	yes	yes
Any journal that reviews books	yes	no	yes	yes
Any article that mentions or describes a book	yes	yes	yes	yes
Standard selection lists	yes	yes	no	no
Baker & Taylor, Brodart, Blackwell	yes	yes	no	no
Advertisements and flyers	yes	yes	yes	yes
Books-in-print	yes	yes	yes	no
Catalogs	yes	yes	no	yes
CD-ROM in print	no	no	no	yes
Newsletters and bulletins	yes	yes	yes	yes
Reviewing periodicals	no	no	no	yes
Circulation statistics	no	no	no	no
Reference questions	no	no	no	yes
Patrons requests	yes	yes	no	yes

Format of Digital Collection

Table 12 shows available format of digital collection at the participating libraries. Out of eight formats listed in the Table, six (75.0%) were provided by MMU, while IIUM, MPPJ and MPSJ reported the availability of five (62.5%). On the other hand, all the surveyed libraries stated the availability of “text” and “multimedia” formats; however, none of them provided “three dimensional files” and “archival finding aids”. Likewise, the IIUM, MPPJ and MPSJ did not provide “moving image” format, while the MPPJ was the only participating library found not providing “print & photographs” format.

Table 12. Format of Digital Collection

Format	IIUM	MMU	MPPJ	MPSJ
Text	yes	yes	yes	yes
Image	yes	yes	yes	yes
Sound	yes	yes	yes	yes
Moving Image	no	yes	no	no
Print & Photographs	yes	yes	no	yes
Multimedia (text, image, sound, graphic, animation)	yes	yes	yes	yes
Three Dimensional Files	no	no	no	no
Archival Finding Aids	no	no	no	no

Magnitude of Interlibrary Lending

Analysis of the data showed that MMU library made or received 56 interlibrary lending (ILL) requests over a period of five years, 1997-2001, with a success rate of 100 percent. IIUM library requested 1745 and received 924 items during the same period with a success rate of 52.9% (Table 13). Public libraries, MPPJ and MPSJ, did not make or receive any interlibrary lending request in that period. It is not clear why these two libraries are not providing this service to users.

Table 13. Request Made for ILL (1997-2001)

Library	Request made	Document received	Success rate
IIUM	1745	924	52.9%
MMU	26	30	100%
MPPJ	N/A	N/A	N/A
MPSJ	N/A	N/A	N/A

Libraries Contacted for ILL

The selected libraries were also asked to indicate libraries they contacted for ILL requests. It was noted that academic libraries in Malaysia (mean score = 1.2) or overseas libraries (mean score = 1.0) were the most preferred sources for this purpose. The participating libraries also frequently made ILL requests to national libraries (mean score = 0.7) and public libraries (mean score = 0.2) in Malaysia (Table 14). Among the overseas libraries contacted for ILL are the British Library, the National University of Singapore and Nanyang Technological University in Singapore.

Table 14. Libraries Contacted for ILL

Library	Mean score
Academic libraries in Malaysia	1.2
National libraries in Malaysia	0.7
Public libraries in Malaysia	0.2
Libraries outside of Malaysia	1.0

Notes: scale: 0 = Not at all; 1 = Les Frequently; 2 = Frequently; 3 = Very Frequently
N = 4 libraries

5 Conclusion

It is clear that the process of transforming Malaysian libraries into digital libraries is painfully slow. If our classification of a digital library is followed (that is, at least 60% of its total collection is in electronic/digital form), the four libraries in this study would be a long way from reaching digital library status. Even the MPSJ library, which has been classed as a digital library by other researchers, may have to be reclassified. One of the main causes for the sluggishness in the development of digital libraries is the lack of financial support for electronic materials from governing authorities. For example, MMU library spent only 29% of its annual budget on electronic materials in 2001 and MPSJ library only 5%. At this rate, it will take a long time for these libraries to become digital. Coupled with the lack of financial support is the high cost of acquiring electronic materials. Analysis of the data showed that libraries have to spend at least twice as much for an electronic item as for a printed item. For example, in the case of the MMU library, the unit cost for electronic materials is around RM300 while a printed item costs only RM150. This problem underscores the need to allocate huge sums to build up electronic resources.

The findings also revealed that tools and criteria used for selecting electronic materials are different from those used for printed materials. It is useful to point out that relevancy of tools and criteria used is somewhat influenced by the typologies of the library. For example, of the 18 criteria listed for selecting electronic materials, the public libraries use thirteen or fewer while the academic libraries accepted all 18 criteria.

Interlibrary lending is an important means to provide users with materials not available in the library's collection. In the case of IIUM library, ILL is practiced on quite a large scale. 1745 items were requested under ILL services during the study period. However, there is no evidence that the investigated libraries act in concert to fill up the lacunae in their collection. There is no notion of banding together in the form of a consortium. Such a scenario is not in keeping with the rising expectations of users and the technological capabilities needed to service those needs. We therefore suggest that the investigated libraries consider the question of establishing an electronic consortium. Most libraries have begun to realize that membership in a consortium "is the only anchor for entry into the choppy waters of database licensing of exorbitant electronic products" [7]. For the good of the community and nation, Malaysian libraries have to strive hard to improve accessibility and diversity in the digital age. They can do it provided they possess affordable technology, the power to negotiate through a cartel, and unwavering support from the authorities. Otherwise, they are liable to fall by the wayside, as traditional as they are now.

References

1. Brody, Fern. *Planning for the balance between print and electronic journals in the hybrid digital library; lessons learned from large ARL libraries,* Ph. D. Dissertation, University of Pittsburgh, 2001, pp. 79.
2. Jewell, Timothy. *Selection and preservation of commercially available electronic resources : issues and practices*, Washington : Digital Library and Council on Library and Information Resources, 2001, pp. 12.
3. Muller, Rush G. Shaping digital library content, *Journal of Academic Librarianship, 28*, 3 (2002), pp. 97-103.
4. Lesk, Michael. *Practical digital libraries : books, bytes and bucks*, San Francisco; Morgan Kaufmann, 1997, pp. 2.
5. Harloe, B and Budd, J. Collection development and scholarly communication in the era of electronic access, *Journal of Academic Librarianship, 20* (1994), 2, 83-7.
6. Baker, A. The impact of consortia on database licensing, *Computers in Libraries* 2000, pp. 46 - 50.
7. Hiremath, Uma. Eletronic consortia : resource sharing in the digital age, *Collection Building , 20* (2), 2001, pp. 80 – 88.

Digital Divide: How Can Digital Libraries Bridge the Gap?

Gobinda G. Chowdhury

Graduate School of Informatics
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, UK
gobinda@dis.strath.ac.uk

Abstract. Recent developments in Information and Communication Technologies (ICT) have, while making our life easier, created a social divide known as the digital divide. Statistics show that there are significant disparities among the populations in the developed and developing world in terms of accessibility to, and use of, ICT. Research and development in digital libraries do not only require sophisticated ICT, they also call for huge investments in terms of money and intellectual resources. Developing countries are lagging behind in digital library research and development, due to the digital divide as well as the lack of appropriate resources required for research and development. As a result, users in the developing world are being deprived of digital library services. This paper argues that some recent global digital library developments can be used by users in the developing countries : subject gateways, digital reference services, free access to e-journals and e-books in many areas, e-print archives and free digital libraries. The paper ends with an action plan that may be used by library and information professionals in both developing and developed countries to exploit the benefits of these digital information resources and services. This will to some extent help to bridge the digital divide.

1 Introduction

While the information and communication technologies (ICT) in general, and the Internet and the world wide web in particular, have made life easier by facilitating easy communication with virtually everyone, and easy access to information located virtually anywhere in the world, they have also widened the gap between the rich and the poor, the 'have' and the 'have nots'. In other words, new technologies, while improving our life in many ways have created what is called the 'digital divide'. The digital divide has become a popular phrase to describe the perceived disadvantages of those who are either unable, or do not choose, to use the appropriate ICT in their day-to-day activities, decision making, learning and pleasure [1]. It is caused by a number of divides, such as the infrastructure, hardware, running costs, manpower, and information and digital literacy.

Digital libraries make use of ICT and the web to provide access to the local and remote digital information sources and services. Therefore, accessibility to the basic

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 379-391, 2002.

ICT and the Internet is a pre-requisite to the development and use of digital libraries. There are many other factors too. The two most important issues are (1) the cost of building and maintaining sustainable digital library systems and services, and (2) achieving the required information literacy standards so as to exploit the full benefits of digital libraries. This paper discusses the problems facing the developing countries in these two areas the consequence of which is the widening of the gaps between the developed and the developing countries in terms of accessibility to digital information sources and services. This paper argues that some recent global digital library developments can be used by users in the developing countries, and thus digital libraries can play a significant role in bridging the gap. Finally the paper shows how digital library services can be set up with minimum initial investments in the developing counties in order to make some of the global digital library sources and services available to the users.

2 Digital Divide: Some Facts and Figures

The Digital Divide Basics Fact Sheet [2] shows that there are an estimated 429 million people (only 6% of the world's entire population) online globally, with the following distribution:

- 41% of the global online population is in the United States and Canada
- 27% of the online population lives in Europe, the Middle East and Africa
- 20% of the online population logs on from Asia Pacific
- Only 4% of the world's online population are in South America.

Digital divide, however, is not necessarily a developing country phenomenon. In Britain, it is estimated that more than 60 percent of the richest ten percent of the population have household access to the Internet, whereas among the poorest 10%, only 6% have household access to the Internet [3]. In the Fall of 2000, the US Department of Commerce found that only 41.5% of all US homes had Internet access [4].

However, disparities in the least developed countries are mind-boggling; most people don't even have a phone, let alone an online connection, either at work or home. According to BBC [5], 'more than 80% of people in the world have never even heard a dial tone, let alone surfed the Web'. United Nations Secretary General Kofi Annan recently said that the Internet is used only by five percent of the world's population [6].

3 Money Spent on Digital Library Research and Development

Huge amount of money has been spent in digital library research and development. In the United States, over US$ 24 million was awarded in 1994 as part of the first Digital Library Initiative (DLI-1), and over US$55 million has been allocated so far in the second phase which will be much more if the funding for the recent joint international projects is counted. There have been several other research projects on digital libraries outside the DLI funding in the United States (for details see [7]). Research on digital libraries in the United Kingdom has largely been funded by the Electronic Libraries (*eLib*) Programme. The three phases of eLib had a costing in excess of £20 million [8].

These figures are just the tip of the iceberg as far as the total amount of money spent for digital library research and development is concerned. A large number of digital library projects have been funded by libraries, universities and other institutions the total cost of which is difficult to calculate. Harvard University has a budget of $12 million over a period of five years in the Library Digital Initiative [9]. University of Central England, UK has a capital allocation of over £1 million for a period of three years for the digital library systems and services [10]. This are just some figures to give an idea of the amount spent at the institutional level. A large number of similar digital library projects have been undertaken in Europe and in other parts of the world.

Developing countries, especially the least developed countries, that struggle to meet the basic human needs, cannot afford to spend such huge amount of money required for research and development of digital libraries. In addition, there are many other problems that stand in the way of digital library development in the developing countries.

4 Problems of Digital Library Development and Use in Developing Countries

Library development has not been a priority of governments in the developing countries. Governments struggle to meet the basic human needs like food, water, health, electricity, sanitation, transportation, etc. Consequently, libraries have long been suffering from financial and other crises such as lack of the appropriate technology, trained manpower, etc. Libraries have also been affected by a number of social problems, the primary ones being the poor literacy rates. While governments are struggling to improve the levels of basic literacy, proper use of library and information services call for another level of literacy – the information literacy that is absolutely necessary for people to become good information users. Due to the lack of suitable technologies and trained manpower, and above all due to the lack of financial resources, most libraries in the developing countries do not even have fully developed and up to date OPACs, let alone full-fledged automated library management systems, and digital libraries. Hence compared with the developed world scenario, libraries in the developing countries are already left behind by at least one generation. Digital divide, and lack of resources for digital library research and development, may increase the gap far more significantly between library and information services in developed and developing countries.

The following is a short list of problems or issues that stand in the way of digital library research and development in the developing countries:

- Shrinking library budget that forces the library management to struggle to maintain a minimum standard of services leaving no room for new ventures and developments
- Lack of financial support specifically for digital library research and development

- Absence of fully developed and up to date OPACs, and little access to online information resources – online databases, e-journals, etc.
- Poor ICT – computers and networks
- Poor facilities for access to ICT, especially the Internet
- Stringent government and institutional policies on Internet access
- Lack of trained manpower
- Poor information literacy rate that causes lack of appreciation of modern information services and their use.

The list may go on and on. Any experienced library manager from a developing country can surely add quite a few more points to the above list. In short, existing libraries in the developing countries are struggling for their mere existence. Of course there are many reasons for the lack of resources for library development. In countries where citizens still struggle for reliable sources of food, water, medical care and educational opportunities, bridging the digital divide may seem like a lofty goal, and that’s is why digital library development is way down the list of priorities of governments and institutions.

What is the solution then? Should the library and information professionals sit behind and watch their developed counterparts embrace new technologies and excel in the provision of information services? Should the majority of the world population, who live in the developing and least developed countries, become information poor day by day? Or are there any hopes?

This paper highlights some recent developments in the library and information world that may be quite encouraging for information professionals and users in the developing countries. In fact, over the past few years many new services have appeared that can be used by anyone, anywhere and are free at the point of use. Many of these new services are the consequences of digital library research, while others are the results of new economic models of the information industry, and some are the results of good gestures to help people. The rest of this paper highlights some of these developments, and proposes some measures to be taken by the library and information professionals in the developing countries, or even those in the developed countries, to reap the benefits of the global digital library research and development activities.

5 New Services and Facilities

The following sections briefly mention the features of some new services and facilities that can be used by anyone for free. The list is by no means exhaustive, and one can easily find many more such services and facilities. Nevertheless, the following sections show that library and information professionals who do not have access to huge budgets and other resources to venture on digital libraries, can still provide good quality services to their users at almost no additional cost. Of course, the basic IT infrastructure and Internet access are the pre-requisites to these services, but the cost to this is negligible compared to the contents, and, above all, the benefits that the end users may get by availing these services.

5.1 Information from Government, Regional, and International Organizations

One of the direct impacts of the Internet on the governments and the regional and international organizations has been that they are now trying to make as much information available on the net as possible. As a result, end users can get access to the up to date (as much as possible) information for which, even a few years ago, they had to wait for long and had to go through a number of hassles. Asian governments are well represented on the web – in most cases detailed information about the government ministries, departments, policies, activities, publications, etc., are available on the web. The e-ASEAN Task Force was created in 1999 by the Association of Southeast Asian Nations (ASEAN) to develop a broad and comprehensive action plan for an *ASEAN e-space* with an aim to give directives to the ASEAN governments for establishing an ASEAN Information Infrastructure (AII) [11].

Regional and international organizations also make useful information about them – organization, activities, publications, etc.— available on the web. In addition, thousands of NGOs (Non-governmental organizations) have set up their website in order to provide access to a number of useful information resources about government, and various socio-political and economic issues.

Information and library professionals in the developing countries may provide access to all of these, or preferably to selected and the most appropriate, web information resources to their users by creating a simple webpage providing links to the various resources. A better and more useful approach will be to organise the resources into various categories according to their content, sources, and/or the user requirements. Users can easily navigate through such an organised structure of information sources.

5.2 Information through Subject Gateways and Virtual Libraries

One of the most prominent, and useful especially from the end-user perspectives, outcomes of the recent digital library research has been the development of a number of subject gateways. These gateways select and organise valuable subject-specific information resources available on the web, and let the user access to those resources through a custom-built interface. The following are some examples of subject gateways:

- Art and Architecture: ADAM (http://www.adam.ac.uk)
- General: NOVAGate (http://novagate.nova-university.org), BUBL Link (http://link.bubl.ac.uk/)
- Engineering: EEVL (http://www.eevl.ac.uk/) and EELS (http://eels.lub.lu.se/)
- Business and economics: Biz/ed (http://bized.ac.uk)
- Health and medicine: OMNI (http://www.omni.ac.uk/)
- Social science: SOSIG (http://www.sosig.ac.uk/)

A detailed list of various subject gateways and virtual libraries appears in Chowdhury and Chowdhury [12]. Library and information professionals can make one or more of these subject gateways accessible to their users by pointing to the appropriate service websites from then own webpage. Alternatively, they may point the users to a general subject gateway like BUBL that would allow the user to browse or search the web information resources by subject or discipline.

5.3 Digital Reference and Information Services

A number of reference and information services are now available on the web. Interestingly, many of these services are provided by non-library and commercial organisations. While some of these services are free, others need the users to pay. Detailed discussions on such services are available in a number of recent publications (see for example, [12-15]). McKiernan [16] maintains a site that provides categorised listing of libraries that offer real-time reference services using chat software, live interactive communication tools, call centre management software, bulletin board services and other Internet technologies. Most of these services are designed for registered users of the libraries concerned.

Table 1 provides a quick overview of some online reference and information services that are currently available. This is not an exhaustive list, but the table shows the different types of services that are now available, and also their major characteristics. In addition to those mentioned in Table 1, there are also some web-based reference services where users need to conduct search with a reference query. Such services provide free access to various online reference sources, and allow users to either select a specific source or conduct a search on a range, or all, of the reference sources. Examples of such services include the following:

- Internet Public Library (http://www.ipl.org)
- Infoplease (http://www.infoplease.com)
- Britannica (http://www.britannica.com)
- Bartleby reference (http://www.bartleby.com/reference)
- Internet Library for Librarians (http://www.itcompany.com/inforetriever/)
- Electric Library (http://ask.elibrary.com/refdesk.asp)
- Mediaeater Reference Desk (http://www.mediaeater.com/easy-access/ref.html)
- ReferenceDesk (http://www.referencedesk.org/)
- Xrefer (http://www.xrefer.com/)

While most of these web-based reference services are available for free, some charge small amount of fees. For example, the Electric Library charges $79.95 for an entire year of unlimited access.

Janes, Hill and Rolfe[17] report a study of 20 web-based ‘expert services’. By asking 240 questions to 20 selected expert services, they noted that the sites gave verifiable answers to 69% of factual questions. The high rate of success of factual questions shows that these expert services can be used to find answers to simple (ready reference) type of questions.

While the digital reference and information services are designed specifically for the end-users, library and information professionals may tap on them to provide services to their users. This is particularly true for the developing countries where the end users may not have access to the Internet from their home or office, and where the access is rather expensive. LIS (library and information science) professionals may select one or more web-based reference services according to the nature and need of their users, and either can use those services on behalf of the users or can let the users use on their own.

Table 1. Characteristics of Some Web-based Reference Services

Service	**Subject**	**Payment**	**Service Providers**	**Question Input**	**Mode of Delivery**
Askme (Askme.com)	All	Free	Volunteer Experts	Select a subcategory and Input query through a Web-based query form	e-mail
AllExperts (Allexperts.com)	All	Free	Volunteer Experts	Select a subcategory and enter query through a Web-based query form	e-mail
Live Advice (Liveadvice.com)	All	Fee-based (Each advisor sets a per-minute rate for Phone and Recorded Advice)	Registered Experts	Select a subcategory and enter query through a Web-based query form	e-mail
Find/SVP (Findsvp.com)	Business	Fee-based (Users can choose a cost band)	Business Experts	Enter query through a Web-based query form	e-mail, phone, fax, courier
Professional City (Professional-City.com)	Law, Accounting Marketing	Fee-based	LIS Professionals	Enter query through a Web-based query form	e-mail

5.4 Access to Electronic Texts – Books, Journals, Theses, etc.

One of the most prominent benefits of digital libraries is that users can get online access to books, journals and other publications such as conference proceedings, theses, etc. Indeed many digital libraries and other services have been set up in the recent past that provide free access to a number of electronic books, journals, theses, etc. The following sections provide some examples free access to e-journals, ebooks, theses, etc.

e-journals. While most of them are accessible only through payment, some e-journals and books are available for free. Some publishers and associations/organisations are now making journals available free to the readers in some countries. For example

- Blackwells (www.blackwells.co.uk) is making all 600 of its journals freely available to institutions within the Russian Federation [18]

- World Health Organisation (WHO; http://www.who.int) is spearheading an initiative to enable 100 of the world's poorest countries to access 1000 of their top biomedical journals
- Academic Press's Ideal service (www.idealibrary.com) is making 300 science, technology and medicine journals available to research centres across Senegal in west Africa.
- PubMed Central (http://www.pubmedcentral.nih.gov/is) is a digital archive of life sciences journal literature managed by the National Center for Biotechnology Information (NCBI) at the U.S. National Library of Medicine (NLM). It is free for use from anywhere in the world.

A number of free e-journals are now coming up in different subjects. For example, in information science some of the very good and free e-journals include: the D-Lib Magazine, Ariadne, Information Research, and so on. Many such free e-journals are also available in other subjects. One may be sceptical about the qualities of these free e-journals. However, a study by Fosmire and Yu [19] shows that one need not worry about this. They conducted a survey of 1,209 e-journals and noted that 213 (18%) of scholarly journals were free. They further noted that these journals have very high impact factors (a measure used to assess the quality of a journal), and each of them has a reasonable numbers of articles published.

e-print Archives and the Open Archives Initiative. e-prints are seen as a means to fighting the high costs of scholarly publications from publishers. The idea is that the fastest and cheapest way for authors to make their papers available is to store the electronic copies of their papers on e-print servers [20]. Success and rapid growth of arXiv e-print server (http://www.arxiv.org) has given birth to many new e-print services, such as CogPrints (the Cognitive Sciences E-print Archive) (http://cogprints.soton.ac.uk/), front end for the mathematics arXiv (http://front.math.ucdavis.edu/), and WoPeC (Working Papers in Economics) (http://netec.mcc.ac.uk/WoPEc.html). These e-print services can be excellent sources for authentic and up to date research information for users anywhere in the world.

ebooks. Although electronic books are not yet as common as e-journals, ebook service provided by the netlibrary (http://www.netlibrary.org) is growing very fast. While the netlibrary requires payment for use, many electronic books are now available for free. Some examples are given below:

- The Dictionary.com (http://www.dictionary.com/) site provides access to a number of dictionaries, thesauri, writing resources, and other tools including the automatic webpage translation services
- The eLibrary reference desk (http://ask.elibrary.com/refsearch.asp) provides access to a number of dictionaries, encyclopaedias and almanacs
- The Classic Book Shop (http://www.classicbookshelf.com/SiteMap.htm) provides access to a number of classic books available in electronic format
- The webbooks.com site (http://www.web-books.com/cool/ebooks/Library.htm) provides access to over 2200 electronic books in different subjects.

Theses, Dissertations, etc. The two most prominent digital libraries that provide access to electronic theses, dissertations and scholarly publications are NDLTD (Networked Digital Library of Theses and Dissertations; http://www.NDLTD.org) and NCSTRL (Networked Computer Science Technical Reference Library; http://www.ncstrl.org). These services aim to increase the availability research information to scholars, academics and students. They have grown dramatically over the first few years. NDLTD is now becoming a global access point to electronic theses and dissertations. At the time of writing (May 2002) the site provides access to the theses and dissertations from 138 members – 122 universities from around the world and 16 internationally renowned institutions.

5.5 Other Free Digital Libraries

Many digital libraries provide free access to a variety of digital information resources. The Greenstone Digital Library (GDL; formerly New Zealand Digital Library, NZDL) in New Zealand is a free digital library service that may be particularly valuable for users in the developing countries. Witten et al [21] list five specific areas where digital libraries can promote developments in the developing countries:

1. In the dissemination of humanitarian information
2. in facilitating disaster relief by providing the appropriate information
3. in the preservation and propagation of indigenous culture
4. in building collections of locally-produced information, and
5. in creating new opportunities to enter the global marketplace.

To this we can add another important point that can be applicable to any digital library: digital libraries can facilitate life long learning which is the key to success in this fast changing world.

While discussing the collection and services of the Greenstone Digital Library, and arguing how the digital library can meet many of the above objectives in developing countries, Witten et al. [21] comment that digital libraries provide a golden opportunity to reverse the negative impact of ICT on developing countries.

6 Summary and Some Action Plans

While developing countries are threatened by the growing digital divide, and their information professionals and users feel that they are lagging behind the digital library revolution due to the high cost of digital library research and development, as well as other factors such as the lack of technology, manpower, and other resources, this paper has shown that many new developments are taking place that may reverse the situation if they are used properly. This paper has pointed out several new services and facilities that are particularly suitable for users in the developing countries. However, the pre-requisite to these services is the availability of basic ICT and Internet facilities. Since this is one of the major problems in the developing world and many users may not have reliable and affordable Internet access from home, or even from work, library and information centres can play a very important role. The following are some simple guidelines that may enable library and information professionals in developing countries to make some of the digital library services available to their users without much cost or hassle. The main activities should include the following.

6.1 Building and Linking Local Digital Libraries

Building digital libraries of local and indigenous materials, is an important step in bridging the digital divide. Many such digital libraries are now being built in the developing countries. Some Asian and African countries, for example Hong Kong, Singapore, Malaysia, India, South Africa, etc., are ahead of others, but other countries are following suit. A recent example is the development of the Ganesha Digital Library (GDL) Network in Indonesia [22].

Many of the projects on building digital libraries of indigenous resources have been initiated and funded by institutions and governments locally, while some have been developed though international collaboration. The *Million Books Project*, funded by the US National Science Foundation and the governments of China and India is an example of building a digital library of indigenous materials through international collaboration. The project began in 2001 at Carnegie Mellon University, in collaboration with six Chinese universities and ten research centres in India, with an aim to provide full-text searching to one million books, making them accessible to anyone anywhere. The project has received the support of the National Science Foundation (NSF) and the governments of India and China [23].

6.2 Digital Outsourcing

Some free services are mentioned in this paper. However there are many more. Information professionals in the developing countries should spend time on outsourcing of free digital information sources and services. The task of selection should include a number of activities including (a) identification of the appropriate sources and services based on the subject, sources/authority, user requirements, etc. (2) evaluation of the sources in order to assess the suitability of the selected sources and services in the light of the user requirements vis-à-vis the technical requirements to access and use them, and (3) to create some sort of surrogate for each source and service to facilitate organization, etc.

6.3 Organisation of Digital Information Sources and Services

This may require basic web design skills. Simple web design skills may be acquired easily, and a number of free courses and guides for web design are available on the web. In the absence of anything else, the editor that come with the web browsers (Netscape Navigator, or Internet Explorer) may be used to design simple web pages. The major professional skills will be required in the organisation of the identified digital resources and services. An understanding of the users and their information needs vis-à-vis the content, format, etc., of the selected sources and services may help the information professionals organise them properly. Appropriate tools used for information organisation in traditional libraries (classification schemes, thesauri, etc.) may be used for the purpose.

6.4 Using Freely Available Digital Library Software and Support

Free software and support are available from a number of international digital library research groups, such as NDLTD, GDL, etc., which may be used for building local digital libraries.

6.5 Improving Information Use

Information use rather than access is a major problem in many developing countries. Paul (2002) comments that ‘a major issue in debating the digital divide in ASEAN countries centres more on usage than on wired access or computer ownership.’ There are many reasons for poor information usage despite having good access. One of the major reasons is poor information literacy (discussed below). The other most important reason is the work culture and habits. In many countries, more so in the developing world, the work culture does not allow people to spend more time on the Internet, and the day-to-day activities are based more on the traditional approach through the use of paper documents and telephone or written communications.

6.6 Improving Information and Digital Literacy Skills

Poor information and digital literacy is a major problem in the developing countries. Widharto [24] while discussing the problems facing information services in Indonesia comments that training remains a key to the future of the Indonesian libraries. This statement can be generalised for other developing countries too. Information or digital literacy training may be organised at different levels. Because of the limitation of resources, information professionals may begin with a simple approach of providing training to the users at different levels – basic, advanced, etc. Nevertheless, to keep pace with the rapid changes in ICT and digital library systems and services, such training should be provided on a regular basis in order to help the users keep up to date and thereby make the optimum use of the sources and services made available to them.

7 Conclusion

The digital divide is not only a problem of the developing countries; within the developed countries there are significant proportions of the population for whom the digital divide is as prominent as it is between the north and the south [25]. Nevertheless, as Ross Shimmon [25,26], the Secretary General of IFLA, comments, and this paper has justified, library and information professionals, even with their limited resources, can play a significant role to bridge the digital divide.

While library and information professionals in a less fortunate situation can play a great role in making use of the recent digital library developments to the benefit of their users, there are some deeper issues too. The digital divide can only be reduced when the users actually make use of the information for the purpose of making informed decisions, and in every aspect of their daily lives. Paul [27] comments that the digital divide can be bridged by improved:

- access, measured by access indicators
- usage, measured by usage indicators, and
- outcome, measured by impact indicators.

Nevertheless, these are difficult parameters to measure, and painstaking research is needed to develop the measuring yardsticks and best practice standards.

References

1. Cullen, R.: Addressing the digital divide. Online Information Review. 25(2001) 311-320.
2. Digital Divide Basics Fact Sheet http://www.digitaldividenetwork.org/content/stories/index.cfm?key=168
3. Cronin, B.: The digital divide. Library Journal. 127(2002). 148.
4. Digital Divide Network. http://www.digitaldividenetwork.org
5. Information Rich Information Poor (1999). BBC News. Oct. 14, 1999. http://news.bbc.co.uk/hi/english/special_report/1999/10/99/information_rich_information_poor/newsid_466000/466651.stm
6. Conhaim,W. W.: The global digital divide. *Information Today*. 18(2001), http://proquest.umi.com
7. Chowdhury , G.G. and Chowdhury, S.: Introduction to digital libraries. Facet Publishing, London (2002).
8. Rusbridge, C.: After eLib. *Ariadne,* Issue 26 (2001) http://www.ariadne.ac.uk /issue26/chris/intro.htm
9. Harvard University Library. Digital Library Initiative. http://hul.harvard.edu/ldi/html/costs
10. About UCEEL. http://diglib.uce.ac.uk/webgate/dlib/templates/about.asp
11. e-Task Force. http://www.e-aseantf.org/
12. Chowdhury , G.G. and Chowdhury, S.: Information sources and searching on the world wide web. Library Association Publishing, London (2001).
13. Chowdhury, G.G.: Digital libraries and reference services: present and future. Journal of Documentation. 58(2002), 258-283.
14. Lankes, D., Collins, J.W. & Kasowitz, A.S. (eds): Digital reference service in the new millennium: planning, management, and evaluation, Neal-Schuman (2000).
15. Sherman, C.: Reference resources on the Web. Online, 24(2000), 52—56.
16. McKiernan, G.: LiveRef(Sm): a registry of real-time digital reference services. http://www.public.iastate.edu/~CYBERSTACKS/LiveRef.htm
17. Janes, J; Hill, C., Rolfe, A.: Ask-an-expert services analysis. Journal of the American Society for Information Science and Technology. 52 (2001), 1106-21.
18. Smith, G. Closing the digital divide. Information World Review. (2001). Online. Downloaded from http://proquest.umi.com
19. Fosmire., M. and Yu, Song: Free Scholarly Electronic Journals: How Good Are They? Issues in Science and Technology Librarianship, Summer 2000. Online: http://www.library.ucsb.edu/istl/00-summer/refereed.html
20. Day, M.: E-print Services and Long-term Access to the Record of Scholarly and Scientific Research. Ariadne. Issue 28 (2001). http://www.ariadne.ac.uk /issue28/metadata/intro.html
21. Witten, I. H., Loots, M., Trujillo, M.F., and Bainbridge, D. : The promise of digital libraries in developing countries. The Electronic Library. 20(2002), 7-13.

22. Fahmi, I.: The Indonesian Digital Library Network is born to struggle with Digital Divide. Bulletin of the American Society for Information Science. 28(2002).
23. Michalek., G.: The Universal Library and the Million Book Project. D-Lib Magazine. 8(2002), http://www.dlib.org/dlib/june02/06inbrief.html#MICHALEK
24. Widharto: Challenges in accessing scientific and technological information in Indonesia during the economic crisis. Bulletin of the American Society for Information Science and Technology. 28(2002), 25-27.
25. Shimmon, R.: From digital divide to digital opportunity. (2001). Online. http://www.unesco.org/webworld/points_of_view/shimmon.html
26. Shimmon, R. Can we bridge the digital divide? Library Association Record. 103(2001), 678-679.
27. Paul, J.: Narrowing the digital divide: initiatives undertaken by the Association of South-East Asian Nations (ASEAN). Program, 36(2002), 13-22.

Digital Libraries in Academia: Challenges and Changes

Anne Adams[1] and Ann Blandford[2]

[1] Research institute for digital libraries (RIDL), Middlesex University, Trent Park,
London, UK. N14 4YZ
a.adams@mdx.ac.uk
http://www.cs.mdx.ac.uk/RIDL/aadams/default.html

[2] UCL interaction centre, 26 Bedford way,
London, UK. WC1H 0AB
a.blandford@ucl.ac.uk
http://www.uclic.ucl.ac.uk/annb/

Abstract. Although web accessible digital libraries (DLs) have greatly increased potential information accessibility within academia, the use of these resources varies widely across disciplines. This study, within contrasting departments (Humanities, Computing and Business) of a London university, reviews the social and organisational impacts of DLs across these disciplines. In-depth interviews and focus groups were used to gather data from 25 lecturers and librarians, and results analysed using the grounded theory method. Web-accessible DLs are identified as changing the roles and working patterns of academic staff (i.e. lecturers, librarians and computer support staff). However, poor accessibility due to inappropriate implementation strategies, access mechanisms, searching support & DL usability reduces the use of these resources. Consequently, web and personal collections without guarantees of quality are widely used as an accessible alternative. One conclusion is the importance of implementation strategies (e.g. giving feedback on document context, collection boundaries, ownership, accountability and support) in informing DL design.

1 Introduction

In the past, academic libraries were totally bound by their physical parameters. Library users initiated the interactions by going to the library. They physically walked around the library and searched or browsed for information, or asked a librarian for help. When successful, they read the hard copy information or took it out of the library. This model supported a wide variety of users from many different disciplines. With the introduction of library technology, those physical boundaries have slowly changed. The searching was done via microfiche or CDROMs then information was either photocopied or read in the library online (e.g. via CDROMs or Library IP based computers).

With the advent of web-accessible digital libraries and remote authentication (e.g. Athens password), users' physical interaction with the library could completely

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 392-403, 2002.

change. Digital libraries (DLs) have the potential to transform aspects of the education process, with remote access to specialized information in a format that is easily updated, and speedy searching and access facilities. However, the invisible presence of these resources, their poor usability and user support, has made their impact less dramatic [6, 16]. A key element in the successful design and implementation of digital libraries has, in the past, been identified as their social context [7,8,9]. The social contexts of organisational systems can have an important impact on the community's involvement in resulting technology systems [10]. The role of the librarian and the changing impact of DLs across all the academic disciplines, although crucial, have not been fully researched. This paper presents the findings from an in-depth analysis of lecturer and librarians' perceptions within contrasting disciplines and the impact of digital libraries within those social contexts. The resulting design implications for digital libraries are also presented.

1.1 Background

Digital libraries (DLs) are a major advance in information technology that frequently falls short of expectations [8,17,11]. Crabtree *et al* [9] identified problems with digital libraries through research into physical academic library interaction patterns with regard to information searching strategies. There are two principal aspects of their findings:

1. the importance of collaboration between the librarian and the user in the searching activity, and
2. the significance of social context in digital library design.

However, Crabtree *et al* [9] concentrated on one aspect of library interaction (i.e. information searching) within the confines of a physical library and with library assistants rather than subject librarians. Covi & Kling [8] argue that understanding the wider context of usage is essential to understanding digital library use and its implementation in different social worlds. Negative reactions to digital libraries are often due to inappropriate system design and poor implementation [1,2,7]. However, there may be other less obvious social and political repercussions of information system design and deployment. Symon et al [15] have identified, within a clinical setting, how social structures and work practices can be disrupted by technology implementation. Although academic DL systems do not deal with sensitive, personal information, apparently innocuous data can be perceived as a threat to social and political stability [1,2]. There are several accounts that detail the importance of social context for digital library design and implementation [7, 8]. To understand the impact of DLs within academia, an in-depth evaluation is required of the implementation and use of these applications across disciplines, within their specific social and organisational settings. However, as Covi & Kling [8] have highlighted, there are few high-level theories that aid designers in understanding the implication of these issues for DL design and implementation.

1.2 Social Context and Roles

DL research increasingly focuses on the importance of directing DL design towards work practices and the communities they support [12 ,8]. Covi and Kling's [8] research into patterns of usage for DLs within an academic context identified the importance of roles within effective DL design. A parallel study to this research completed within the clinical domain found that DL technology was perceived as a threat to senior staff members' roles due to their poor training and support [1,2]. Traditional organisational norms and roles were reversed by DLs, allowing junior clinicians easier access to information than senior clinicians.

It is also important to understand the user's informal practices and how they interact within organisational dynamics, changing situations, evolution of task definitions, or social and political aspects (e.g. staff motivation, hierarchies). Adams & Sasse [3] found that systems that do not take into account these practices and are perceived to restrict them would be circumvented. DL designers must therefore design their systems around user practices, understanding both social and organisational norms. The electronic dissemination of information within various settings can be used and interpreted in politically sensitive ways. Digital libraries, in particular, can change the context of people's work practices, and can therefore restructure their relationships with both each other and the task in hand [13,15]. The restructuring of these professional relationships can have far-reaching social and political consequences. Ultimately, system designers should be aware of social and political motivations within an organization in order to develop and implement more sensitive design strategies.

2 Research Method

Focus groups and in-depth interviews were used to gather data from 25 academics and librarians from 4 different campuses within a London university. 10 of those interviewed were from Humanities, 10 from Computer Science and 4 from Business with the split of the sample being approximately 50% librarians / academics. The final respondent was from a key managerial role within library services. The academics were selected from all levels within their department (i.e. lecturer, Senior Lecturer, Reader, Professor). There was a representative sample from each department of teaching and non-teaching staff. Of the 13 librarians interviewed, the majority were subject librarians with authorization to acquire and support digital resources for their discipline.

Four issues guided the focus of questions:

- Perceptions of their role within the academic setting and information requirements.
- Perceptions of how information is currently accessed, and how these processes accommodate or inhibit current working practices.
- The impact of organisational social structures and patterns of interaction on information resource awareness, acceptance and usage.

- Technology perceptions (specifically of DLs) and how these affect other issues already identified.

A pre-defined concept for a ‘Digital Library’ was not employed so that users were allowed to explore what they perceived of as a digital library. This resulted in a discipline distinction between the different definitions given to similar electronic resources (a summary of which can be seen in the results). Although various electronic resources were reviewed, three main DLs were discussed - the ACM DL, PROQUEST and LEXUS.

An in-depth analysis of respondents’ information and technology perceptions was conducted using the Grounded Theory method. Grounded Theory [14] is a social-science approach to data collection and analysis that combines systematic levels of abstraction into a framework about a phenomenon which is verified and expanded throughout the study. Once the data is collected, it is analysed in a standard Grounded Theory format (i.e. open, axial and selective coding and identification of process effects). Compared to other social science methodologies, Grounded Theory provides a more focused, structured approach to qualitative research (closer in some ways to quantitative methods) [14]. The methodology’s flexibility can cope with complex data, and its continual cross-referencing allows for the grounding of theory in the data, thus uncovering previously unknown issues.

3 Results

The results identified different perceptions about electronic resources not only between disciplines but also, more importantly, between the librarians and the lecturers. Key to these differences were the current and past roles of the library and how lecturers and students interacted with it. Web-based digital libraries, while alleviating most library resource and interaction problems, require a change in the librarians’ role if they are to be implemented effectively. Without this role change, lecturers and students were found to have a poor awareness and understanding of digital resources, and resorted to the web and online personal collections as an accessible alternative. Finally, DL interactions were marred by their design and inadequate support. Many digital libraries were found to be inappropriately designed for users’ needs, marginalizing their importance in the educational process for both students and researchers.

3.1 Digital Resource Perceptions

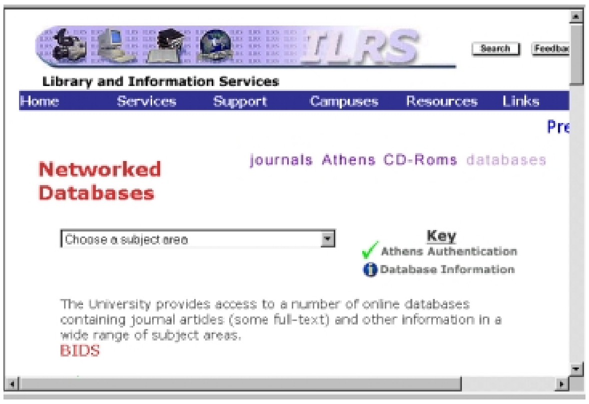
The respondents’ perceptions of and definitions for a wide range of electronic resources were identified. The results showed that, with slight variations, the users had a uniform perception of what a digital library, a database, and an archive were. The definition of a DL usually included that it was a large store of general but up-to-date information in various media with current usage. However, an archive was invariably denoted as a subject specific historical collection with clearly defined parameters, which is not in current use. A database was described by most as a way of structuring and organizing information, which could be accessed either by CDROM, local networks or the web. One librarian added that a database contained only summarized, abstract or citation information while a digital library contained the full text.

An analysis (using normalized data) of how often the users referred to different terms for digital resources within the interview is shown in Table 1. There were some interesting differences in how often the librarians and lecturers from different disciplines commented on these resources. A divergence can be seen between electronic resources discussed by lecturers and librarians. Computer science and business lecturers referred to similar electronic resources as ‘digital libraries’ while in humanities they were identified as ‘archives’. However, the librarians invariably referred to all these resources as ‘databases’.

It may be noted that the library web site (see Figure 1) reflects the librarians’ database terminology, without any reference to digital libraries or archives. Several of the digital libraries are located under the headings ‘databases’ and ‘journals', despite (in the case of, for example, the ACM digital library) storing more than journal publications. Many of the resources under the title ‘Networked Databases’ were then identified in the description on the site as digital libraries (e.g. Medline, EBSCO).

Table 1. Percentage of each digital resource referred to by users

	Digital Library	Database	Archive	Web
Lecturers				
CS / Business lecturing	**31%**	10%	6%	53%
Humanities lecturing	7%	3%	**32%**	58%
Librarians				
CS / Business librarian	17%	**35%**	24%	24%
Humanities librarian	3%	**66%**	3%	28%



The University provides access to a number of online databases containing journal artides (some full-text) and other information in a wide range of subject areas.

Fig. 1. Library resource interface

3.2 Current and Changing Roles

The results showed that lecturers tend to perceive the librarians as being tied to books and hard copy resources. For example:

> "They would keep the list of the recommended books with them; if there was Internet URL's then they would print them out and put those in the library as well." *(CS lecturer - teaching)*

It was highlighted by some that library systems also reflected this book-orientated approach:

> "… if they go into the library and they punch into the machine I want something on this subject and it will come up with some books in that area because of the keywords in the title or keywords in that area. It won't come up with journal articles." *(Humanities senior lecturer - teaching / research)*

Ultimately, some library users perceived that the librarians were centred on and possessive of the resources rather than supporting and understanding the users:

> "… the librarians are not user-centred they're information resource centred … they want to protect their resources not to gain access to them." *(CS lecturer - teaching / research)*

The subject librarians, when detailing their role, always mentioned resource acquisition as the first priority in their job role, and then training. Few highlighted the marketing of or on-going support for electronic resources:

> "Provision of materials which involves book selection, journal selection and I suppose even online resource selection … A major part of our work is on library education." *(CS librarian)*

The perceived roles of librarians were identified as relating to current interaction patterns between lecturers, librarians and students. Interactions between librarians and lecturers or students occurred primarily within the physical boundaries of the library. Lecturers and librarians interacted on an informal ad-hoc basis either by 'bumping into' one another (primarily in the library) or by direct initiation from the lecturers or students themselves. Library initiated interaction was email based, usually regarding hard-copy resource acquisitions or discontinuation:

> "But I haven't spoken to a librarian directly for at least 3 years." *(CS lecturer – teaching / research)*

> "I filled in an email two days ago if that counts saying what was good and bad journals. But no, not on the whole. We send in our requests for books." *(CS lecturer - teaching / research)*

Within the humanities department, some interaction became more pro-active with librarians arranging meetings with the lecturers. However, the interaction was always focused on the students and course requirements. Librarians across all the departments were not perceived as identifying and supporting the lecturers' own needs, while some lecturers noted that they needed to know more:

> "It's an area of enormous ignorance for me. If I knew more, I would know better how to advise people." *(Humanities senior lecturer - teaching)*

The librarians often discussed training sessions, but these tended to centre on student training, as the lecturers were notoriously bad at attending these sessions. It was suggested by librarians across the disciplines that this was because the lecturers were embarrassed by their poor electronic resource skills:

> "So if you're running one on medieval studies - the medieval lecturer will come and sit at the back of the class and you know that they're not trying to keep an eye on their class, they're trying to actually learn without appearing not to know." *(Humanities librarian)*

The library department has, over recent years, employed an electronic resource librarian (a new position within the university). Although the job specification is yet to be clarified, the role concentrates on electronic resources across disciplines but is confined to one of the university's many sites. The job requires more pro-active interaction with the lecturers (e.g. attending research group meetings to promote current electronic resources, organizing external training from DL providers) than that of a subject librarian. It was this librarian that highlighted the different demands on librarians in terms of supporting the user required by electronic resources as opposed to print resources.

There is a perception that some librarians may have problems changing to this role. However, the benefits from this change were evident from these results. Only one lecturer interviewed reported having received information about current systems available. This lecturer was based at the electronic resource librarian's site.

The results also showed that DL resources require more interaction between the library and other departments (e.g. IT Support) necessitating further role changes:

> "… now there is also a bit of a barrier with the computing staff about whether they should be bothered with this … they're loathe to see that their role is also changing." *(Electronic resource librarian)*

With current roles and interaction patterns, students who need support have to leave their searches wherever they are (e.g. computer room) and go to the library to seek advice.

In summary, the roles and expectations of students, teaching staff, librarians and IT support staff are all being forced to change as digital libraries and similar resources are introduced.

3.3 Digital Libraries or the Web

As can be seen from Table 1, lecturers across the disciplines frequently note the importance of the web as an electronic resource. The CS lecturers, in particular, highlighted the importance of using the web as the main supplement to core books. However, it was noted that the level of plagiarism from the web had dramatically increased over recent years. It was also found that students had very poor skills in searching and identifying reputable sources. Lecturers and librarians were both seeking to address these inadequacies with training to improve searching and information discerning skills. Neither the lecturers nor librarians mentioned the benefits of DLs (i.e. guaranteed reputable resources, discipline focused) as opposed to the web. Lecturers frequently commented about the attractiveness of the web for both themselves and the students. The mystical and consumable qualities of searching the web were identified as key incentives. Web searching and the wide variety of web resources were noted as an easy consumable that led the user towards large quantities of information for relatively little effort:

> "Some people when they use search engines they type in a question and if they don't get the answer that they are looking for they type in another one. Just like prayers" *(CS lecturer – teaching / research)*

> "I mainly use Amazon for books a lot. I find the books and download and print off the summaries for the students" *(CS Professor – research)*

Ultimately, the lecturers had poor awareness of what digital libraries were available via the web or subscribed to by the university:

> "no like advice, certainly no tailoring of information from the library service." *(Humanities senior lecturer – teaching)*

Several of the humanities lecturers, for example, suggested that a useful digital resource would be an online newspaper archive so that the students did not have to travel to the physical newspaper library to complete their research. However, all the librarians noted the successful acquisition of this same resource for the past year. This example highlights the importance of not only acquiring the right resources but also adequately marketing them. The main library approach to marketing these resources was by links on the library web-page, induction courses, word of mouth or handouts within the library:

> "We explain about it at the skills session they get and we have sheets at the desk that we give out." *(humanities librarian)*

> "When new staff come in we make them aware of the databases that are available which we think they'll probably like in their subject area." *(CS librarian)*

A survey conducted by the library department, however, had shown that few lecturers knew about the library web site. Also, both the students and lecturers were found to rarely attend the library.

> "Problems with students - they just tend to be library phobic." *(CS lecturer – teaching)*

> "And they don't really use it [the library] themselves [lecturers]. Because they use the same journal articles every year. So in that sense there is no, very little connection between academics and us." *(Humanities librarian)*

3.4 Digital Library Design to Support User Needs

Even if lecturers were aware of digital library resources, their poor design relative to user needs discouraged usage. Some humanities lecturers, for example, noted that they needed to see the whole page of a newspaper, including advertising and other articles, to assess advertising and marketing strategies undertaken. However, most newspaper digital libraries assume that only content is important and that even this does not have to remain accurate to the printed version (i.e. to a specific electronic version). This makes the resource an inadequate replacement for hard-copy versions. However, this same discipline was eager to gain access to digital libraries for a variety of multimedia resources (e.g. visual media, television programmes, films, music) that are not currently being provided.

Other digital formatting issues related to hyperlinks within documents. For some disciplines, this can aid in effective information gathering. However, in disciplines where the flow of the content is important (e.g. literature, journalism), hyperlinks can be a disadvantage:

> "The only thing I worry about with the digitising of works are the way that it changes our interpretation of the documents. I've seen it with our students as soon as you have something on the web with hyper-links they read the information in a different way. Speed reading and focusing in on the keywords that will take them forward to further information. Their own interpretations of the document are taken out of their hands and moulded by the person who digitised the document rather than the author." *(Humanities reader – teaching / research)*

Whatever the disciplines, there were continual references to reading large quantities online being difficult, and printing expensive:

> "Most people if they're going to do it, serious reading, download it and print it off." *(CS senior lecturer – teaching)*

> "The sort of texts they need to have access to are widely available and the quantity we expect them to look at would be a problem as far as

> reading them on a computer or expensive to print up." *(Humanities Reader – teaching / research)*

Another issue that arose with all the digital libraries across the disciplines was one of poor usability and support for infrequent users. Many of the interfaces were continually changing, so a gap in usage meant relearning the system, which sometimes outweighed the benefits of its use:

> "So it doesn't encourage students or staff to use it because unless you're using it every week you lose it. So if you don't go into it for a couple of weeks you've lost it." *(Humanities librarian)*

> "You have to keep training every year, every 6 months. It's not static" *(CS librarian)*

Even frequent users complained that they were often lost when interfaces changed without notification or links to support them in re-learning the new interface:

> "I mean the other thing about ACM and many of the other databases is that they don't even tell anyone about it - they just change it. ACM just went and you couldn't get into it there were all these phone calls what's happened to the ACM we can't get into it. And the other thing was they didn't answer anyone. They didn't tell anyone and then they wouldn't answer back to queries." *(Electronic resource librarian)*

Some in library services summarized the design problem succinctly:

> "Electronic libraries and digital libraries in the UK very seldom descends to looking at the way that these things might be used, how they might fit into people's work patterns and needs. So far we have been inventing tools and then trying to find markets for them rather than doing the market research and then providing the right tools." *(Library management)*

The importance of a pro-active role across the departments supporting and understanding the users' needs is evident from this research. However, understanding users' needs requires increased communication and collaboration to aid in an understanding of potential successful and unsuccessful implementation strategies. Ultimately, as one lecturer pointed out, there is no communication from the library about strategies that they might be taking. This respondent identified this as due to the culture of library systems as a whole:

> "Because librarians have the skill or disability of making difficult answers." *(CS lecturer – teaching / research)*

4 Discussion

Bishop's [4] study into DL users from different social and economic backgrounds found that they can be easily deterred from DL usage and that poor awareness of library coverage prevents a full understanding of DL potential. The findings detailed in this paper have highlighted the importance of organisational roles and interaction patterns. Lecturers were found to have a poor awareness of DLs available as well as an understanding of their potential. They were also found to be deterred from using these resources because of poor relevance to their needs, support and usability for infrequent users. A link was identified between these issues and the current poor interaction model in place between lecturers and librarians. The disparity between librarians' and lecturers' attention to different electronic resources highlights diverse perceptions of information resources.

A key factor in the poor interaction patterns between librarians and lecturers, to which they referred frequently in interviews, was their ad hoc, informal nature. While these patterns may have, in the past, been effective, the perceived benefits of online resources have initiated a change in users' requirements from librarians and library resources. Whether librarians are or are not conceptually bound to books, it is important to highlight that across the disciplines lecturers' requirements are not. Lecturers are, however, inadequately aware of what electronic resources are available, and require support in their learning and use of these facilities [4]. The driving force for the changing role of librarians should be the attractive yet unreliable reputation of web information. Students can be lured into this fantasy information consumable world and, without lecturers acquiring the knowledge to guide them towards DLs, the impact of the web will increasingly dominate academic resource use. Ultimately, the librarians' role must change into one that is more pro-active and flexible – for example, attending small research group meetings, and helping to develop and support the resources that users need. The feedback to and from both developers and users provide the information and contextual knowledge that each requires [9]. The increased importance of electronic resources also means a role change for IT support, with increased collaboration required between IT and library services.

Finally, the design and implementation of DLs should cater for the discipline differences that were clearly highlighted by these results. For example, the importance of context and full text original documents required by the humanities will not be fulfilled if systems are designed to present content only via the abridged hyper-linked documents required by other disciplines [5].

5 Conclusion

This research has highlighted how related social and organisational issues can impede effective technology deployment. To counteract these problems, DL designers and implementers must first identify the social context prior to technology design and deployment [7,8,9,10]. There is a need within this context to increase awareness of digital resources available and their potential, within specific academic contexts and disciplines [9]. There is also a need to strongly supporting training for some lecturers with a supportive and non-judgemental approach [1,2]. Ultimately, lecturers and students need the services of an information expert role to support and inform them.

References

1. Adams, A. & Blandford, A.: Acceptability of medical digital libraries. In the Health Informatics Journal (2002) 8 (2) 58 – 66.
2. Adams, A & Blandford, A.: Digital libraries in a clinical setting: friend or foe. Proceedings of ECDL'2001 Springer (2001) 231-224.
3. Adams, A. & Sasse, M. A.: The user is not the enemy. In Communications of ACM. ACM Press (Dec. 1999) 40 – 46.
4. Bishop, A. P.: Making Digital Libraries Go: Comparing Use Across Genres. Proceedings of ACM DL '99, ACM Press (1999) 94-103.
5. Bishop, A. P.: Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components. Proceedings of ACM DL'98, ACM Press (1998) 29-39.
6. Blandford, A., Stelmaszewska, H. & Bryan-Kinns, N.: Use of multiple digital libraries: a case study. Proceedings of JCDL'01, ACM Press (2001) 179-188.
7. Caidi, N.: Technology and values: Lessons from central and Eastern Europe. Proceedings of JCDL'01, ACM Press (2001) 176-177.
8. Covi, L. & Kling, R.: Organisational dimensions of effective digital library use: Closed rational and open natural systems model. In Kiesler, S (ed) Culture of the Internet. Lawrence Erlbaum Associates, New Jersey (1997) 343-360.
9. Crabtree, A., Twindale, M., O'Brien, J. and Nichols, M.: Talking in the library: Implications for the design of digital libraries. Proceedings of DL'97, ACM Press (1997) 221-228.
10. Kling, R.: What is social informatics and why does it matter? D-lib Magazine, (1999) 5(1), January. *www.dliborg/dlib/january99/k/ing/01/<lmg.hgml*
11. Marchionini, G. & Maurer, H.: The roles of digital libraries in teaching and learning. In Communications of ACM, ACM Press (April. 1995) 67 – 75.
12. Marchionini, G., Nolet, V., Williams, H., Ding, W., Beale Jr., J., Rose, A. Gordon, A., Enomoto, E. and Harbinson, L.: Content + Connectivity => Community: Digital Resources for a learning community. Proceedings of ACM DL'97, Philadelphia, ACM Press (1997) 212-220.
13. Schiff, L., Van House, N. & Butler, M.: Understanding complex information environments: a social analysis of watershed planning. Proceedings of ACM DL'97, Philadelphia, ACM Press (1997) 161-168.
14. Strauss, A. & Corbin, J.: Basics of qualitative research: grounded theory procedures and techniques. Sage, Newbury Park (1990).
15. Symon, G., Long, K & Ellis, J.: The Coordination of work activities: co-operation and conflict in a hospital context. Proceedings of Computer supported cooperative work, ACM Press (1996) 5 (1) 1-31.
16. Theng, Y.L., Duncker, E., Mohd Nasir, N., Buchanan,G. & Thimbleby, H.: Design guidelines and user-centred digital libraries. In Abiteboul, S. & Vercoustre, A.(Eds.), Proceedings of ECDL'99 (1999) 167 – 183.
17. Wyatt, J.: The clinical information access project, New South Wales: lessons from a NeLH precursor. In proceedings of Advances in clinical knowledge management, Presented at ACKM 3 (2000).

Building Digital Libraries for Children: Reviewing Information Literacy of Students and Teachers

Marcia Mardis[1] and Ellen Hoffman[2]

[1] Center to Support Technology in Education, Merit Network, Inc.,
Ann Arbor, MI 48105, USA
mmardis@merit.edu

[2] Teacher Education, Eastern Michigan University
Ypsilanti, MI 48197, USA
ehoffman@online.emich.edu

Abstract. Research on student and teacher information literacy from the US is reviewed in relation to the issues needed to ensure that digital libraries will be able to adequately meet the needs of education. Student use of Internet for research while increasingly common is not efficient or critical. Teacher skills in information literacy instruction are critical for improving student capabilities. Recommendations are made for potential tools and services that will be a key for improving library functionality.

Children in the US today are growing up in an age when computers are as common in the home as television was half a century ago. In fact, more children aged 8-17 chose the Internet over television as the medium they could least live without (33% Internet versus 26% television) [1]. New capabilities of less expensive devices such as handheld devices for network connection are increasingly lowering the barriers to the ubiquitous presence of the networked digital world as a space open to children as one foundation of their learning environments. In this growing digital world, the potential for digital libraries to have a massive effect on education is now no longer a theoretical future limited by access. In fact, informal studies indicate that US students are using the Internet for school research more than traditional print sources even when this is not a requirement for classroom assignments, with 94% of teens in a recent survey indicating that they used the Internet for school assignments and 71% noting it was their primary source [2]. However, use does not necessarily equal efficiency or literacy in the use of information sources. Despite the attention given to the topic of information literacy, it has yet to be attained among elementary and secondary students in a widespread manner.

Although students are the ultimate beneficiaries of information skills, gap analysis indicates that teachers in classroom settings are key factors in the transfer of information literacy skills. An examination of the literature about teacher Internet mastery and of the perceived barriers to information literacy reveals some factors digital collection developers will want to consider. A recent survey conducted by the National Center for Education Statistics (NCES) found that 99 percent of full-time regular public school teachers reported they had access to computers or the Internet somewhere in their schools. Sixty-six percent of public school teachers reported using computers or the Internet for instruction during class time; thirty percent reported assigning research

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 404-405, 2002.

using the Internet to a moderate or large extent. The ways teachers direct students to use computers or the Internet varied by instructional level, school poverty level, and hours of professional development. Elementary school teachers are more much likely than secondary school teachers to assign students practice drills using computers and to have their students use computers or the Internet to solve problems. Secondary school teachers, however, are more likely to assign independent work that involves using the Internet [3].

This sense of technology confidence affects the ways teachers control knowledge in the classroom. Because information literacy is an outcome that is linked to technology related skills, it requires a shift to self-directed and inductive instruction. Teaching strategies, therefore, need to be revamped. This need for alteration in teacher practice can cause mental and practical confusion and discomfort [4]. Most often, this discomfort manifests in perceptions about introducing electronic resources skills including insufficient class time, teacher fear of failure, and perception that information literacy skills are the role of another class [5]. Yet, when educators are able to counter this uncertainty and inexperience with mastery skills, teachers can develop a "capacity to suspend belief, take risks, and experience the unknown [that] are essential to learning (p. 16)" [4] There is little research that investigates potential barriers to student use of online resources from the student perspective.

Digital information tools are developed with the best of intentions. However, these all too frequently fail to consider how students will perceive their importance. Because students are familiar with computers, assumptions are made about their knowledge of information sources. Since students tend to lack experience with relatively unstructured information sources like journal articles and databases, the degree to which students process and synthesize subject matter is impacted. Students do not have foundational information seeking and synthesis skills, despite their technological facility [6].

When digital library developers begin to acknowledge and address these factors, educators and students will benefit.

References

1. The Home Technology Monitor. How Children Use Media Technology 2001. Menlo Park, CA: Knowledge Networks/SRI, 2001.
2. The Pew Internet Trust. The Internet and Education: Findings of the Pew Internet & American Life Project. Washington, DC: The Pew Research Center, 2001.
3. United States Department of Education. Teacher Use of Computers and the Internet in Public Schools. http://nces.ed.gov/pubs2000/quarterly/summer/3elem/q3-2.html
4. Fullan, Michael. Change Forces: Probing the Depths of Educational Reform. London: The Falmer Press, 1993.
5. Farmer, D.W. Information Literacy: Overcoming Barriers to Implementation. *New Directions for Higher Education* (Summer 1992), 103-112.
6. Hartmann, E. "Understandings of Information Literacy: The Perceptions of First year Undergraduate Students at the University of Ballarat." Australian Academic & Research Libraries 32 (April 2001): 110-122.

The Use and Functionality of the Environmental Data Registry: An Evaluation of User Feedback

Panayiota Polydoratou[1], Michael Pendleton[2], and David Nicholas[1]

[1] The Internet Studies Research Group,
City University, Department of Information Science London EC1V 0HB, UK
(p.polydoratou@city.ac.uk, nicky@soi.city.ac.uk)

[2] United States Environmental Protection Agency
Office of Environmental Information
1200 Pennsylvania Avenue, NW
Mail Code 2822-T
Washington, DC 20460
pendleton.michael@epa.gov

Abstract. The Environmental Data Registry (EDR) is a cornerstone of the U.S. Environmental Protection Agency's efforts to manage and integrate environmental information for the purposes of safeguarding human health and protecting the environment. In order to gain an understanding of how the EDR can better support its intended audience, EPA hosted a user conference in January 2002. We developed and distributed a questionnaire to the conference attendees, and gathered their input regarding the use and functionality of the EDR. Nineteen of the fifty attendees provided feedback on EDR services and products, and efficiency of information retrieval. Survey results identified difficulties these participants encountered while searching for information, as well as the utility of the information they retrieved. Participants also provided suggestions for existing and additional services. This paper is a first of its kind assessment of the use and functionality of an active metadata registry system based on the views of its users.

1 Introduction

"The Environmental Data Registry (EDR) is a comprehensive, authoritative source of reference information about the definition, source, and uses of environmental data. The EDR catalogs the Environmental Protection Agency's (EPA) major data collections and helps locate environmental information of interest." Development of the EDR began in 1993, and made significant strides under the auspices of EPA's Reinventing Environmental Information (REI) Initiative that began in the late 1990's. The EDR is EPA's primary resource for metadata pertaining to data within the Agency's major information systems. The EDR also serves as a clearinghouse for EPA's data standards. It is closely allied with several other registry systems including the *Terminology Reference System*, the *Substance Registry System*, the *Chemical Registry System* and the *Biology Registry System*. Those systems provide search tools for re

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 406-418, 2002.

trieving information on how environmental terminology, and physical, chemical and biological substances are represented in the Agency's regulations and data systems.

Since its inception, the EDR has continually evolved in order to serve the needs of its users. The User's Conference held in January 2002 provided a wealth of input toward continued improvements, and additional functionality has been added as a direct result of input provided by conference participants.

The maturity of the EDR in relation to other existing metadata registry systems makes it a suitable candidate for evaluating how people use these systems; it also allows us to explore how such systems might be enhanced to better serve their customers. This paper discusses results of a study, which surveyed EDR users for feedback regarding system use and functionality. This is a first of its kind assessment of user opinion -- such information is vital for further evolution in the field of metadata management.

2 Literature Review

The application of metadata appears to provide an effective answer to discovering and retrieving networked information (Dempsey, 1996, 2000; Dillon, 2000). Woodward (1996) and Vellucci (1998) have conducted detailed literature reviews on the evolution and use of metadata formats; Milstead and Feldman (1999) bolstered those studies by over viewing emerging metadata projects and standards. Dempsey and Heery (1996), within the requirements of the DESIRE project to create a generic format for use by Internet Based Subject Gateways, have produced a comparative description of several metadata formats. Caplan (2000) further discussed the challenges for metadata schema implementers with reference to some metadata schemas.

The diversity of communities on the Internet and the information needs of each of those communities indicated the application of different metadata schemas - sometimes simpler and sometimes more sophisticated - for the description of resources. Interest in metadata registries arose from the need to be able to search across diverse metadata schemas among information systems. Leviston (2001) distinguishes between two different types of metadata registry system prototypes. Systems that usually serve as reference points by listing URLs that point to Web sites of metadata initiatives and projects and systems, which are concerned with the management of the evolution of metadata vocabularies over time and provide with mappings between schemas. The latter are usually developed to satisfy more sophisticated needs (p.2) and they are based on the application of ISO/IEC 11179 standard that refers to the Information Technology: Specification and Standardization of Data Elements. Metadata registries are registration authorities associated with the description, discovery, storage and exchange of metadata, and as such they address data sharing and standardisation problems often associated with networked resources.

Research on metadata registries is still in its formative stages. In addition to the EPA, the UK Office of Library and Information Networking (UKOLN), the University of Goettingen in Germany, and the Australian Institute of Health and Welfare have also been actively involved in developing and implementing metadata registries.

In Europe, UKOLN has taken the lead on metadata registry systems research through the EU funded projects DESIRE I & II that led to the development of ROADS software, which is being used by many Internet-based subject gateways to support resource description and to facilitate cross searching and interoperability among resources. Further to that, the SCHEMAS project, completed in December 2001 provided with a forum for metadata schema designers and implementers and launched the SCHEMAS metadata registry system.

The University of Goettingen in Germany developed MetaForm, which is described as "…*a database for metadata formats with a special emphasis on the Dublin Core (DC) and its manifestations as they are expressed in various implementations*". MetaForm is an outgrowth of the META-LIB project (Metadata Initiative of German Libraries). It facilitates the registration of metadata formats along with a description of their elements with a particular emphasis on DC. Metaform provides three services: Crosswalks, Crosscuts and Mappings. Throughout those services it allows for comparisons of Dublin Core with other formats, provides with specification of DC elements applications within different schemas and supports mapping among those elements applications.

The Australian Institute of Health and Welfare's Knowledgebase is described as *"…an electronic register of Australian health, community services, housing and related data definitions and standards. It includes the relevant National Data Dictionaries, national minimum data sets, and the National Health Information Model."* Knowledgebase incorporates for the first time in electronic version the National Health Data Dictionary and the National Community Services Data Dictionary, which have been respectively published in hard copies since 1989 and 1998. It bases the data element definitions used from the previous resources on ISO/IEC 11179. Additionally it supports the use of National minimum data sets and facilitates links among current information agreements on collection and provision of data and data elements.

The DC Metadata Registry is defined as *"… a collection of RDF schemas, application profiles and related semantics belonging to various resource communities. Its goal is to promote the discovery, reuse and extension of existing semantics, and to facilitate the creation of new vocabularies."* Recent developments involve the launch of the second registry prototype that supports multi-lingual searching of DC elements. For more information on the DC metadata registry prototype, see Heery & Wagner (2002).

From current prototypes we see that there is a variation in metadata registry systems functionality. All of the above though support functions that are considered essential of the role and requirements of metadata registry systems. Those include the description of data elements, provision of guidelines for their use, mappings among elements of different metadata schemas and, for the case of DC registry, facilitation of multilingual searching. Consistent update of registry content is essential for validity and credibility. For views on the role and functional requirements see Leviston (2001) and the DCMI Open Metadata Registry Functional Requirements (2001) [1] and for discussion on users' expectations of metadata registry systems see Baker et al. (2001). A description of the purpose and scope of DC with a particular emphasis on multilingual support is provided on the DC Registry Web site. Information on minimum levels of metadata description to resources hosted on Internet based subject gateways is included in DESIRE II Handbook.

[1] This document is currently under review. Available at: http://dublincore.org/groups/registry/fun_req_ph1.shtml

3 Aims and Objectives

The aim of this survey research effort was to evaluate the EDR for ease of use, functionality, as well as user expectations for additional functionality. In particular, the objectives were:

- To establish frequency and duration of use.
- To identify which services and products provided by the EDR are used most often, and to rank their usefulness. (We defined usefulness as the extent to which the service provides effective solutions and/or answers questions).
- To assess user satisfaction with regards to information resources coverage and identify those of greatest interest to them.
- Find out how information retrieved from the EDR has been used and question whether their organisation has a requirement for a metadata registry system.
- To assess the EDR's functionality by asking conference participants to describe how they retrieve information and to indicate its relevance.
- To rank various services from the standpoint of content and interface usability.
- To receive feedback regarding how the EDR could support their work.

4 Scope and Limitations

This survey effort is part of ongoing research examining the use and applications of metadata registry systems. For the purposes of this paper, we are adapting DESIRE project's definition of *metadata registry systems* as:

> "…formal systems that can disclose authoritative information about the semantics and structure of the data elements that are included within a particular metadata scheme. Registries would typically define the semantics of metadata elements, give information on any local extensions in use, and provide mappings to other metadata schemes" (http://www.desire.org/html/research/deliverables/D3.5)

Although we did not extensively survey EDR Users,[2] we believe that these results provide insight regarding user views, and are sufficient to meet our objectives. Interpretation of results needs to be considered in light of two important observations concerning the audience, specifically:

- One third of responses came from first time users (Note: The conference was designed to accommodate users with varying degrees of experience).
- With the exception of two individuals, all respondents were either EPA employees or EPA contractors.

5 Methodology

The Users' conference was publicised among the 800 registered parties through the Standard Update Newsletter and on the EDR's website for a period of three weeks

[2] We aim to enhance these results with web usage statistics of the service and interviews with both users and employees at EDR.

between the 1st and 23rd of January, 2002. The number of participants was limited to fifty people due to the conference facility size restriction.

A questionnaire was designed and distributed to attendees of EPA's 1st EDR Users' Conference, which was held at the U.S. Bureau of Labor Statistics' Conference Center on January 24, 2002 in Washington, DC.

Of the fifty people that attended the conference, nineteen (38%) completed and returned the questionnaire. EPA reports that half of the attendees were Agency employees, and that "*Others in attendance were primarily contractors supporting EPA programs; interested individuals from other government agencies, including the Veterans Health Administration, the Defence Information Systems Agency, and Environment Canada; and representatives from the Chemical Abstracts Service*".[3] The objective of the conference was to demonstrate EDR services and to obtain user feedback regarding these services.

The questionnaire consisted of four parts:

- Part 1: User familiarity with electronic resources, frequency of access, and difficulties experienced while searching online resources;
- Part 2: How frequently they use the EDR, and how familiar they are with the products and services provided;
- Part 3: The usefulness of information retrieved from the EDR and;
- Part 4: General comments and feedback on the registry. [4]

6 Results

6.1 Characteristics of Population

The EDR users' conference attracted attention from individuals representing governmental organisations in the U.S. (Veterans Health Administration, the Defence Information Systems Agency) and Canada (Environment Canada), representatives from the Chemical Abstracts Service and EPA contractors and other employees with an interest in the development and implementation of data standards.

Part 1 of the questionnaire facilitated two aims. The first aim was to draw an indicative profile of the conference attendees' information seeking behaviour. Questions were designed to assess user familiarity with online information resources, frequency of use, and difficulties experienced while searching for information.

Sixteen of the respondents noted that they access and use online information on a daily basis, although only about half considered themselves to be advanced users. The majority (fourteen respondents) regards searching and retrieving online information as a relatively easy process. Despite that, they have pointed out restrains they encounter (Table 1[5]) from time to time indicating that *"Large amount of unorganised informa-*

[3] Standard Update: better data through standards. EDR Users' Conference. Winter 2002. Available at : http://www.epa.gov/edr/wint2002.htm

[4] Results from this part are presented throughout the paper instead of forming a separate section.

[5] Please note that results in this table represent eighteen of the nineteen responses. One person did not reply.

tion" and *"Badly designed sites that are difficult to navigate"* are the two most common reasons that prevent them from easily accessing the information they seek.

Table 1. Difficulties obtaining information of interest

DIFFICULTIES EXPERIENCED WHILE TRYING TO OBTAIN INFORMATION FROM ONLINE SOURCES	RESPONSE
Too much unorganized information	11
Badly designed sites that I find difficult to navigate	11
Lack of supportive services	6
Too much information	5
Other (please specify) – Too many irrelevant sites returned on a fairly general search – Confusing search interface – Info[rmation] lacking – EPA firewall limiting (e.g. access to US NRC ADAMS system or US DOE directives)	4
Lack of time required for the search	3
Lack of online help facilities	2
Cost of services/information to obtain	2
Unfamiliar with searching methods	2
Lack of guidance tools	1

The second aim was to monitor user interest in specific EDR predefined areas[6] in order to provide baseline for evaluating our current services. We targeted three services on the EDR – systems development, standards implementation and data harmonization that are accessible via the *How to…Facility*. All three services share the same interface while providing links to different information, relevant to each of the three areas. Most of the respondents indicated an interest in all three areas[7] rather than any one in particular. This result was emphasised through comments that the respondents made stressing the importance of understanding how metadata registry systems work, network with other interested parties in the area of metadata standardisation, and promote the use of the particular registry system.

6.2 Functionality

The second part of the questionnaire sought to gather user opinions on the EDR's functionality. Questions were designed to assess 1) the frequency and longevity of EDR use, 2) familiarity with services and products provided and 3) rank their usefulness and coverage of information contained. Furthermore, we assessed whether the organisation that the users represented has a requirement for a metadata registry system and asked the conference participants to describe in what way.

[6] Under the How to… facility, EDR distinguishes between Data Harmonization, Standards Implementation and Systems Development.

[7] All categories also include the combination of two of the predefined categories plus an indication of other – non-specified by us – areas of interest.

Results showed that one third of respondents were first time users. Five people replied that they have been using the Environmental Data Registry occasionally (a period greater than one month between sessions) and another four indicated that they hardly ever use it. The remaining four responses were from those who have been using the EDR for more than six months (responses ranged from one to five years). One person did not specify. We feel that the diversity in frequency and longevity of EDR use reflects the conference structure to address a broad range of user familiarity and experience supported by the fact that metadata registry systems are still at their formative stages.

We asked the conference attendees to identify the EDR services they've used, and to rank their usefulness. We defined usefulness as the ability of the service to retrieve effective information to the initial user query. We used a scale of 1 (poor) to 5 (excellent) to assess it. *Search Facility* came up as most used service while *MetaPro* (only one person noted that they had used it) [8] came up as least used service. Other services identified included the *Download, How to…* (eight people) and *Help* (seven people) features (Fig. 1). We disregarded results that ranked services, which have not been indicated as "used services".

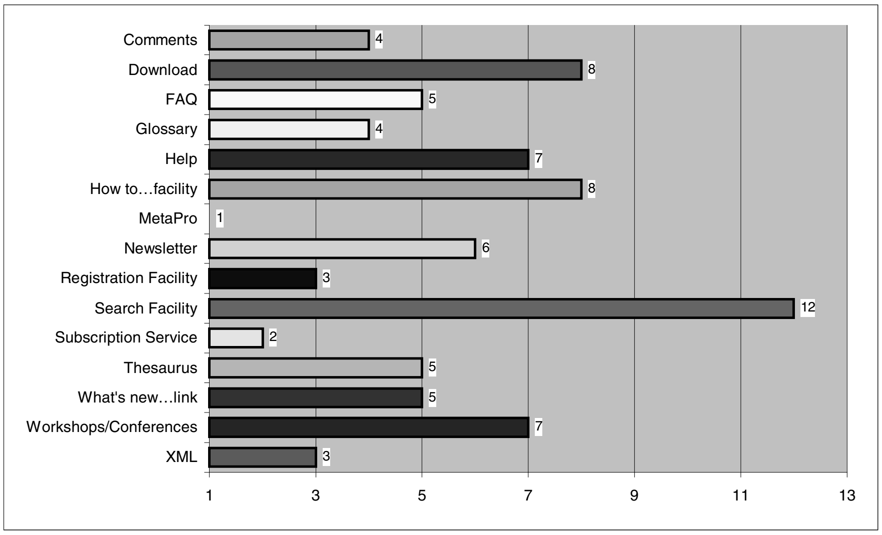


Fig. 1. Services used

The *Search* facility, which has been the most used EDR feature among conference attendees, has been ranked by half of the respondents as a poor or fair service. Among suggestions for improvement were that there should be additional search text for the whole site and Boolean search support. Also, it has been noted that the service speed is slower than desired. *Download* was considered to be a very good service by eight respondents. Only one of the respondents regarded it as fair. Overall, the conference was considered to be an excellent event by the majority of respondents.

[8] Please note that MetaPro is currently under development phase. EPA hopes to progress further this tool in the next fiscal year.

We also wanted to rate user satisfaction of the information resources coverage within the EDR. The majority of respondents replied that although they are happy with the content, they would welcome additions. Most appealing information resources as shown in Fig. 2 are:

- Data standards generally
- Data elements
- Business rules documents

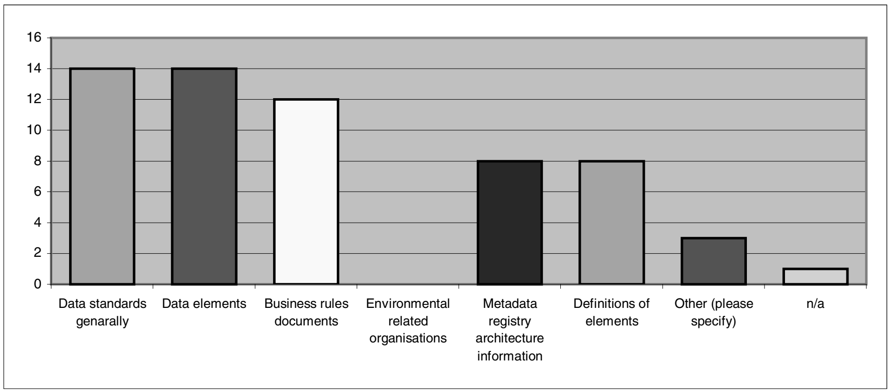


Fig. 2. Most appealing information resources

Business Rules Documents are essential for the implementation of data standards as they provide guidelines for the use of data elements. In particular, respondents specifically indicated an interest in:

- “Harmonisation, "future" services for harmonisation and XML e.g. list management”
- “How the Environmental Data Registry will integrate EPA's ability to enforce regulation[s]”

In order to identify potential applications, we asked EDR users to indicate how they use the information they retrieved. Almost half of the respondents have used the retrieved information to meet needs of their programs. Some of the specifications referred to *“[use of the EDR] services, as Environment Canada is categorising 23.000 substances on persistence, bioaccumulation and toxicity”, “application of data standards to conform to systems development” and “I use the EDR information to ensure the systems I develop conform to EPA standards”*. This result is in line with the outcome that all respondents with the exception of two people declared some kind of association with EDR. Three respondents noted that they have used information from the EDR for their systems development work, or for *[enterprise*] architecture research. Other (21%) included responses such as:

- Audit of data
- Incorporated into system development facility registry
- To check details of data standards as needed for meetings on systems development, standard implementation and data integration
- For data standards implementation in NARSTO and Environment

The majority (74%) of respondents replied that the organization they represent has a requirement for a metadata registry system. This finding is of particular interest given that all but two respondents were associated with the organization hosting the conference (U.S. EPA).

6.3 Usability (Initial Evaluation of Particular Functions)

Part 3 gathered feedback on the EDR's usability. Conference attendees were asked to describe 1)the process of obtaining information from the EDR and 2) the relevance of information that they retrieved, 3) to rank the coverage of services and 4) to evaluate the content and interfaces associated with some of the facilities. Attendees were also asked to indicate future services that they deem most important. Conclusions drawn from this section are:

Almost half of the respondents found and downloaded the information they sought relatively easy but it was noted that more guidance would be helpful. Three respondents indicated that they were unable to find what they wanted, while the remaining eight respondents were divided between those who did not answer the question, and those who considered the whole process easy and straightforward. This result is consistent with our ranking results pertaining to the *Search Facility*, which were found to be poor or fair by half the respondents. One respondent that found it easy to obtain the information that they wanted commented that "*[It has been quite easy and straightforward] except when [the] search did not work."* Another respondent attributed the difficulty with the Search Facility to "*too many links made it difficult to grasp the various features."*

More than two-thirds of respondents were satisfied with the relevance of the information they retrieved from the Environmental Data Registry specifying that it was either relevant or adequately covering their query.

We also asked EDR users to note their satisfaction towards the services provided to help them through their searching. We called these services supportive, as we wanted to emphasise their role in providing help and guidance while searching. Those services included the glossary, the help guide, the search facility, and the site map. A full two-thirds (68%) of respondents replied that while they are generally satisfied with the services provided, they would welcome additional improvements. (Fig.3)

The search facility was a service identified as needing improvement by almost half the respondents, followed by the site map and the glossary. Recommendations from respondents included:

- Needs to standardise terms
- *[Incorporate]* Text search for entire site
- Organisation search should yield ordered (alphabetical?) results
- Add systematic diagram
- Need more context sub settings help- e.g., click on a term & get the definition, an example, etc.
- *[Improve]* Navigation - Hard to get back to where you came from
- Improved search function; acronyms

We asked respondents to indicate their satisfaction in terms of content of the *Business Rules Documents*, *Data Standards,* and *Information Resources.* We also asked them to assess the interface of the *Comparison Matrix* and the *Help*, *Search* and

Site Map facilities. Results are presented in Figures 4 and 5. We should note that almost half of the respondents did not rank the content of the *Business Rules Documents,* although it was one information resource that they indicated strong interest in.

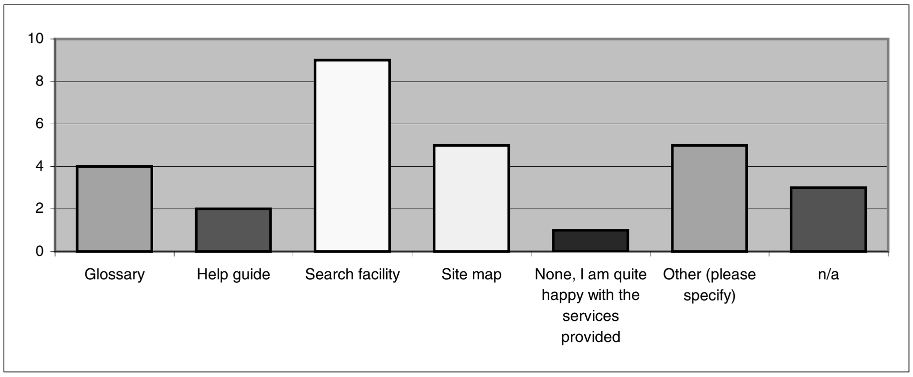


Fig. 3. Suggested services improvements

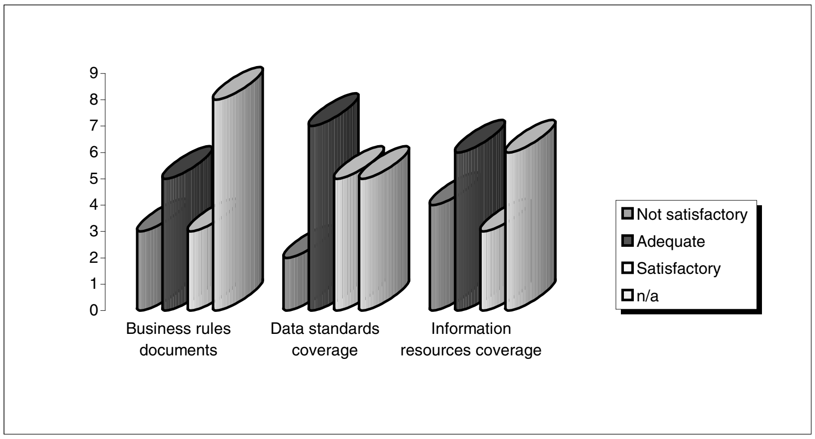


Fig. 4. Ranking of resources in terms of content

7 Conclusions

The 1st EDR Users Conference was an informative event where EDR products and services were effectively presented to the users community. The conference format included two hands-on sessions, giving attendees an opportunity to become familiar with different services through the use of guided exercises.

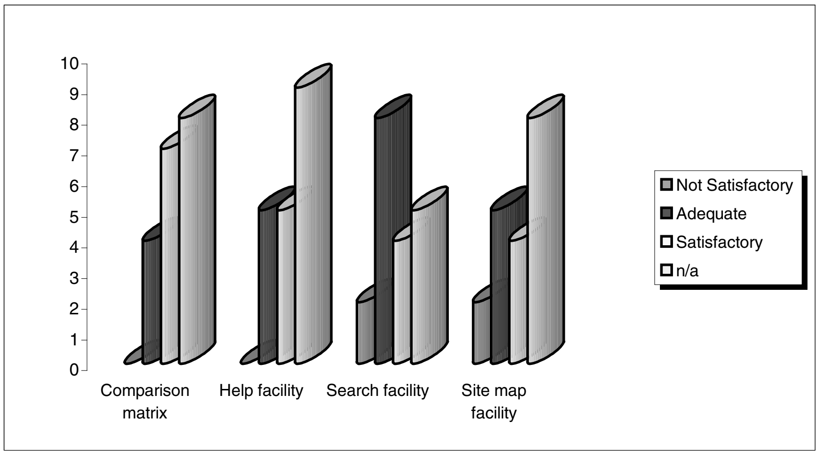


Fig. 5. Ranking of resources in terms of user interface

An indicative profile of the respondents' information seeking behaviour shows that in their majority are people familiar with the use and search of online resources on a daily basis. Almost half described themselves as advanced users who, although they find the process of obtaining information of interest relatively easy, also encounter some difficulties which they attribute mostly to unorganised information and poor web site design and resultant lack of supporting services. Approximately one-third of respondents were first time users, and all with the exception of two people declared a relation with the Environmental Protection Agency, which developed and maintains the Environmental Data Registry. Nine of those who have used EDR in the past are seen as regular users of the system, and four have been using EDR for a period of time that ranges between one and five years. It appears that although the EDR is not used on a very frequent basis, those who do use it do so on a regular basis.

Almost two-thirds of respondents expressed their interest in all three predefined areas (develop systems, implement standards and harmonise data) as opposed to focussing on one particular service. This could be interpreted along with the number of first time users that EDR users are still more interested in learning about how a metadata registry system works. The development of an internal gateway – as we see the *How to…Facility* within EDR - to resources for specific areas of interest represents EDR's ongoing interest and research in the area of metadata standardisation and its user satisfaction.

The most used services among conference attendees were the *Search Facility* followed by *Download* and *How to…*features. Half of the respondents ranked *the Search Facility* as a poor or fair service, while *Download* was considered to be a very good service. The majority of respondents are satisfied with the EDR's coverage of resources, but they would welcome future additions. Of particular interest are the Agency's *Data Standards*, *Data Elements* and *Business Rules Documents.* Again, the majority of respondents replied that there is a requirement for a metadata registry system within their organisation and 42% of respondents have used the information they retrieved from the EDR within their organisation. In particular, applications have been associated with systems development and data standardisation for purposes of ensuring data quality and system interoperability.

Less than half of respondents view the process of obtaining information from the EDR as relatively straightforward, but they would appreciate more guidance during their search. Suggestions refer mainly to the improvement of *Search Facility* and the *Site Map*. Most respondents noted that the information they retrieved from the EDR has been either relevant or adequately covering their initial query, which suggests that the EDR provides users with valuable information, and is an appropriate means for information retrieval needs, in spite of findings that suggest that continued improvement is needed. One key improvement identified by respondents was the need for inclusion of "*Boolean, acronyms and text search for entire site*."

We believe that metadata registry systems are a vital part in the process of data standardisation and metadata management and we feel that research in the area is going to grow in the future, particularly after the establishment of the software to support the mappings and interoperability among different schemas. We feel that results from this survey effort provide with an initial indication of users' expectations of the registry system. This paper is a first of its kind assessment of the use and functionality of an active metadata registry system based on the views of its users. The combination of other research methods such as web usage statistics and interviews with the systems users would enhance our insight of metadata registry systems usage. The analysis of EDR's web usage statistics is the following step in our research effort of the use and functionality of EDR.

Acknowledgments. The authors would like to thank Larry Fitzwater for his useful suggestions and comments while writing this paper.

References

1. Baker, Thomas et al. (2001). What terms does your metadata use? Application profiles as machine - understandable narratives. Paper presented in DC-2001, Proceedings of the International Conference on Dublin Core and Metadata Applications 2001. Available at: http://www.nii.ac.jp/dc2001/ (last visited on the 27/06/2002)
2. Caplan, Priscilla (2000). International Metadata Initiatives: Lessons in Bibliographic Control. Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the challenges of networked resources and the web. Available at: http://lcweb.loc.gov/catdir/bibcontrol/caplan_paper.html (last visited 27/06/2002)
3. Caplan, Priscilla. (1995). You call it corn, we call it syntax-independent metadata for document like objects. *The Public - Access Computer Systems Review*, Vol. 6 (4). Available at: http://info.lib.uh.edu/pr/v6/n4/capl6n4.html (last visited 27/06/2002)
4. Dempsey, L. (2000). The subject gateway: experiences and issues based on the emergence of the Resource Discovery Network, *Online Information Review*, 24(1), 2000, pp. 8-23
5. Dempsey, L. (1996). ROADS to Desire: some UK and other European metadata and resource discovery projects. *D-Lib Magazine*, July/August, 1996. Available at: http://www.dlib.org/dlib/july96/07dempsey.html (last visited 27/06/2002)
6. Dempsey, L. and Rachel Heery (1996). A review of metadata: A survey of current resource description formats. Available at: http://www.ukoln.ac.uk/metadata/desire/overview/ (last visited 26/04/02)

7. Dillon, Martin (2000). Metadata for Web Resources: How Metadata Works on the Web. Library of Congress Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the challenges of networked resources and the web. Available at: http://lcweb.loc.gov/catdir/bibcontrol/dillon_paper.html (last visited 27/06/2002)
8. Duval, Eric, …et al. (2002). Metadata principles and practicalities. D-Lib Magazine. Vol. 8, No. 4, April 2002. Available at: http://www.dlib.org/dlib/april02/weibel/04weibel.html (last visited 27/06/2002)
9. EPA's Data Standards Process. *Standard Update Newsletter*, Vol. 1 No.2, Summer 1999. Available at http://www.epa.gov/edr/1033summer99.pdf (last visited 27/06/2002)
10. Heery, Rachel and Harry Wagner (2002). A metadata registry for the semantic Web. D-Lib, Vol. 8, No. 5. Available at: http://www.dlib.org/dlib/may02/wagner/05wagner.html (last visited 27/06/2002)
11. Heery, Rachel. (1996). Review of metadata formats. *Program*, Vol. 30 (4), pp.345-373
12. Leviston, Tony. (2001). Discussion Paper: Describing metadata registry requirements and realities. January 2001. Available at: http://www.dstc.edu.au/Research/Projects/Infoeco/publications/registry-discussion.pdf (last visited 27/06/2002)
13. Metadata--Understanding its Meaning and Uses *Standard Update Newsletter*, Vol. 1 No.3, Fall 1999. Available at http://www.epa.gov/edr/2016fall99.pdf (last visited 27/06/2002)
14. Milstead, J. and S. Feldman. (1999). Metadata: Cataloguing by any other name. *Online*, Vol. 23 (1). Accessed through Dialog (Gale Group Trade and Industry DB). Also available at: http://www.onlinemag.net/OL1999/milstead1.html (last visited 27/06/2002)
15. Partners. *Standard Update Newsletter*, Vol. 2 No.1, Spring 2000. Available at http://www.epa.gov/edr/2016aspr2000.pdf (last visited 27/06/2002)
16. Standard Development Activities. *Standard Update Newsletter*, Vol. 2 No.2, Summer 2000. Available at http://www.epa.gov/edr/2016bsum2000.pdf (last visited 27/06/2002)
17. Standards Stewardship and Implementation EDR *Standard Update Newsletter*, Vol. 3 No.2, Summer 2000. Available at http://www.epa.gov/edr/spr2001.pdf (last visited 27/06/2002)
18. The New Integrated EDR *Standard Update Newsletter*, Vol. 2 No.3, Fall 2000. Available at http://www.epa.gov/edr/wint2000.pdf (last visited 27/06/2002)
19. The Substance Registry System EDR *Standard Update Newsletter*, Vol. 3 No.1, Spring 2000. Available at http://www.epa.gov/edr/spr2001.pdf (last visited 27/06/2002)
20. Vellucci, Sherry L. (1998). Metadata. *Annual Review of Information Science and Technology (ARIST)*, Vol. 33, pp. 187-220.
21. What's a standard? *Standard Update Newsletter*, Spring 1999. Available at http://www.epa.gov/edr/spring99.pdf (last visited 27/06/2002)
22. Woodward, Jeannette. (1996). Cataloging and classifying information resources on the Internet. *Annual Review of Information Science and Technology (ARIST)*, Vol. 31, pp.189-220.

Adding Semantics to 3D Digital Libraries

Anshuman Razdan[1], Jeremy Rowe[2], Matthew Tocheri[1,3], and Wilson Sweitzer[1,3]

[1]PRISM, Arizona State University, Tempe AZ 85287-5906 USA
[2]Information Technology, Arizona State University, Tempe AZ 85287-0101 USA
[3]Department of Anthropology, Arizona State University, Tempe AZ 85287-2402 USA

Several efforts have begun to archive 3D data in an organized manner to create 3D Digital Libraries. An important challenge in creating an intelligent archiving mechanism is the problem of adding semantics to the original content. Under the NSF funded 3D Knowledge Project (3DK), PRISM researchers have devised a patent pending process to add semantic content to 3D data [1][2][3]. This paper details some aspects of the process including data acquisition, geometric modeling, representation, analysis, adding semantic content, and visual query using a database of 3D bones as an example.

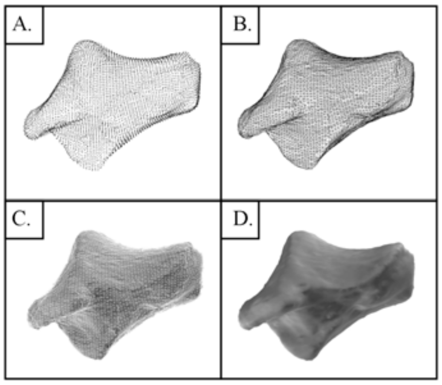


Fig. 1. 3D point cloud (A); wireframe of triangular mesh (B); flat-shaded model (C); smooth-shaded model (D) of a wrist bone

Data Collection. Bones can be scanned using 3D laser scanners. The resulting point cloud data are represented as 3D objects in the form of triangle meshes, a collection of geometry (i.e. points (x, y, z) in space) and topology (i.e., how these points are connected to each other). These meshes are the most common representation method for 3D objects and can also be shaded to assist visualiza-tion of the object (Fig. 1).

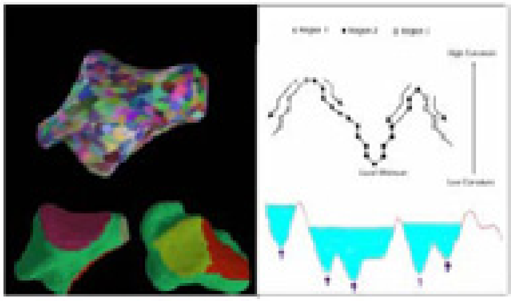


Fig. 2. Before (top left) and after (bottom left) merging similar regions on a wrist bone by changing the watershed depth threshold (right)

Segmentation and Feature Extraction. To raise the level of abstraction of the data (i.e., to be able to describe the object in terms of its various parts), the triangle mesh must be segmented into distinct regions or features. In the case of bones, this involves identifying features such as joint surfaces and muscle attachments, which can then be quantitatively described and cataloged along with the 3D model. Semi-automatic feature extraction and segmentation is accomplished with

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 419-420, 2002.

a watershed-based hybrid segmentation algorithm. Researchers can vary the watershed depth, allowing the algorithm to define regions at various resolutions (Fig. 2).

Feature Editing and Adding Textual Semantics. After segmentation, defined regions are then tagged by the researcher. This additional information, or semantic content, can include contextual and/or 3D data. Using a wrist bone, the trapezium, as an example, associated information such as biographical data (age, sex, population, etc.), data collection information (scanner and software used, etc.), and anatomical names of segmented features, along with 3D data (measurements, spatial relationships between features, etc.), are linked with the raw 3D data. These metadata are recorded using Extensible Mark-up Language (XML) schema. The data are then ready for archiving using any off-the-shelf database software.

Visual Query. Bones are 3D objects and only a fraction of the information they contain can be described textually. The 3DK visual query system is designed to support a variety of modes including text and interactive 2D and 3D models [2][3]. In addition to traditional textual inputs, the interface includes fast interactive visualization capabilities for inputting 3D search criteria. The results interface combines display of raw data and supplemental semantic data from the XML archives with quantification tools to extract additional 3D data directly from the search results. We are implementing 3D compression techniques for faster delivery of 3D data to the client/user via a java interface. The goal is to use Java3D for displaying the 3D data.

Conclusion. 3D digital libraries are steadily gaining popularity. Several 3D catalogs are currently available, but much of the library creation process in terms of shape description remains manual. The approach to digital archiving and analysis of 3D objects outlined herein shows great promise for research in physical anthropology and other fields. The tools provided allow researchers to quantify irregular shapes, improve reproducibility, and efficiently share data between researchers.

Acknowledgements. This work was supported in part by the National Science Foundation (grant IIS-9980166). For more information on the PRISM 3DK project visit http://3dk.asu.edu. The authors would like to thank various members of the PRISM lab, in particular Ms. Mary Zhu for help with the figures.

References

1. Razdan A., Liu D., Bae M., Zhu M., Farin G., Simon A., Henderson M.: Using Geometric Modeling for Archiving and Searching 3D Archaeological Vessels. CISST 2001, Las Vegas.
2. Rowe J.: A Model Digital Library for 3D Pottery Data. Coalition for Networked Information Spring 2001, Washington, D.C.
3. Rowe J., Razdan A., Collins D., Panchanathan S.: A 3D Digital Library System: Capture, Analysis, Query, and Display, 4th International Conference on Digital Libraries (ICADL), Bangalore India, 2001.

INEXP: Information Exchange Protocol for Interoperability

Hongzhi Wang, Jianzhong Li, and Zhenying He

Department of Computer Science and Technology
Harbin Institute of Technology
Harbin, China, 150001
whongzhi@0451.com lijz@banner.hl.cninfo.net
hzy_hit_cn@sina.com

Abstract. Interoperability is required in digital libraries to enable them to query a large number of distributed, heterogeneous and autonomous data sources. In this paper, INEXP, an information exchange protocol for mediator-based interoperability, is presented. INEXP provides for the communication between mediator and wrapper, including query to data and schema, control command and result delivery. The protocol is designed to make the communication effective, efficient and secure. Some implementation issues are considered.

1 Introduction

Modern digital libraries need to use information from not only local, homogeneous data sources, but also heterogeneous data sources in distributed environments, even unknown data sources hidden on the web. In other words, interoperability is required.

Our contribution in this paper is that we present a protocol for information exchange between mediator-wrapper/proxy and proxy-wrapper. We give the format of the communication, and security and efficiency are considered. Our protocol uses XML as the basic format of data exchange and representation. This brings new problems, as the size of data to communication increases because of the repetitive tags of XML.

2 Protocol Details

As a protocol in application level, effectiveness and security should be considered for constituting the protocol. Three kinds of sub-protocols should be defined for the wrapper: data-query protocol: the protocol for querying wrapper for data and wrapper returning data to mediator; schema-query protocol: the protocol for wrapper and mediator querying for schema and returning schema as result; control command protocol: the protocol for mediator and wrapper exchange command. (Details on each sub-protocol can be found in [3].)

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 421-422, 2002.

Keeping the mediator and wrapper communication in a uniform transfer channel, the first byte of every package of the protocol is to identify the type for the package.

Issue of Encryption. To ensure the security of the communication in distributed environments, the protocol supports encryption of queries and data. The IDEA algorithm [1] is used for encryption. Mediator and wrapper have a common privacy key to protect the queries of users and the system. There are two encryption levels of the answer of users. The Diffie-Hellman [1] algorithm is chosen to generate the cryptographic key.

Issue of Compression. The information retrieved by the wrapper may be so massive that the communication time is too slow for users to tolerate. In this instance, the wrapper must compress the results.

The strategy discussed in [2] is used to compress the result. It is improved to satisfy the needs of this application. The data of communication is a data stream and they are treated one by one in order.

Considering efficiency, the result is not clustered while the data in the same schema is compressed as one cluster in [2]. The center is chosen randomly. The order of wrapper sending the XML result is called sending order, which is decided by the wrapper. The □between two near XML result in sending order is computed. □$_{ij}$ is the edit distance between x_i and x_j in sending order. The sending format of XML documents with sending order {x_1, x_2, x_3, ……x_n} is x_1□$_{12}$□$_{23}$……□$_{n-1n}$.

3 Conclusion and Future Work

In this paper, we have described a simple but very powerful and flexible protocol, INEXP, which provides for the communication between components of interoperability in digital libraries. We have considered some issues about the communication, including format, semantics, security and efficiency.

The next step is to perform more experiments to test the capability of the protocol. Another problem to consider is the fusion of INEXP and the whole digital library.

References

1. Bruce Schneier: Applied Cryptography: Protocols, algorithms, and source code in C (Second Edition). John Wiley & Sons, Inc (1996).
2. Hongzhi Wang, Jianzhong Li, Zhenying He: A Storage Strategy for Compress XML Warehouse. NDBC(2002).
3. Hongzhi Wang, Jianzhong Li, Zhenying He: INEXP: Information Exchange Protocol for Interoperability. Technical Report of Harbin Institute of Technology.

Study on Data Placement and Access Path Selection in an FC-SAN Virtual Storage Environment[1]

Chao Li, Chun-xiao Xing, and Li-zhu Zhou

Department of Computer Science and Technology
Tsinghua University, Beijing, 100084, China
li_chao00@mails.tsinghua.edu.cn,
{xingcx,dcszlz}@mail.tsinghua.edu.cn

Abstract. Because of the limitations of DAS (Directly Attached Storage), network storage methods (e.g. FC-SAN) with virtual storage technology have been replacing DAS in digital libraries and other massive storage applications. But in some cases, FC-SAN may perform no better than sharing data on LAN. This paper highlights some criteria and gives recommendations for data placement and access path selection, for a storage system of co-existing FC-SAN and DAS.

1 Introduction

Today the traditional storage method DAS can no longer meet the dramatically increasing requirements of digital libraries. Network storage methods (e.g. FC-SAN) with virtual storage technology have been replacing DAS [1][2]. In developing digital libraries supported by FC-SAN virtual storage systems, its dominant easy-to-manage and easy-to-use feature of virtual storage often leads to the tendency to deploy it in a digital library to achieve better performance. In fact, this practice strongly depends on some attributes of stored documents. In some cases, FC-SAN may perform no better than sharing data on LAN.

We carried out a study on data placement and access path selection in an FC-SAN virtual storage environment. The paper first presents a linear time-consuming model of data access through the analysis of the virtual storage principle. Then the concept of *equivalent of virtual storage cost* is defined to evaluate the cost paid in FC-SAN virtual storage environment. Lastly a decision-making method is given for data placement and access path selection.

2 Main Idea and Conclusions

Based on the analysis of the virtual storage principle in FC-SAN environment [3][4], we work out the linear time-consuming model about data access:

[1] Supported by the National Natural Science Foundation of China under Grant No. 60221120146; the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 423-424, 2002.

$$T_1=aL_{DATA}+C_1,\quad T_2=aL_{DATA}+C_2,\quad T_3=bL_{DATA}+C_3\ . \qquad (1)$$

Here, T_1 is the time spent in data access by MDC [4], T_2 is the time spent in data access by Host [4] via SAN, and T_3 is the time spent in data access by Host via LAN. C_1, C_2, C_3, a, and b are constants, with $C_2 < C_1 < C_3$ and a << b. L_{DATA} is the size of the accessed data. In Fig.1. the thick black line shows the access path.

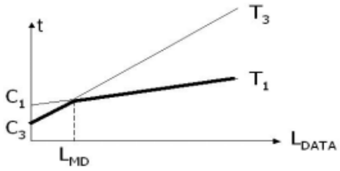


Fig. 1. L_{MD} and Access Path Selection

Our study shows that the value of L_{MD} in Fig.1. reflects the cost paid in FC-SAN virtual storage environment, and we call it the *equivalent of virtual storage cost*. With the concept of *equivalent of virtual storage cost*, the decision-making method is developed: (1) When the size of one data file is far greater than the *equivalent of virtual storage cost*, we choose to store this data file in the FC-SAN virtual storage environment; and for the most frequently accessed data of a storage server we should designate MDC on the server itself. (2) When the size of one data file is less than or near to the *equivalent of virtual storage cost*, we choose to store this data file in the traditional storage methods, such as DAS or sharing data on LAN. (3) As for Hosts, when $L_{DATA} < L_{MD}$, take the way of accessing data through LAN connections relayed by MDC; and when $L_{DATA} > L_{MD}$, just directly through FC connections.

Our experimental results indicate that this method is a simple and practical way for data placement and access path selection in a massive storage system of coexisting FC-SAN and DAS to achieve better performance.

References

1. Clark, Tom.: Design Storage area network. Addison Wesley Longman, Inc, 1999
2. Marc Farley.: Building Storage Networks. Chinese edition by China Machine Press, 2000
3. Storage Networking Industry Association. http://www.snia.org/
4. Tivoli SANergy Administrator's Guide Version 2 Release 2

Building an OAI-Based Union Catalog for the National Digital Archives Program in Taiwan

Chao-chen Chen[1] and Hsueh-hua Chen[2]

[1] Professor, Graduate Institute of Library and Information Science,
National Taiwan Normal University
cc4073@cc.ntnu.edu.tw
[2] Professor, Department of Library and Information Science,
National Taiwan University
sherry@ccms.ntu.edu.tw

Abstract. On January 1st 2002, the National Science Council of Taiwan launched a National Digital Archives Program (NDAP). To share the digital collections of all the archive participants, search via a union interface, and allow the general public access to the collections, it is urgent to build a union catalog of the National Digital Archives. In this article, we define its functions and the system architecture, and explain the problems we encountered in developing the OAI-based union catalog system.

1 Introduction

The National Digital Archives Program (NDAP) has, as participants, Academia Sinica, National Taiwan University, National Central Library, National Palace Museum, National Museum of Natural Science, National Museum of History, Historica Sinica, Taiwan Historica, and dozens of other academic groups. How the digital resources of the participants should be shared, how to allow searching of the entire archives from one single interface, and how to show the public the entirety of the archives, are very important issues.

To share the digital resources built by the participants, the construction of a union catalog is of priority. The next question is how to show the digital resources (fulltext, image, sound, and visual) through metadata. A union catalog can be built on two models: a collective union catalog or a distributed virtual union catalog. The former has the advantage of offering better search results, but has a high construction cost.[1] The latter has a low construction cost, but poor search results. To retain the advantages of both models and avoid their drawbacks, a new protocol for distributed harvesting of metadata – Open Archives Initiative (OAI) was born in the digital era. Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides a simple solution for harvesting the metadata of different databases automatically, in batches, or in different distributions, and also constructs a collective union catalog.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 425-426, 2002.

2 OAI-Based Union Catalog for the NDAP

The program invited several participant representatives to form the OAI test-bed team, to build the NDAP union catalog with OAI-PMH, and to handle system technology. Although OAI-PMH is a simple and easily designed protocol, some problems have not yet been considered in the actual union catalog system design. For example, how should the databases of different units be connected, how to convert metadata of different formats to Dublin Core, how to harvest digital objects through metadata, and how to design the data service end interface. Figures 1 and 2 show the system structures of the service and data providers that we have defined for the national union catalog.

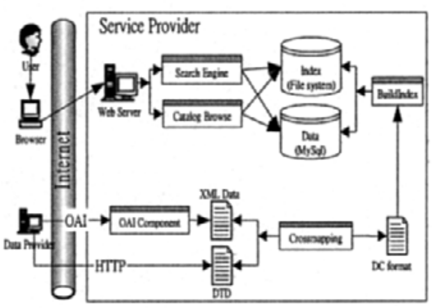


Fig. 1. System structure of service provider

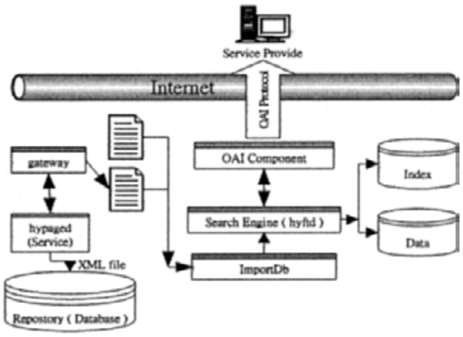


Fig. 2. System structure of data provider

3 Problems Encountered

For the union catalog, participants have to convert their metadata to Dublin Core and XML, and then deposit them in the Data Provider. Without consensus about the format and content value of metadata, most institutions used various formats to develop their systems and consequently have difficulty following completely the above procedure. This makes it hard for the Service Provider to harvest metadata and causes imprecise results when searching the union catalog.

Acknowledgements. We would like to acknowledge the support for this project from the National Science Council of Taiwan under contract NSC 91-2422-H-001-3203. We also thank Mr. Ya-ning Chen and Ms. Shu-jiun Chen for their advice, and all the research assistants of this project: Kuo-shen Su, Ming-jie Liu, Huei Oyang, Chen-ji Chang, Hsing-ron Pan, Hsueh-yun Yian, Jao-ru Chen, Hsiao-ying Hung, Huai-wen Chang, and Tai-hsin Hsu.

Intergenerational Partnerships in the Design of a Digital Library of Geography Examination Resources

Yin-Leng Theng[1], Dion Hoe-Lian Goh[1], Ee-Peng Lim[2],
Zehua Liu[2], Natalie Lee-San Pang[1],
Patricia Bao-Bao Wong[1], and Lian-Heong Chua[1]

[1]Division of Information Studies
School of Communication and Information
Nanyang Technological University (Singapore)
{tyltheng,ashlgoh,nataliepang,patriciawong,clhchua}@ntu.edu.sg

[2]Centre for Advanced Information Systems
School of Computer Engineering
Nanyang Technological University (Singapore)
{aseplim,aszhliu}@ntu.edu.sg

Abstract. This paper describes the engagement of intergenerational partners in the design of a digital library of geographical resources (GeogDL) to help prepare Singapore students for a national examination in geography. GeogDL is built on top of G-Portal, a digital library providing services over geospatial and georeferenced Web content. Scenario-based design and claims analysis were employed as a means of refinement to the initial design of the GeogDL prototype.

1 Introduction

Having completed the first phase of development of GeogDL (digital library of geography examination resources), a study was conducted to engage a group of intergenerational partners involving designers, secondary school students and usability-trained evaluators for the purposes of reinforcing and/or refining the initial design of GeogDL. GeogDL [2] is a Web-based digital library application containing past-year examination questions and solutions, supplemented with additional geographical content.

We wanted to design GeogDL with and for students taking a Singapore national examination in geography (the GCE 'O' level geography examination) with a strong underpinning user-centred design rationale. GeogDL is built above G-Portal [6], a digital library providing services over geospatial and georeferenced Web content.

Beyond summarizing the design of GeogDL, a main contribution of the paper is making explicit the use of Carroll's scenario-based design and claims analysis [1] that inspired recommendations for the refinement of the initial design of GeogDL.

The remainder of this paper describes the study and discusses the implications of the findings in relation to design and implementation issues for GeogDL as well as geospatial digital libraries (DLs) in general.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 427-439, 2002.

2 GeogDL: Design Philosophy and Initial Design Choices

In this section, we briefly revisit our previous work on G-Portal and GeogDL so that their methods and findings can provide a background for the body of this paper and the issues explored within it.

2.1 G-Portal

G-Portal [6] is an on-going DL project at the Centre for Advanced Information Systems in Nanyang Technological University (Singapore). The aims of the project include identification, classification and organization of geospatial and georeferenced content on the Web, and the provision of digital services such as searching and visualization. In addition, authorized users may also contribute resources so that G-Portal becomes a common environment for knowledge sharing. G-Portal resources are defined as Web content, annotations and metadata.

G-Portal also provides a platform for building applications that use geospatial and georeferenced content. This is achieved through *projects* which are user-defined collections of related resources. Resources within projects are further organized into *layers* which allow finer grained organization of content.

2.2 GeogDL

G-Portal is used to build our first DL on geography examination resources (GeogDL). GeogDL [2] is not meant to be a replacement for textbooks and classroom education, but an alternative to printed past-year examination solutions developed to help students revise for their GCE ‘O’ level geography examination.

In GeogDL, past-year examination questions (with their solutions) are created as separate G-Portal projects. Each project consists of Web resources, at least one of which contains the solution to the question. Other resources contain information to related topics and are used as supplementary material for further exploration. Resources may be further organized into layers depending on the needs of the teacher. For example, the solution to an equatorial region question could appear as a resource in a layer while a separate layer might contain supplementary vegetation resources found in equatorial climates.

In the initial version of GeogDL, examination questions are first accessed through the classification interface that organizes questions by year. Upon selection of a question, the associated project, its resources, and the corresponding map are loaded. Currently, resources are divided into three categories: question, solution and supplementary resources, each of which is accessible separately via the classification interface.

3 Scenario-Based Design and Claims Analysis

Our study was inspired by Carroll's work on the task-artifact cycle, user-centred strategies such as scenario-based design and claims analysis [1].

The task-artifact cycle explains why design is never completely "done". At the start of any software development, tasks help articulate requirements to build artifacts, but designed artifacts create possibilities (and limitations) that redefine tasks. Hence, managing the task-artifact cycle is not a linear endeavour with starting and ending points [1]. There will always be a further development, a subsequent version, a redesign, a new technology development context. The design scenarios at one point in time are the requirements scenarios at the next point in time. Carroll [1] stresses the importance of maintaining a continuous focus on situations of and consequences for human work and activity to promote learning about the structure and dynamics of problem domains, thus seeing usage situations from different perspectives, and managing tradeoffs to reach usable and effective design outcomes.

Claims analysis was later developed by Carroll [1] to enlarge the scope and ambition of scenario-based design approach to provide for more detailed and focused reasoning. Norman's influential model of interaction [7] is used as a framework in claims analysis for questioning the user's stages of action when interacting with a system in terms of goals, planning, execution, interpretation and evaluation.

3.1 Experimental Protocol

We engaged a group of intergenerational partners involving secondary school students, designers and usability-trained evaluators. The concept of intergenerational partnership, in which design partners of varying ages, needs, expectations and experience negotiate design decisions, is especially crucial in systems designed for children and teenagers [e.g. 3; 9; etc.]. One of the challenges of this kind of partnership is for children/teenager users to trust adult designers to listen to their contributions. Druin et. al. [3] found that this kind of idea-elaboration process takes time to develop, but they found it to be extremely important to work towards in a design partnership [3], and hence towards a better design that would cater to the needs of the prospective children/teenager users.

Brainstorming session among usability-trained evaluators and designers

Four usability-trained evaluators were involved in the study. Two of the evaluators were Masters of Information Studies students at Nanyang Technological University (NTU, Singapore) who had completed a course on Human-Computer Interaction (HCI) with a working knowledge on scenario-based design and claims analysis. The other two evaluators were lecturers at NTU who taught HCI and Systems Analysis/Design respectively. Since there is little literature available on the practicalities of applying claims analysis to evaluate and improve the usability of DLs [5], the evaluators met for a brainstorming session prior to the sessions with the student design partners to make concrete and agree upon the procedures in carrying out claims analysis [1].

Identifying possible goals or scenarios of use of GeogDL
To situate claims analysis within the context of use, the session began with the evaluators identifying the possible goals or scenarios of use prospective users might have when using GeogDL.

Ellington et. al. [4] propose four basic factors to match the natural learning processes of humans, and thus ensure the successful learning experiences of learners by: (F1) making learners *want to learn*; (F2) incorporating sufficient activities to help learners experience *learning by doing*; (F3) providing sufficient channels of *feedback* to learners; and (F4) enabling learners to *digest and relate* what they have learned to the real world.

Since the main goal of GeogDL is to help students prepare or revise for the GCE 'O' level geography examination, the following sub-goals were postulated to provide the possible scenarios of use with the inclusion of the four basic factors proposed by Ellington et. al. [4] for successful learning experiences of learners:

- *Goal #1: Practice/revision* on multiple-choice (MCQs), short structured and essay-type questions. Model answers and hints to tackle these questions should also be provided (applying F2). Feedback should be provided (applying F3).
- *Goal #2: Trends analysis*. The idea is to give information on when and what questions are being asked over the years. This would help students identify trends in the types of questions asked and the topics covered. This may increase their motivation to want to learn (applying F1).
- *Goal #3: Mock exam*. This would help students better manage their time in answering questions. To make it fun, a scoring system could be incorporated for MCQs (applying F4), while hints/model answers could be provided for structured and essay questions (applying F3).
- *Goal #4: Related links and resources*. This could include related topics, teachers' recommendations, etc., thus showing relationships of concepts, and linking concepts to the real world (applying F4).

To protect against potential distortion of the scenarios, the above four sub-goals or scenarios of use were validated with the two designers of GeogDL. Designer 1 was in charge of the architecture of G-Portal; while Designer 2 was in charge of populating GeogDL with geography examination resources.

At the time of carrying out this study, only Goals #1 and #2 were implemented. Goals #3 and #4 are currently being implemented. Therefore, for the purpose of this study, we were only interested to examine GeogDL in terms of Goals #1 and #2. In identifying the scenarios of use with good coverage and minimal bias, we made use of the participatory design approach where prospective users were involved as design partners.

Modifying questions as used in Claims Analysis
Space constraints, however, do not permit us to write in detail the changes made to the questions tailored for the specific goals. Based on the four goals identified, the evaluators modified the original nineteen questions formulated by Carroll [1] so as to "speak the students' language" and to make them more relevant to the specific goals in question. For example, the original question "How does the artifact evoke goals in the user?" was modified to reflect Goal #2 (Trends analysis), and was changed to

“How does the system (screen) help you to decide what to do to analyse trends or spot questions?”

Sessions with student design partners
A group of eight secondary students (ages between 13 – 15 years old), consisting of four boys and four girls, were invited as design partners. The purpose of the session was to reinforce the initial design and/or gain insights from what the student design partners said they wanted or what they wanted, as a means of refinement of the initial design. The session with the four girls was held in the morning while the session with the four boys in the afternoon, each lasting approximately two hours. Every student was assigned to one usability evaluator, and they were asked to carry out claims analysis on either Goals #1 or #2.

The session was divided into three parts. Part 1 began with getting to know the students in terms of their experience with Web-based interface, searching/browsing skills and study habits. The interview session ended with a discussion on the possible scenarios of use for students preparing for the GCE ‘O’ level geography examination. Part 1 lasted approximately forty-five minutes. The evaluators stepped through GeogDL with the students responding to the stages of actions when interacting with GeogDL in Part 2 of the session. They were asked to identify the positive outcomes as well as negative consequences of the features provided in GeogDL in supporting either Goals #1 or #2. Part 2 also lasted approximately forty-five minutes. In Part 3, all four students together with the four evaluators congregated for a focus group discussion. The purpose was to confirm and/or refine the four goals identified by the evaluators described earlier, and brainstorm, if any, other goals that students might have when preparing/revising for GCE ‘O’ level geography examination.

3.2 Findings and Analyses

3.2.1 Profiles, Study Habits, and Scenarios of Use

Students’ Profiles
Our student design partners came from a local secondary school in Singapore and would form a representative sample of prospective users, according to a secondary school teacher who was also one of the evaluators involved in this study.

We wanted to capture students’ profiles to help us understand, for example, not only what they said they liked about a certain feature, but also why they said they liked it. Studies have shown users’ backgrounds in terms of their experience with Web-based interface and searching/browsing skills might affect their acceptance of a system [5]. Since GeogDL aims to provide users with a successful learning experience, an understanding of the subjects’ study habits, in particular, examination techniques adopted would also be useful.

Boys
The boys (denoted as S1 to S4) were between 13 – 14 years old, and were generally more confident Web users compared to the girls. They rated themselves as

intermediate to advanced users spending a considerable amount of time everyday on the Web, ranging between two to six hours, playing games, emailing or chatting with friends. Except for S2, all believed that their searching/browsing skills commensurate with their usage of the Web. S2, though a self-believed advanced Web user, thought of himself a novice in searching/browsing on the Web. The boys rated themselves as novice or intermediate in terms of library searching/browsing skills.

Girls

Although the girls were one year older than the boys, they were comparatively less confident Web users. The reason, according to a teacher of the school, was that the girls did not have the benefits of being introduced to simple HTML/XML programming in the revised lower secondary curriculum. The girls (denoted as S5 – S8) rated themselves as novice or intermediate users of the Web. They used the Web mainly for emailing or chatting with friends. Except S5 who rated herself "intermediate", the rest of the girls rated themselves novices and commented that their searching/browsing skills were "poor". Similar to the boys, library searching/ browsing skills were not good, ranging from novice to intermediate.

Study Habits

In general, the students were less motivated to explore beyond what was required of the syllabus. All the students relied heavily on textbooks, exam questions with model answers, teachers' worksheets and notes taken during lessons to prepare for exams. In particular for geography, atlases and maps were constantly referred to.

Scenarios of Use

Enumerating typical and critical use scenarios characterizes the scope of an artifact's actual use or the anticipated use of an artifact still in design [1]. The students reinforced the relevance of the four goals identified by the evaluators to achieve the main goal of preparing/revising for GCE 'O' level geography examination. As suggested in Carroll's task-artifact cycle hypothesis, the GeogDL artifact also provided a platform for students to add on/modify the goals of GeogDL. Because the scenarios provided a working representation for exploring and altering the design, the students also saw GeogDL not only as an examination resource DL, but also as an interactive teaching aid.

3.2.2 Stages of Actions and Design Consequences

Since we were interested in how users complete a task successfully, we made use of the method "questioning stages of actions" to elicit claims about the design of GeogDL. In this method, theories of human activity were thought to be effective in facilitating systematic questioning. Based on Norman's execution-evaluation cycle, Carroll [1] developed a set of questions as a heuristic for comprehensively interrogating the tradeoffs implicit in scenarios. We modified the original set of questions designed by Carroll [1] to make them specific to the goals in question and also in simpler English so that the student designers could understand the questions.

Capturing and analyzing students' responses

The students performed iterative walkthroughs of the system together with the respective evaluators to achieve Goals #1 or #2. This was done by questioning stages of actions in Norman's execution-evaluation model of task completion broadly divided into these three phases [1]:

- *Before executing an action.* This phase intends to prompt claims on the design before users perform an action. Two stages of users' actions that address formation of goals (Stage 1a) and planning (Stage 1b) are involved. A total of seven questions were used to prompt claims.
- *When executing an action.* This phase (Stage 2) obtains claims by questioning users on how well the system helps them to perform the action. We used two questions instead of the original three in Carroll's set because we felt that one of the questions was redundant.
- *After executing an action.* Two stages (Stage 3a and 3b) prompt users to interpret system's response and evaluate the system's effectiveness in helping to complete a goal. We appended to the original list questions that address also Nielsen's well-established design heuristics. A total of twelve questions were asked.

For each scenario of use, evaluators helped the students to step through the above five stages by framing their goals (Goals #1 or #2 in our study), taking action, interpreting the consequences of their actions, and evaluating action consequences with respect to the instigating goals.

Owing to space constraints, we are not able to show all eight students' responses to the twenty-one questions for all the five stages. As an illustration, Table 1 shows S6's comments in response to the three questions asked in the Goal Stage (Stage 1a) for Goal #2 (Trends analysis). Columns 2 and 3 record S6's claims highlighting positive consequences or negative consequences/risks respectively. The rest of the students' responses were constructed in this manner.

Table 1. Stage 1a (Goal Stage): Student S6's for Goal #2 – Trends Analysis

Stage	Positive Consequences	Negative Consequences
Stage 1a: Goal Stage Questions to prompt: 1. How does the system (screen) help you to decide what to do? 2. How does the system (screen) help you to want to analyse trends or spot questions? 3. How does the system (screen) suggest that spotting questions is: - simple or difficult? - appropriate or inappropriate?	Comments: Statement on the occurrence of the question in the past years helps me to get a vague idea of the question's frequency. Compliance - Feature: Linking of related concepts	Comments: No references to the map. Violation - Feature: Linking of related concepts Comments: I have no idea how to use statement of occurrence to spot question. Violation - Feature: Match between system and real world Comments: Too many windows opened which causes confusion. Violation - Feature: Minimalist design

Table 2. Desirable – Features with Positive Consequences

No.	Features	Positive Consequences							
		Boys				Girls			
		S1	S2	S3	S4	S5	S6	S7	S8
1	Diagnosis and recovery from errors	Not applicable since students did not encounter errors in their interactions.							
2	Visibility of systems status	1a	1a	1b, 2	3a	2, 3a	2		
3	Match between system and real world				3a	3a			
4	Control and freedom for users	1a, 1b	1a, 1b						
5	Consistency and standards			3a			3a		
6	Recognition rather than recall	3a		1b	1b				
7	Flexibility and efficiency of use		3a		1b, 2, 3a			2	2
8	Minimalist design								
9	Speak the users' language			3a		1a	3b		
10	Help and documentation						3b		
11	Provide shortcuts								
12	Links to related concepts					1a, 3a	1a		

Table 3. Undesirable - Features with negative consequences

No.	Features	Negative Consequences							
		Boys						Girls	
		S1	S2	S3	S4	S5	S6	S7	S8
1	Diagnosis and recovery from errors	Not applicable since students did not enter errors in their interactions.							
2	Visibility of systems status	1b	1b		3b	2, 3b		1a	1a
3	Match between system and real world			3b	1a, 3a	1b, 3b	1a		
4	Control and freedom for users	1a	1a						
5	Consistency and standards			3a, 3b	3a, 3b	3a, 3b	3a, 3b		
6	Recognition rather than recall			3b		3b	3b		
7	Flexibility and efficiency of use	1a	1a	1a, 2, 3b	1a, 1b, 2, 3a, 3b	1a, 1b, 3b	2	1a, 1b, 2b, 3b	1a, 1b, 2b, 3b
8	Minimalist design			1a, 1b	1a, 3b	1b, 3b	1a	1a	1a
9	Speak the users' language	1b	1b		1a, 1b, 3b	3a, 3b	1b	1b	1b
10	Help and documentation			2, 3b	3b	3b	2		
11	Provide shortcuts	1b		2, 3b	3b	3b	3b		
12	Links to related concepts		1b	3a, 3b	1a, 1b, 3a, 3b	1a, 1b, 3a, 3b	1a, 3a		

Analyzing design consequences

Since students' comments were made in response to the design of GeogDL where the method of operation was not fully predictable, and where the students were not completely novices in the use of Web-based interactive systems, we turned to the following well-accepted design heuristics to categorize students' comments [e.g. 8; etc.]. We made these assumptions: students' comments with positive consequences suggest compliance with the design heuristics (see Table 1, Column 2); while comments with negative consequences/risks indicate violation of design heuristics (see Table 1, Column 3).

By categorizing all eight students' comments in this manner, a list of claims with positive outcomes in relation to design heuristics was generated (see Table 2). Table 3 shows combined students' comments on the negative consequences/risks violating design heuristics, obtained from similar tables like Table 1. Unless properly dealt with, negative consequences/risks could potentially affect usability of a system [1]. Section 4 discusses recommendations made to GeogDL to eliminate or at least alleviate the negative consequences or risks imposed by these current features that might hinder the completion of Goals #1 and 2.

4 From Analysis to Refinement

In this section, we identified areas for refinement grouped according to violations against the following design heuristics (see Table 3):

1. *Diagnosis and recovery from error.*
   Students' Comments: No comments from students since we did not encounter errors. Comments such as "don't know what to do or how to proceed" were common.
   Recommendations: An examination of GeogDL showed that no error messages were provided. Error messages should be clear, indicating precisely the problem, and constructively suggesting a solution.
2. *Visibility of system status.*
   Students' Comments: "I'm not sure if I have completed my goal"; etc.
   Recommendations: The system should always keep users informed of what is going on through appropriate feedback within reasonable time. The student was not sure whether she had already accomplished the goal. She expected something different and not just a question with a phrase to signify the types of questions asked.
3. *Match between system and real world.*
   Students' Comments: "I'm not sure what to do"; "lack of a legend on the map, which failed to provide linkages to topics"; "mouse-over text is also missing to provide context to potential mouse clicks"; etc.
   Recommendations: Follow real-world conventions, making information appear in a natural and logical order. Instead of a map-based interface only, a list of questions could be created also as a point of access to GeogDL. The map should not be the main window. There could be graphical representations of occurrences of questions, and information should be organized by topics. A legend should be provided on the map.
4. *Control and freedom for users.*
   Students' Comments: "Lack of a clear map between different features in the system (e.g. questions and relationship to map)"; "don't know how to exit"; etc.

Recommendations: Users often choose system functions by mistake and will. They need a clearly marked “emergency exit” to leave the unwanted state without having to go through an extended dialogue. Perhaps an explorer-like presentation to organize different information and content could be implemented in GeogDL. Users would be familiar with its use, and also be able to tell at a glance, the relationships between different functions in GeogDL.

5. *Consistency and standards.*
   Students’ Comments: “Links are not designed using Web formats”
   Recommendations: Users should not have to wonder whether different words, situations, or actions mean the same thing. Follow platform convention. Recommend that GeogDL be designed using the standards of the Web – as it is perceived by the users that GeogDL is a Web-based system (using Internet Explorer to access the system). Icons and taxonomy used should also be that of the Microsoft Windows environment to increase acceptance and familiarity.
6. *Recognition rather than recall.*
   Students’ Comments: “I don’t know how to start using GeogDL”; etc.
   Recommendations: Make objects, actions and options visible. The user should not have to remember information from one part of the dialogue to another. Instructions for use of the system should be visible or easily retrievable whenever appropriate. Students were unable to identify with the newness of the geospatial-like interface in GeogDL. Perhaps a virtual tour of the system would be useful, which can also be supported with careful implementation and training.
7. *Flexibility and efficiency of use*.
   Students’ Comments: “There is a lack of instructions and explanatory notes to help me to navigate”; “No indication that it is the final screen. Found the question window by mistake”; “Overlapping windows causes confusion”; etc.
   Recommendations: Help could be provided to users by giving instructions and explanatory notes. GeogDL should also provide feedback to users when the final screen has been reached by providing ‘previous’ or ‘next’ buttons. Re-design interface such that windows are neatly arranged to make GeogDL more efficient and flexible to use. Fig. 1 is a recommendation for a revised interface to GeogDL by tiling the windows neatly, and also making the map-based and classification interfaces prominent as equal points of access to GeogDL. Accelerators, unseen by the novice users, may often speed up the interaction for expert users to cater systems to both inexperienced and experienced users.
8. *Minimalist design.*
   Students’ Comments: “Too many windows opened, causes confusion”; etc.
   Recommendations: Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility. Improve the design by integrating certain functions together in one window. Please see Fig. 1 for a recommended revised interface to GeogDL.
9. *Speak the user’s language.*
   Students’ Comments: “I don’t understand what windows ‘layers’ do”; etc.
   Recommendations: System should speak the user’s language with words, phrases and concepts familiar to the user, rather than using system-oriented terms. Use “legend” instead since this term is familiar to geography students used to reading maps and atlas.

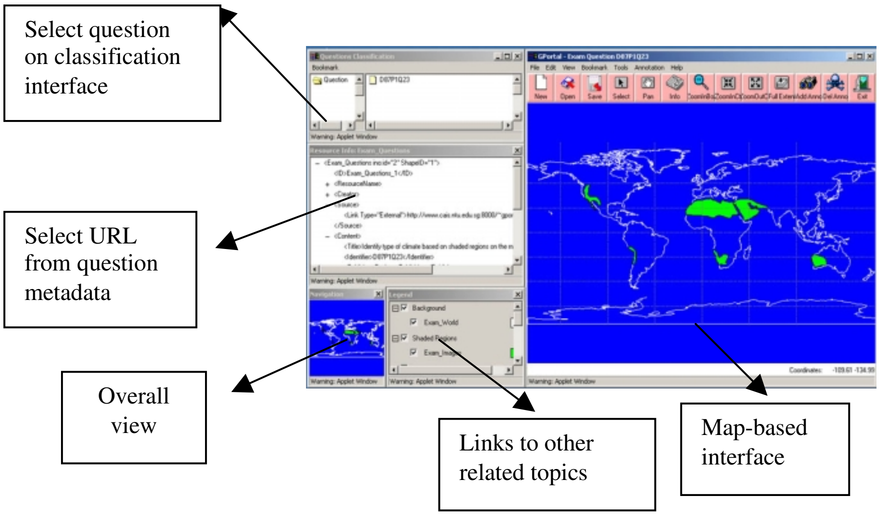

Select question on classification interface
Select URL from question metadata
Overall view
Links to other related topics
Map-based interface

Fig. 1. One of the recommendations for improving the interface addressing the violations against design heuristics "flexibility and efficiency of use" & "minimalist design"

10. *Help and documentation.*
    Students' Comments: "There is a lack of help and documentation"; etc.
    Recommendations: Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large. A virtual tour of the whole system could aid users' exploration and familiarization with the system. Help mascot, as suggested by the students, could monitor and guide users' actions.
11. *Provide shortcuts.*
    Students' Comments: "No shortcuts available for more experienced users"; etc.
    Recommendations: The features that make a system easy to learn such as verbose dialogues and few entry fields on display are often cumbersome to experienced users. Clever shortcuts, unseen by novice users, may often be included in a system such that the system caters to both experienced and inexperienced users. Shortcut buttons/quick jump menu could be designed for experienced users.
12. *Links to related concepts.*
    Students' Comments: "No references are made to the map"; "Climate identification on map is not related to similar topics, questions, and has no references to links"; "No other links from the questions exist, to also prompt for further exploration"; etc.
    Recommendations: To help users achieve a successful learning experience, not only should information appear in a natural and logical order, inter-connectivity between concepts should also be captured. Perhaps there should be links and references to the map. The map interface should also tell users 'where' they are. This is to allow users to see in an organized fashion the organization and taxonomy of GeogDL map. A suggested list of related links from each section for

further exploration by users could be created. Also, links should be provided to the model solutions of questions, and to tips from teachers in answering such questions/similar questions; etc.

5 Conclusions and On-Going Work

This paper described the engagement of intergenerational partners and the novel use of scenario-based design and claims analysis as a means of refinement to the initial design of the GeogDL prototype. The study also showed that through a process of aggregation, a team of eight design partners could produce a comprehensive, rich set of data, of which we presented only some of the findings in this paper.

This is on-going work for us. The initial work has created useful findings to refine the initial design of the GeogDL prototype. It will be interesting to repeat this work with other age groups and control for factors such as Web skills, gender and study habits/preferences.

Compared to other forms of usability evaluation, say heuristic evaluation, claims analysis is powerful and more strongly theory-based. In our study, we showed how usability problems could be detected by analyzing claims made by users stepping through stages of actions in Norman's execution-evaluation cycle model of task completion. Claims sharpen the understanding of relationships that may only be suggested by the scenarios themselves [1], highlighting just how GeogDL in use affords actions, suggests explanations, signals progress and highlights problems for refinement. Unlike other usability evaluations, claims analysis situated in the context of use together with the emphasis to generate likely scenarios, make evaluators focus not only on problems but also on solutions.

However, Carroll's claims analysis is not intuitive to use since the questions to prompt claims are quite difficult to understand, and using it well requires a competent level of "craft skills". More can be done to make scenario-based design and claims analysis practical and easy to use.

Acknowledgements. The authors would like to thank the students from Yio Chu Kang Secondary School for their participation and ideas. The project is partially supported by the SingAREN 21 research grant M48020004.

References

1. Carroll, J.: Making use: Scenario-based Design of Human-Computer Interactions. The MIT Press. (2000).
2. Chua, L.H., Goh, D., Lim, E.P., Liu, Z., Ang, R. A Digital Library For Geography Examination Resources. Proceedings of the Second ACM+IEEE Joint Conference on Digital Libraries, pp. 115-116. (2002).
3. Druin, A., Bederson, B., Hourcade, J.P., Sherman, L., Revelle, G., Platner, M., and Wong, S.: Designing a Digital Library for Young Children : An Intergenerational Partnership. Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 398 – 405. (2001).

4. Ellington, H., Percival, F. and Race, P.: Handbook of Educational Technology. Kogan Page. (1995).
5. Keith, S., Blandford, A., Fields, B. and Theng, Y.L.: An Investigation into the Application of Claims Analysis to Evaluate Usability of a Digital Library Interface. Accepted for Workshop on “Usability of Digital Libraries” in the Second ACM+IEEE Joint Conference on Digital Libraries (Portland, Oregon, July 2002).
6. Lim, E.P., Goh, D., Liu, Z., Ng, W.K., Khoo, C., Higgins, S.E. G-Portal: A Map-based Digital Library for Distributed Geospatial and Georeferenced Resources. Proceedings of the Second ACM+IEEE Joint Conference on Digital Libraries. pp. 351-358. (2002).
7. Norman, D.: The Psychology of Everyday Things. Basic Books. (1988).
8. Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T.: Human-Computer Interaction. Addison-Wesley. (1994).
9. Theng, Y.L., Mohd-Nasir, N., Buchanan, G., Fields, B., Thimbleby, H. and Cassidy, N.: Dynamic Digital Libraries for Children. Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries. pp. 406 – 415. (2001).

Pie Charts for Visualizing Query Term Frequency in Search Results

Terry Anderson[1], Ali Hussam[2], Bill Plummer[2], and Nathan Jacobs[2]

[1] School of Computing and Mathematics, University of Ulster at Jordanstown,
Newtownabbey, Co. Antrim BT37 0QB, UK
tj.anderson@ulster.ac.uk
[2] Computer Engineering and Computer Science Dept., University of Missouri-Columbia,
201 Engineering Building West, Columbia, MO 65211, USA
{HussamA, PlummerB, nate}@missouri.edu

Abstract. Search engine queries are normally brief but often return unmanageably long results, with which users struggle to determine document quality and relevance. In recent years, many studies have enhanced search results with metadata displayed as visual cues. Their success in helping users make faster and more accurate document judgments has been uneven, reflecting the wide range of information needs and document selection strategies of users, and also the relative effectiveness of different visualization forms.

We identify the frequency with which query terms are found in a document as a straightforward and effective way for users to see the relationship between their query and the search results. In our prototype, we display these frequencies using simple pie charts. Despite performance limitations, evaluation with 101 users has been promising and suggests future developments.

1 Introduction

Modern Web and intranet search engine performance for document retrieval is technically impressive and functional improvements will doubtless continue. The frustratingly long lists of semi-relevant results often returned by search engines are due more to the ambiguity inherent in brief, under-specified queries than to any inadequacies of the information retrieval software. The casual user finds difficulty in employing even simple Boolean operators and strongly prefers to enter queries composed of just two or three terms or a short phrase [1] [2]. This tendency is reinforced by the responsiveness of search engines, which encourages users to formulate and reformulate queries quickly. Thus the user trades effort at the query formulation stage for the often significant effort of assessing document surrogates. Even the most carefully written query may result in an information glut. Daunted by the quantity of results, users generally resort to the simple filtering strategy of relying on their search engine's relevance scores and examining only the initial documents. Empirical results from web-based surveys suggest they typically consider the first 20 or 30 documents at most because reading the textual descriptions is a cognitively demanding task [1] [3]. Given that many potentially valuable results are likely to be ignored, Spink et al. [1] assert "We need a new generation of Web searching tools that

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 440-451, 2002.

work with people to help them persist in electronic information seeking to resolve their information problems."

Our aim is to produce a search interface that allows users to scan results more efficiently than is currently possible with conventional text output. Appropriate visual cues, removing the need to skim each document surrogate, should give searchers a quick but reliable impression of which documents might best match their queries.

This paper begins by considering how a user assesses document relevance. It then reviews a range of attempts at visualizing document space, and reflects on their varied degrees of success in relation to user search. Drawing on these two sections, we have selected key metadata to visualize. Our prototype which displays query term frequencies is then presented, followed by a user-based evaluation, conclusions and ideas for continuing work.

2 Document Relevance

80% of search engine users say they find what they want all or most of the time, according to one survey (reported in [4]). Ranking algorithms, although often impressively successful at placing highly relevant hits on the first page, can only poorly reflect the wide variety of purpose and richness of domain knowledge that underlie human relevance judgments. The problem persists of unmanageably long results from which users struggle to determine document quality and relevance. Users need effective interfaces to help them interpret and sift results swiftly.

From the limited information in a document surrogate - title, excerpted text containing query terms, URL and perhaps file type, file size, date modified and cached status - the user must determine the quality and relevance of an item. He or she must first translate an information need into a query, and the preference for using just a few terms has already been noted. However, the need may range from simple to complex and subtle. For example, a user may be looking for a very specific document, or collecting background material or just browsing to keep informed [2]. Many such task and context considerations may be lost in this need-to-query translation, but they will nevertheless continue to guide the user’s interaction and selection of material [5].

In the process of assessing document surrogates for relevance, it appears that users initially scan for their query terms. Wilkinson et al. [6] noted that users intuitively consider the ’best’ match to be where the document has the most matches to their query. Essentially they understand the ’coordination level’, the number of query terms in a document. Indeed, major search engines (including Google, Alltheweb and Altavista) now use bold font styles to make such matches prominent.

Mizzaro [5] suggests five other criteria that determine information needs: comprehensibility (the intellectual level of the text), recency, quantity of information (number of documents and/or page length), publication language, and fertility (how useful a document will be for finding other documents, e.g. number of references or links). Trustworthiness is another important criterion suggested by Sellen et al. [2], while others have identified user concern with likely download times [7].

The title and description of a document are naturally the primary means for assessing comprehensibility and publication language. Recency and quantity may be fairly obvious when a file date and size are shown. However, users are often highly resourceful in interpreting additional information from the URL [8]. They may be able to infer, depending on their internet experience and domain knowledge, location (and therefore likely download times), trustworthiness (based on domain and link to a physical organization [2], nature of content (possibly including comprehensibility and fertility) and recency (e.g. conference names often include a year date).

3 Related Work on Visualization

Spink et al. [2] suggest that visualization, because of the powerful human ability to understand graphics, may prove to be one of the most fruitful areas for generating new web searching tools. The previous material indicates the content of the visualization, but the range of possible formats is vast (see [9] [10]). Highly graphical information retrieval user interfaces have enjoyed over a decade of experimentation and inventiveness. Now, with the growth of evaluative studies, we can begin to understand why some solutions are more effective and realistic than others and to determine which ones may best help users persist in information seeking. The following section reviews a range of attempts at visualizing document space to inform the choice of visualization for this research.

3.1 Visualization: Set Level

Visualization of search engine results operates at two levels; the ’set level’ and the ’document level’ [11]. At the set level the aim is an overall visualization of many documents, typically hundreds if not thousands. Major examples include self-organizing maps, Venn chart abstractions, scattergraphs, concept clusters as physical maps or perspective walls, hierarchies seen in hyperbolic trees or cone trees, graphs, grids and spheres [9] [10].

Despite many extremely innovative and technically sophisticated visualizations at the set level, few have progressed beyond the prototype stage. There appear to be two main reasons for this. First, the communication overheads can be substantial, even though computational delays that caused bottlenecks only a few years ago are reducing rapidly as application servers become cheaper and more powerful. Second, users tend to perform better with simpler visual-spatial interfaces [13].

Users vary considerably in their spatial ability. In a study of information retrieval performance using 2D and 3D interfaces, Westerman and Cribbin [14], conclude that three-dimensional displays are more cognitively demanding, particularly when the data involved is abstract. Information about documents is inherently abstract, textual and non-visual, unlike the continuous variables of scientific visualization [10]. Sebrechts et al. [12] also reported that users had difficulty in coping with 3D navigation. Similarly, Cockburn and McKenzie [15] found that subjects were slower at locating files in cone tree visualizations than in a standard tree-view in a Windows or Mac interface. Performance deteriorated markedly as the amount of information increased, though Cockburn and McKenzie [15] also noted that many subjects liked the visualization because it helped them appreciate the overall information space. Although there may be exceptions, two dimensional representations are likely to be more effective for most information retrieval.

3.2 Visualization: Document Level

Turning to document-level visualization, the format has usually been to provide for each item a compact graphic representation that in some way conveys key information. Sometimes the text of the document surrogate is shown, sometimes reduced, sometimes omitted, but in all cases it is available, usually by clicking on the corresponding document graphic.

One of the best known attempts is TileBars [16]. Each document’s horizontal bar-shaped graphic shows the distribution of keyword terms throughout a text as tiny colour-coded blocks, so that frequency and co-occurrence of terms can be rapidly gauged. The TileBars display indicates "(i) relative document length, (ii) frequency of query terms in the document and (iii) the distribution of the terms with respect to the document and to each other" [17]. In the same paper Hearst also emphasized the importance of providing "the user with intuitive, descriptive interfaces that indicate the relationship between the query and the retrieved results". In TileBars, the visualization component is shown side by side with a one line description of each item. On a given screen, it is possible to display between 10 and 20 documents. Generating the detailed graphics involved made considerable computational demands.

Byrd [3] argues that the TileBars display is overly complex, supporting not one but two independent and sequential decisions - which documents to view and which passages to view. He himself has developed a more simple colour-coded bar representation for each document, displaying document term weights (not raw term occurrences). The user is only shown the keyword distribution (in the scrollbar area) when a document is selected. Byrd also notes that his distribution-displaying scrollbar is of limited value with short documents if the keywords in the main text display of the document are already colour highlighted. Although he tried to evaluate his visualization scheme, quantitative results were unreliable due to poor system response times. Subjective feedback, however, clearly showed that users liked the graphical representations. In situations where the visualization was not considered useful, it was unobtrusive and users were able to rely on the textual descriptions with little or no distraction. As with TileBars, Byrd’s system allows one-line information for at most a few tens of documents to be displayed on a screen.

Veerasamy and Belkin [18] developed a very compact visualization, with one screen displaying up to 150 documents, each represented as a vertical column composed of multiple clearly labelled bars, one for each query word. The height of a bar corresponded to the weight of that query word for a given document. Mouse movement over a column caused the corresponding document title to be highlighted. In a fuller evaluation-oriented study, Veerasamy and Heikes [19] found that this visualization tool helped users identify document relevance about 20% more quickly. They noted that users consulted the visualization before they read the title. In many cases the visualization alone provided enough information for users to quickly eliminate non-relevant documents.

3.3 Thumbnails

Many attempts to display documents as thumbnails or miniatures have been made. The appearance of a document can provide cues as to its nature, e.g. scholarly or popular, and remind users if they have seen it before. However, a simple miniature of the first page of a document may have limited value, given the standard templates used by many corporate web sites, online magazines and academic papers [20] [21].

Czerwinski et al. [21] evaluated a prototype thumbnail display in which the small graphics allow many thumbnails to be fully visible (perhaps 20 on average) and partially obscuring a large number of others (potentially well over 100) in a semi-3D format. Although subjects expressed a preference for the thumbnail images, they appeared able to perform about as well with textual descriptions. Ogden et al.'s [20] system, capable of showing 10 thumbnails per screen and using colour to highlight keyword locations, likewise failed to show any clear advantage of thumbnails over text. While they concluded that thumbnails will only be relevant for some user tasks, they also commented on user preference for the graphical display.

In 'Web page caricatures' Wynblatt and Benson [22] attempt to develop a visual summary of a document. The summary displays a textual description and additional document metadata including: size and number of embedded media (images, animations or audio clips) and the presence of web page and e-mail links - making the point that it is useful to know whether a document is essentially an index, with links to content, or has substantive content itself. Where a document is illustrated the caricatures incorporate a mini-representation of what appears (according to a set of heuristics) to be the most significant image. The resulting caricatures are quite large, so that typically just 6 might be visible on a screen. The system also suffers from poor response time, because images are not stored by the search engine but must be downloaded and processed during the search.

Woodruff et al. [23] report similar findings to those of Ogden et al. [20] - that sometimes textual document descriptions are quicker to use than plain thumbnails, and sometimes vice versa. However, Woodruff et al.'s [23] 'enhanced thumbnails', at approximately 10 to a screen, incorporate query terms clearly superimposed on each miniature and highlighted with colour and/or font size, and appear to offer considerable speed improvements. Reporting a time reduction of about 30% for enhanced thumbnail use compared with text, Woodruff et al. [23] conclude that enhanced thumbnails result in more consistent user performance than either text summaries or plain thumbnails. Although the enhanced thumbnails were well liked by users, the designers acknowledge that in a production system there would be substantial bandwidth and processing overheads involved in downloading images and generating thumbnails from raw HTML.

3.4 Design Pointers

In the document-level interfaces discussed above, only the designs by Veerasamy and Belkin (multiple-bar columns) [18] and Woodruff et al. (enhanced thumbnails) [23] demonstrated speed improvements over textual displays. Although it is difficult to detect any clear pattern to help identify an underlying cause, it is worth noting that Veerasamy and Belkin's [18] columns of vertical bars were visually the least complex of all the solutions. They were much less detailed than Hearst's [16] TileBars, and arguably easier than Byrd's [3] graphics to interpret quickly, with their clear positioning of query terms beside the bars. The prominent display of query terms in Woodruff et al.'s [23] visualization mark it out from the other thumbnail solutions. The enhanced thumbnails provide content cues apparent even during a rapid visual scan, whereas the other thumbnail solutions appear to provide either too little text or too much. Tentatively, it seems as though an information retrieval interface using uncluttered graphics clearly portraying the relationship between query terms and document content can help reduce search time.

4 System Design Considerations

Having examined user search practice and a range of visualization strategies, in this section we seek to elucidate our own design strategy for the system reported in this paper. Our intention is to augment with visual cues the document surrogates that are central to the success of current search engines.

Given that most queries are brief and underspecified, we will provide a set-at-a-time view in order to encourage users to consider increased numbers of results. To avoid impatient users abandoning a search, and to be feasible in practical contexts, the computational and communication overheads need to be lightweight. We have chosen simple and familiar 2D representations - pie charts - and results will be available in standard browsers. We have ruled out the use of thumbnail displays since user benefits are variable and system demands significant.

In the studies examined so far, it is assumed that displaying additional information will help users locate the relevant documents. Stanyer and Procter [8] state that, "it is metadata that is the key element in improving users' capacity to make informed assessments of a document's quality and relevance". In their prototype 'magic lens', the user can specify which metadata should appear. However, with the exception of the document abstract, they find no consensus about which metadata would be most useful. This confirms the earlier observation that users vary their decision-making strategies, depending on their needs. As Ogden et al. [20] note, "the number and types of decisions that need to be made vary from task to task… A decision making strategy that works well for one type of information request may not work for another."

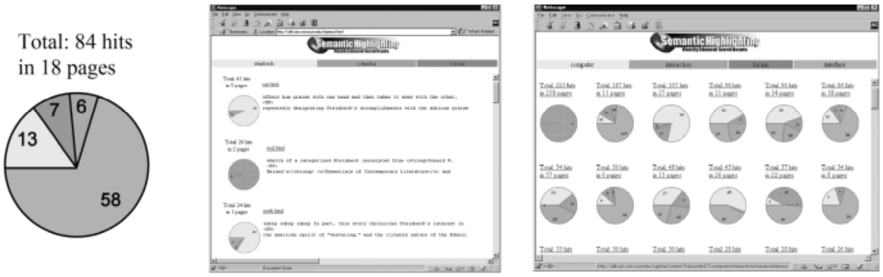
Since provision of a wide range of metadata would have major computational costs, it is worth reviewing the main types in an attempt to decide which would have greatest benefit for judging document relevance. Although thumbnails are not strictly speaking metadata, we include them here since they provide information about document appearance.

Query term distribution helps with locating relevant passages, not relevant documents. Therefore it is less appropriate to display it at the document assessment stage. *Thumbnails* have shown mixed effectiveness, are computationally quite demanding and generally allow no more documents to be represented on screen than do textual document surrogates. *Relative document length* is nowhere mentioned as of primary importance, though file byte size could help distinguish, for example, an authoritative report from a brief comment. *Number and size of embedded media types* (images, animations or audio clips) and presence of web page and e-mail links can certainly be helpful cues, but are not stored by most search engines and often seem to be of secondary importance.

Two pieces of metadata seem to support the user's natural wish in every search to understand the relationship between his query and the results. *Query term weights* derive from complex computations, so users cannot interpret them clearly. In contrast, interpretation of *query term frequency* seems natural to users. Clarke et al. [24] report that for short queries users expect a document matching many terms to be ranked before those with fewer matching terms, regardless of the frequency of term occurrence. We feel that because query term frequency relates obviously to both query and document and seems to be readily understood by users, it is the single most valuable piece of document metadata that can be automatically generated. It therefore is the basis for the visualization in our prototype.

5 Prototype

The prototype system, which we have named 'Semantic Highlighting', was designed to work alongside a freely available search engine [25] able to provide the term frequency details. It offers three views, 1) 'Pie and Text', 2) 'All Pies' and 3) 'Full-text'. In the document and set views each document is represented by a colour-coded pie chart, in which segment size represents the relative frequency of a query term, providing a quick visual impression of the query to document relationship (Fig. 1a). The colours resemble those commonly found in highlight pens. A count of term occurrences ('hits') is also shown for when users require more detail, along with estimated page numbers (based on file size) (Fig.1b). Computational overhead for a pie chart gif is small and the increase in bandwidth requirement at this point low, because images are only 6Kbytes. A legend at the top of the views associates each term with a colour. In the 'Pie and Text' view each entry is represented by a pie chart and conventional document surrogate. Users may scan the pie charts and informally apply coordination level ranking, and/or assess the document surrogate in the normal way.



(a) (b) (c)

Fig. 1. (a) Example Pie Chart Showing Hit and Page Totals (b) 'Pie and Text' View, (c) 'All Pies' View

The 'All Pies' view was developed to display pie charts only (Figure 1c). In this mode users can review 18 to 30 documents at a time. Moving the mouse over the pie chart displays the document's URL.

Both views are generated by CGI scripts written in C calling the search engine's API to collect data for the pie charts, which are then produced by a Perl script. The main processing overhead is in handling the search strings. Where a string has more than one keyword - the normal situation - each keyword must be searched for individually. This reduces performance substantially in our prototype, so for testing purposes the corpus has been kept deliberately small (under 200 documents) resulting in minimal delay. In both displays documents are ordered by the total number of query terms matched, in line with anticipated user expectation.

The 'Full-text' view was designed to display the text of a selected document, in which the term distribution is dynamically revealed by a combination of colour highlighting and forwards and backwards navigation arrows (implemented in Javascript) for quickly locating each term (Fig. 2). This allows rapid scanning or searching for particular targets, and co-occurrence of terms is readily spotted.

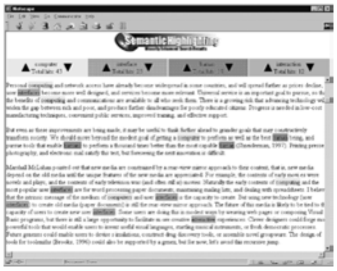


Fig. 2. 'Full-text' View

6 Evaluation

In recent years, a great deal has been written about the difficulty of evaluating interactive information retrieval (IR) [18] [26]. Human variability, including cognitive abilities, context and information needs, has been difficult to factor into conventional IR evaluation studies which have focused on search engine performance. The interactive track TExt Retrieval Conference (TREC) is currently examining the searching process as well as the outcome. However, to have used the large document collections they specify would have resulted in unacceptable response times on our prototype system. Instead we chose a user centred, task-oriented comparative evaluation as advocated by Jose, et al. [27], with a strong focus on user satisfaction as a performance criterion. The evaluation aimed to establish whether the visual enhancements speeded up user assessment of document relevance (and by extension, encouraged users to persist in their search tasks).

The document corpus was taken from different fields - research, medicine, sport and entertainment. It consisted of 163 HTML files, considerably fewer than we would have liked, but limited to guarantee good response times. The documents were carefully chosen to ensure that the subjects could not identify the target by title or citation alone. The corpus was also constructed so that when the pre-defined test queries were entered, usually at least 15 documents would be returned in addition to the target document, so that subjects had to exercise considerable judgment.

The two sets of test queries were designed to mimic real queries, and erred on the underspecified side. They did not include, for example, a document's title, author, citation or URL.

6.1 Subjects

Two groups of subjects were used, representing different levels of computing skill. The ’Students’ were 35 summer school participants studying history at Central Missouri State University whose ages ranged from 19 to 52. Only 4 described themselves as ’computer savvy’, but most were casual WWW users. The ’IT Staff’ were 66 professional information technology staff providing campus-wide computing services at the University of Missouri, Columbia. Ages ranged from 20 to 63, and almost all were heavy web users or developers.

6.2 Experiment Description

Each person was to use one of three search engine interfaces. A ’Conventional’ view was included as a control, displaying only each document’s URL and textual description in a familiar layout. The ’Pie and Text’ view and the ’All Pies’ view were as in Figures 1b and 2c respectively. Subjects were given a ten minute introduction and practice session to familiarize them with their allocated search interface. Each was then given two search tasks and monitored and timed while they used the corresponding queries provided for them. The first task was to answer the question, ’What is the average height to which chrysanthemum flowers grow?’. They were instructed to use the query ’chrysanthemum height’. The second question was, ’What scholar from Columbia University did John Steinbeck have an association with?" and the query was "Steinbeck columbia scholar’.

Timing of responses began once the search results appeared on screen. This effectively removed any distortion caused by varying query execution times. Timing stopped as soon as a subject found the target document. After the searches, subjects were asked to complete a survey, considering how the interface design supported the information retrieval task.

6.3 Analysis

Although there is a noticeable variation between the user groups in the average time to locate the required information (Table 1), both pie chart interfaces exhibit a major time reduction. Most of the users’ time was spent reading document text, so the large efficiency gains were due to earlier identification of the most promising documents. The addition of the pies to the text has decreased time to locate required information by a factor of 2.5 for the IT Staff and by nearly 10 for the Students. The different responses from the IT Staff and Student groups reflect the markedly different levels of competence in online searching. The time improvements for ’Pie and Text’ are even more marked than for ’All Pies’, indicating that the document surrogate is essential to retain. What is surprising is just how well the ’All Pies’ ’ interface worked, given how little information it displayed per document. The same rank ordering of the 3 display interface styles is evident from the average number of documents opened to complete a search task (Table 1, see figures in brackets).

Table 1. Average Time in Seconds to Locate Required Information

View	IT Staff	Students	All Subjects
Conventional	117.5 (1.9)*	272.0 (4.3)	176.1 (3.1)
All Pies	50.4 (1.1)	38.2 (2.1)	57.5 (1.6)
Pie and Text	45.9 (1.0)	28.3 (1.2)	37.6 (1.2)

* (Average number of documents opened to complete a task shown in brackets)

Further analysis of the times taken by ’Conventional’ searchers showed wide variations. The standard deviation for this group was 134 seconds, compared with 30 seconds for the other searchers. This indicates that the availability of pie charts produced shorter and more predictable times.

Survey responses showed that users liked the pie charts, a finding common to many systems that use visualization [3] [21] [23]. The ability to locate the highlighted search terms in the ’Full-Text’ view was also reported by all subjects as helping them find the terms faster, and 63% rated this as ’extremely fast’. When asked, ’Did you understand what the pie chart meant?’, 92% and 88% respectively of the ’All Pies’ view and ’Pie and Text’ view users respectively replied ’yes’, confirming a good cognitive fit between the charts and the user’s task. ’Pie and Text’ users unanimously replied yes to the question, "Did this visual search tool enable you to find/locate your document", compared with 94% for the ’All Pies’ interface. The addition of hit totals to the pie charts was confirmed as helpful by 73% of ’All Pies’ and ’Pie and Text’ users.

7 Conclusion and Continuing Work

There is a mismatch between user’s abilities and needs, and what search engines currently deliver. The unsophisticated brevity of most queries conceals much of the richness of information needs and human decision making. The ’information overload’ of long lists is met with a simple but poor strategy - assess first results only, disregard the rest.

In considering how to design an interface that could enhance popular efficiency in assessing search results, we drew on four observations in the literature. 1) Appropriate visualizations can enable users to scan results very efficiently. 2) People use many and varied ways to determine document ’relevance’, and any interactive IR interface must take this into account. 3) Of the many potentially helpful metadata that can be displayed, there is evidence that query term frequency is particularly beneficial, enabling users to appreciate the query - document link. 4) Simpler 2D visualizations tend to provide a better cognitive fit than complex 3D interfaces.

The prototype we developed had performance limitations (as did many of the systems reviewed) and was therefore restricted in the number of documents it could search at interactive speeds. Although the evaluation had to be constrained accordingly, it showed that query term frequency presented in colour-coded pie charts can speed users in assessing relevance for general information seeking.

Encouraged by our findings, we have redeveloped the prototype to allow a much larger scale and rigorous statistical evaluation of interactive searching. In particular, we want to gather data on how long users persist in information searches. Performance has been radically improved by moving to a major search engine, so that response time on a large intranet of 50,000+ documents is fast, and multiple file types can be indexed. Users can switch seamlessly from the document to the set view and can control the order of results (e.g. by date, by size) to best suit their selection criteria. The new interface also allows users to develop a document shortlist for later review and thus avoid interrupting their rapid scanning. Finally, we are also exploring the potential benefit of incorporating indications of query term synonyms to help users in query reformulation.

References

1. Spink, A., Wolfram, D., Jansen, B.J., Saracevic, T.l.: Searching the Web: the Public and their Queries. Journal of the American Society for Information Sciences, **53**(2) (2001) 226-234
2. Sellen, A.J., Murphy, R., Shaw, K.L.: How Knowledge Workers Use the Web. Proceedings of CHI 2002, **4**, Issue 1 (2002) 227-234
3. Byrd, D. A.: Scrollbar-based Visualization for Document Navigation. Proceedings of Digital Libraries 99 Conference, ACM, New York (1999) 122-129
4. Hearst, M.A.: Next Generation Web Search: Setting our Sites. Bulletin of the Technical Committee on Data Engineering , **23** (2000) 38-48
5. Mizzaro, S.: How Many Relevances in Information Retrieval? Interacting with Computers, **10** (3) (1998) 303-320.
6. Wilkinson, R., Zobel, J., Sacks-Davis, R.: Similarity Measures for Short Queries. Fourth Text REtrieval Conference (TREC-4), (1995) 277-285
7. Dunlop, M. D.: Reflections on Mira: Interactive Evaluation in Information Retrieval. Journal of American Society for Information Science, **51**(14) (2000) 1269-1274
8. Stanyer, D., Procter, R.: Improving Web Usability with the Link Lens. Journal of Computer Networks and ISDN Systems, Proceedings of the Eighth International WWW Conference, Toronto, Elsevier, **31** (1999) 455-466
9. Card, S.K. Mackinlay, J.D., Shneiderman, B.: Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, San Francisco (1999)
10. Spence, R.: Information Visualization. Addison-Wesley, Harlow (2001)
11. Mann, T.M. Reiterer, H.: Case Study: A Combined Visualization Approach for WWW-Search Results. Supplement to Proceedings IEEE Symposium on Information Visualization (InfoVis'99). (IEEE Computer Soc. Press) (1999) 59-62
12. Sebrechts, M.M. Vasilakis, J., Miller, M.S., Cugini, J.V., Laskowski, S.J.: Visualization of Search Results: A Comparative Evaluation of text, 2D, and 3D interfaces. Proceedings of SIGIR'99, Berkeley, California (1999) 3-10
13. Chen, C., Yu, Y.: Empirical Studies of Information Visualization: a Meta-analysis. International Journal of Human Computer Studies, Special Issue on Empirical Evaluation of Information Visualizations, **53**(5) (2000) 851-866
14. Westerman, S.J., Cribbin, T.: Mapping Semantic Information in Virtual Space: Dimensions, Variance and Individual Differences. Int. J. Human-Computer Studies, **53** (2000) 765-787

15. Cockburn, A., McKenzie, B.: An Evaluation of Cone Trees. People and Computers XV: Proceedings of the 2000 British Computer Society Conference on Human Computer Interaction (2000) 425-436
16. Hearst, M.A.: TileBars: Visualization of Term Distribution Information in Full Text Information Access. Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems. Denver, CO. (1995) 59-66
17. Hearst, M.A.: Improving Full-text Precision on Short Queries Using Simple Constraints. Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR), Las Vegas, NV. (1996) 217-232
18. Veerasamy, A., Belkin, N.J.: Evaluation of a tool for visualization of information retrieval results. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96), (New York, ACM) (1996) 85-92
19. Veerasamy, A., Heikes, R.: Effectiveness of a Graphical Display of Retrieval Results. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '97), (New York, ACM) (1997) 236-245
20. Ogden, W.C, Mark, M.W., Rice, S.: Document Thumbnail Visualization for Rapid Relevance Judgments: When do they Pay Off? In Harman, D.K., Voorhees E.M. (eds.): The SeventhText REtrieval Conference (TREC-7). NIST. (1998) 528-534
21. Czerwinski, M., van Dantzich, M., Robertson, G.G., Hoffman, H.: The Contribution of Thumbnail Image, Mouse-over Text and Spatial Location Memory to Web Page Retrieval. In 3D. Proceedings of INTERACT 99, IFIP TC.13 International Conference on Human-Computer Interaction, Edinburgh, Scotland. (1999) 163-170
22. Wynblatt, M.; Benson, D.: Web Page Caricatures: Multimedia Summaries for WWW Documents. Proc., IEEE Int'l. Conference on Multimedia Systems and Computing, (Austin, TX) (1998) 198-205
23. Woodruff, A. Faulring, A. Rosenholtz, R., Morrsion, J., Pirolli, P.: Using Thumbnails to Search the Web. CHI 2001. (2001) 198-205
24. Clarke, C.L.A., Cormack, G.V., Tudhope, E.A.: Relevance Ranking for One to Three Term Queries. Information Processing and Management, **36** (2000) 291-311
25. www.pls.com
26. Draper, S.W., Dunlop, M.D.: New IR - New Evaluation: The Impact of Interaction and Multimedia on Information Retrieval and its Evaluation. The New Review of Hypermedia and Multimedia, **3** (1997) 107-122
27. Jose, J.M., Furner, J.F., Harper, D J.: Spatial Querying for Image Retrieval: A User-Oriented Evaluation. In Croft, B., Moffat, A., Van Rijsbergen, C. J., Wilkinson, R., Zobel, J. (eds.), Proceedings of the Twenty First ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press. (1998) 232-240

Information Therapy in Digital Libraries

Yin-Leng Theng

Division of Information Studies
School of Communication and Information
Nanyang Technological University, Singapore
tyltheng@ntu.edu.sg

Abstract. This paper reports a pilot study to investigate how good current digital libraries (DLs) are in helping users understand their own needs – a kind of information service or therapy traditionally provided by librarians in conventional libraries. A sample group of DLs and subjects were selected for this study. Findings indicate that this sample group of DLs, though providing some form of information therapy, could be more explicit in guiding subjects to understand their own needs and thus help them to accomplish their goals more effectively. Using this study as a basis, the paper highlights insights on how this important service currently provided by conventional libraries could be more efficiently and creatively provided in DLs.

1 Introduction

In the book "Information Ecologies: Using Technologies with Heart", Nardi and O'Day (1999) stress that "helping clients understand their own needs" is one of the most valuable (and unheralded) services provided by conventional libraries [18]. They called this kind of service *information therapy*, in which a reference librarian acts like a good psychotherapist, and through skillful questioning gets clients to talk about their needs. The librarians, also referred to as information therapists, are trained to carry out "reference interviews" to help clients articulate their needs in a precise way using an "impressive deployment of tact, diplomacy, and persistence, as well as skillful interviewing technique" [18].

Unlike conventional libraries, a digital library (DL) is not a building, and it remains an abstract and amorphous thing, even to library professionals [9, 12]. Although no one is really sure who coined the term "digital library", it seems that an early reference came from Al Gore, then Vice-President of the United States, in 1994 [10].

The term "digital library" is a misnomer, conjuring mental expectations of a conventional library [9]. Many expect the roles of DLs to reflect the roles of conventional libraries [12, 22], of which "information professionals making judgments and interpreting users' needs" is one such expected service. Matson and Bonski (1997) advocate that libraries exist, whether digital or otherwise, to help users to find information [15].

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 452-464, 2002.

In a conventional library setting, what users want becomes clearer and more focused after some discussion with the librarians (information therapists). In a DL context, however, no "such librarian" is present.

2 The Study

In contrast to other studies on understanding users' browsing behaviours (for example, [1, 3, 4, 6]), this paper presents a pilot study to investigate how good current DLs are in providing the kind of information therapy proposed by Nardi and O'Day [18], that is, in assisting users to understand their needs in order to achieve their goals. Hence, the objectives of this study are to:

- investigate how good current DLs are in helping users to understand or refine their needs while they are browsing – a kind of information therapy provided by conventional libraries; and
- propose how DLs could proactively help users understand their own needs, and suggest ways in which this kind of information therapy could be incorporated into DLs.

Methodology – Rationale for the Qualitative Approach Taken

Generally, most people feel that in order to get statistically significant results, a number of people (around 20-25) should be asked to carry out the experimental task, in order to pick up a wider range of problems and to get some sense of the frequency and the consequences of each. However, Nielsen and Landauer (1993) conclude from analysing usability problems described in eleven published projects that the maximum cost/benefit ratio for a medium-large software project could be derived from using *three* test users [19].

Since the aim is to learn which detailed aspects of the interface are good and which are bad in providing users with a form of information therapy, and how design could be improved, small numbers of users are more cost-effective as common/frequent problems but not infrequent or minor ones are encountered first. (The problem is — and remains for any methodology — how to find the infrequent disasters!). With 3-6 people, this paper hopes to get qualitative results and impressions.

Video analysis, think-aloud protocol, questionnaires and interviews were used in this study. The sessions were video-taped, and the tapes were analysed to identify potential areas of difficulty experienced by the subjects, and hence identify the lack of provision of features to support information therapy in DLs.

Related Studies on Users' Browsing Behaviour

We define "information therapy" to be a special kind of service provided by libraries, digital or otherwise, to aid users to understand their own needs and to make informed decisions in order to carry out reasoned and reasonable actions.

To help us investigate if current DLs are providing this form of "information therapy", we need to understand the interactions of users when *browsing* and detect when/what problems are encountered. We made the assumption that non-completion of tasks could be due to a lack of provision of features in the DLs to support this form of information therapy.

Browsing is a very common exploration strategy employed by users when navigating systems. However, browsing means different things to different people [5, 16]. Although text browsing is often specific and constrained by the context (books, journals, *etc.*), it is a useful benchmark for a more precise definition of browsing [16]. Some worthwhile sources include: Morse (1973) describing browsing as seeking for new information [17]; Hildreth (1982) suggesting the serendipitous nature of some types of browsing [11]; and Batley (1989) claiming that browsing is more purposeful and focused [2]. These definitions from information scientists seem to suggest that there are different types of text browsing involved, while O'Connor (1985) categorises browsing as systematic, purposeful and serendipitous [20].

Marchionini and Shneiderman (1988) define browsing as "an exploratory, information seeking strategy that depends on serendipity, especially appropriate for ill-defined problems and for exploring new task domains" [14]. Cove and Walsh (1988) use a three-stage model to describe browsing [7]: (i) search browsing where the goal is known; (ii) general purpose browsing where the goal is to consult sources that have a likelihood of items of interest; and (iii) serendipity browsing which is purely random. Bates (1989) sees browsing as a semi-structured, goal-oriented activity, distinct from Boolean search [1].

Browsing, as defined in this paper, is both unfocused and focused [23]. *Unfocused browsing* refers to reading and navigating in a system without a definite or explicit goal for the users to accomplish. *Focused browsing*, on the other hand, is one where users have an idea (which may not be too clear) of what they want to do to accomplish a task.

Selected Digital Libraries

Four DLs were chosen for this study : [1]Networked Computer Science Technical Report Library (NCSTRL; see http://www.ncstrl.org/) and University of Calgary Digital Library (UCDL; see http://pharos.cpsc.ucalgary.ca/); ACM Digital Library (ACMDL; see http://www.acm.org/) and IDEAL On-Line (IDEAL; see http://www.idealibrary.com/). These four were chosen because they are one of the better examples of DLs found on the Web, in terms of their information and coverage.

Subjects

The six subjects (denoted as U1 – U6) selected for the study were: two 1st-year PhD students; one 3rd-year PhD student; one Web developer; one DL designer; and one admin staff. All the subjects were from the Computing Science department.

The two groups of three subjects (Groups A-B) compared two similar kinds of DLs. Group A subjects examined NCSTRL and UCDL; while Group B subjects examined ACMDL and IDEAL. Subjects U1 – U3 (Group A) had between 5 – 9 years of experience using the PC/Mac. They used the Web very often/often. Subject

[1] This study was conducted in early 2001. Since October 2001, NCSTRL has adopted a new approach to incorporate OAI-based technical reference materials. The participating institutions promise in their Web sites that there is little impact initially on the operations at most NCSTRL sites, and the service at the NCSTRL web site should provide similar capabilities to what resulted from previous NCSTRL support.

U3 had not used any DL before taking part in this study. Subject U1 was not certain what a DL is. To him, DLs and Web sites were not that different. Only Subject U2 had experience using DLs, in particular NZDL and IDEAL. Subjects U4 – U6 (Group B) had 5-7 years of PC/Mac experience. They used the Web very often/often. Only Subject U5 had not used a DL before. Subjects U4 and U6 made use of ACMDL, NZDL and NCSTRL in their research.

Focused Browsing Task

To help us understand if current DLs are providing some form of "information therapy", we are interested in subjects' performance when using DLs. Therefore, we provided the subjects with a focused browsing task. The subjects could decide how long to spend on the task. They were asked to think aloud while working.

An example of a focused browsing task is: "Browse all articles by Saul Greenberg from 1996 to 1999."

Procedure

The subjects were briefed on each library before they began the browse task. Group A subjects performed browse task using NCSTRL and UCDL; and Group B subjects on ACMDL and IDEAL.

When the subjects had completed the browse task using the first DL, they were asked to complete an extensive questionnaire commenting on the structure of the DLs in helping them to complete the tasks successfully. *Satisfaction* refers to the "feeling of being pleased with the DLs in helping to complete the task successfully". *Being pleased* is defined in terms of the subjects' perceived ease of use, rate of errors, and time taken to perform the tasks successfully. More details of the questionnaire can be found in [24].

The subjects then proceeded to work on the second DL. At the end of working with the two DLs, they were interviewed on what they thought of the DLs in terms of their usability and usefulness.

Although there are differing views on usefulness and usability (e.g. [8], etc.), *usability* of DLs is referred to in this paper using established dimensions such as: screen design; terminology and system information; system capabilities and user control; navigation; and completing tasks. *Usefulness* is measured in reference to system specifications and not on user performance testing [13].

3 Findings and Analyses

In Section 3.1, we use an example to illustrate how user interactions were analysed using well-known models of interactions. Interaction involves at least two participants: the user and the system. Both are complex and are very different from each other in the way they communicate and view the domain and the task. For interaction to be successful, the interface must therefore effectively translate between them. This translation can fail at a number of points and for a number of reasons. Models of interaction can help us to understand exactly what is going on in the interaction and identify the likely root of difficulties [8]. The two models used in our analyses are:

- *Norman's model of interaction* is perhaps the most influential because of its closeness to our intuitive understanding of the interaction between the human user and computer. The user formulates a plan of action and this is then executed at the computer interface. When the plan, or part of the plan, has been executed, the user observes the computer interface to evaluate the result of the executed plan, and to determine further actions.
- *Interaction Framework* developed by Abowd and Beale addresses the limitation in Norman's model of interaction to include the system explicitly [8]. According to the Interaction Framework, there are four major components in an interactive system: the System; the User; the Input and the Output. The interaction framework is a means to judge the overall usability of an entire interactive system. All of the analysis suggested by the framework is dependent on the current task (or set of tasks) in which the User is engaged. This is not surprising since it is only in attempting to perform a particular task within some domain that we are able to determine if the tools we use are adequate [8].

Section 3.2 reports on the summaries and analyses of subjects' interactions with the respective DLs.

3.1 An Illustration

Space constraints keep me from writing about all the good, usable features these DLs have. It is not the intention of this paper to downplay the effort put in by the designers of these DLs.

Using Subject U3, we illustrate how a subject's interactions with a DL were analysed using the Interaction Framework. The intention of the analysis is to elicit problems, give hints on the form and extent of information therapy provided in the DLs in helping subjects to complete their tasks.

Fig. 1 shows coding of the translations between components proposed in the Interaction Framework [8]. User's browsing interactions can be analysed according to four categories based on the Interaction Framework: (i) user action (UA); (ii) user evaluation (UE); (iii) system display (SD); and (iv) system response (SR).

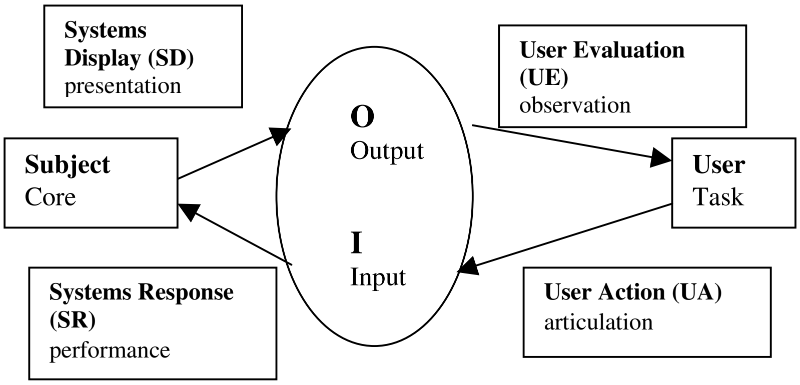


Fig. 1. Coding based on Interaction Framework (adapted from Dix et. al., 1997)

Table 1 tabulates analysis of Subject U3's browsing interactions with UCDL. Column 1 shows a record of user's browsing interactions, and classifies user's browsing interactions according to the four components of the Interaction Framework: (i) UA; UE; SD and (iv) SR. Column 2 draws conclusions on observations made from SR, SD and UE.

Conclusions drawn based on SD give us clues on which design features/aspects of the interface help subjects to achieve their goals (see Table 1). From UE, we get indications of problems that might have prevented subjects from achieving their goals. These problems could either be due to: (1) systems; (2) user; or (3) design faults. *System* problems are machine or server-related problems. For instance, long download time and server not responding to requests are some examples of systems problems. *User* problems refer to errors made by users [21]. This could be due to subjects keying in wrong data or subjects misreading information on the screen. If the problem is a *user* problem, then it is *psychological* and may be due to users' inability to exploit computer screens, and complex information structures. Thus, as a psychological problem, it can be alleviated but not solved by better design.

Table 1. Analysis of Subject U3's Interactions with UCDL

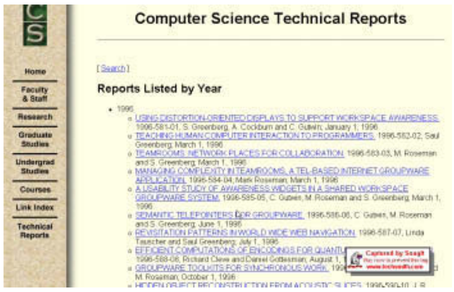
Transcript of Coded Interactions	**Remarks**		
	From SD (before UA)	**From SR (after UA)**	**From UE (after UA)**
Subject U3 started by clicking (UA) onto "Technical reports" (SD), followed by "browse publication by year" (SD). He clicked (UA) onto "1996-2001" to browse collection by year. Too many results (SR) returned. (UE)	Browse features available.	Results returned.	The results returned also included other authors, but not just "Saul Greenberg". He could not find any feature to narrow the search list. Too many results -> design problem
He then went back (UA) to the "Technical reports" page (SD), and clicked (UA) on "browse collection by author" (SD). Owing to server not responding to the request (SR), he was unsuccessful (UE).	Features available to refine browsing terms.	Server problems	Server not responding -> system problem?
He then tried browsing (UA) using "fielded search form" (SD), followed by "browse publication by year" (SD). Results returned (SR). He had no success in locating the articles (UE).	Search and browse features available.	Results returned.	Too many results -> design problem
He then tried (UA) "browse collection by authors" (SD). It worked this time (SR). Results were relevant (UE).	Browse features available.	Results returned.	Task completed.

UCDL uses the server that runs NCSTRL with its own customised, front-end interface. All three subjects did not encounter any difficulty and the search task was successful with them clicking onto different options: "browse by author" (Subject U1); "search by fielded search form" (Subject U2); and "search by keyword" (Subject U3).

Subject U1, a Web developer, was more impatient with the slow response, and made remarks such as "… taking a long time" several times. He completed the task in 3 minutes. The other two subjects did not complain. They took 2 minutes to complete the task. It appeared that the more experience the subjects have, the higher their expectations. A 10-second download time seemed to be the threshold of tolerance for experienced subjects, with 20 seconds for other subjects.

Subject U3 could begin to carry out the browse task because there were features provided in UCDL such as “technical reports”, “browse publication by year” and “1996-2001” (see Table 1; Column “From SD”). Although the first two attempts were not successful, Subject U3 could still try other facilities provided such as “browse collection by author” and “browse publication by year”. At the third attempt, he was successful using the “browse collection by authors” feature. Comments on UCDL fall into these 2 areas:

- *Guiding Browsing Needs.* UCDL provides facilities for Subjects U1, U2 and U3 to re-submit the browse task until the task was complete. However, there was little on the interface to prompt the subjects what to do except for subjects to guess they needed to re-select the different options. Perhaps it would be better for UCDL to give some kind of feedback/instruction to subjects to re-submit browse tasks when they were not satisfied with the results.
- *Hindering Browsing Needs.* Table 1’s “From UE” column provides information on the types of problems hindering subjects in achieving their browsing tasks more effectively. For example, Subject U3 could not complete the browsing task in the first two attempts because too many results were returned, a design problem (see Fig. 2).



Computer Science Technical Reports

[Search]

Reports Listed by Year

- 1996
  - USING DISTORTION-ORIENTED DISPLAYS TO SUPPORT WORKSPACE AWARENESS, 1996-581-01, S. Greenberg, A. Cockburn and C. Gutwin; January 1, 1996
  - TEACHING HUMAN COMPUTER INTERACTION TO PROGRAMMERS, 1996-582-02, Saul Greenberg; March 1, 1996
  - TEAMROOMS: NETWORK PLACES FOR COLLABORATION, 1996-583-03, M. Roseman and S. Greenberg; March 1, 1996
  - MANAGING COMPLEXITY IN TEAMROOMS, A TEL-BASED INTERNET GROUPWARE APPLICATION, 1996-584-04; Mark Roseman; March 1, 1996
  - A USABILITY STUDY OF AWARENESS WIDGETS IN A SHARED WORKSPACE GROUPWARE SYSTEM, 1996-585-05, C. Gutwin, M. Roseman and S. Greenberg; March 1, 1996
  - SEMANTIC TELEPOINTERS FOR GROUPWARE, 1996-586-06, C. Gutwin, M. Roseman and S. Greenberg; June 1, 1996
  - REVISITATION PATTERNS IN WORLD WIDE WEB NAVIGATION, 1996-587-07, Linda Tauscher and Saul Greenberg; July 1, 1996
  - EFFICIENT COMPUTATIONS OF ENCODINGS FOR QUANTU[illegible] 1996-588-08, Richard Cleve and Daniel Gottesman; August 1, [illegible]
  - GROUPWARE TOOLKITS FOR SYNCHRONOUS WORK, 199[illegible] M. Roseman; October 1, 1996
  - HIDDEN OBJECT RECONSTRUCTION FROM ACOUSTIC SLICES, 1996-590-10, J. R.

Too many results returned. No way to refine the results

Fig. 2. Screen shot of UCDL’s browse results by year.

3.2 Summaries and Comments of Subjects' Interactions

Below are summaries of subjects' interactions with NCSTRL, ACMDL and IDEAL interspersed with user evaluation (denoted as UE) to reflect one of the components suggested in the Interaction Framework. A compilation of UEs would give us hints on a list of usability problems encountered.

NCSTRL (Networked Computer Science Technical Report Library)

Subject U1 typed in author's surname "Shneiderman" provided in the "search ALL bibliographic fields" and selected "sort by author" to perform the *search* task. The task was not successful because of server problems. He reloaded the page and tried again. This time many results were returned but could not find the article he was looking for (UE #1). He then clicked onto "Ben Shneiderman" hoping to be brought to the page with a listing of all these articles. But an error message was returned which did not make much sense to Subject U1 (UE #2) except to indicate that the server was not working properly. Subject U1 next clicked on "search collection" in the navigation bar and typed the title to enter the search again. He was successful this time.

Subject U2 encountered more difficulties in trying to accomplish the task. He spent some time trying to understand the interface and did not know where to begin (UE #3). He then selected "search collection". However, he was confused with the two options provided: "search ALL bibliographic fields" and "search SPECIFIC bibliographic fields". (UE #4) He entered in both, but finally deleted the entry for "search ALL bibliographic fields" before executing the search. System registered an error. He was also unclear whether to enter the surname or the first name of the author (UE #5) He then tried again, typing the title of the journal instead of the title of the article. He then asked for help and he completed the task.

Subject U3 completed the search task with less difficulty, first selecting the "search collection", sorting by author and typing author's full name. She encountered network problems during the duration of the task.

Comments

We report the problems gathered from comments/observations made by users (coded as UE) followed with suggestion(s) on how these could be resolved.

- *UE #1 : Too many hits returned.* This is also a problem identified by [2]. Users could be overwhelmed with just too many hits. Perhaps a facility could be provided for users to select the number of hits to be returned.
- *UE #2 : Error message that did not make sense*. It is important that these messages speak the users' language, not containing jargons. They should be simple and concise.
- *UE #3 : Complicated user interface with unclear options provided.* An established design principle is always to keep the interface uncluttered, evoking goals that are central to the purpose of the system [14]. In other words, if the system is designed to be a learning site, then the interface should encourage goals that pertain to learning, and not e-shopping, for example.

- *UE #4 : Cultural differences interfering with usability of system.* An example is the confusion experienced by Subject U2 of not knowing whether to enter the surname or the first name. Though many things may be taken for granted in the western world, this is one case where there are cultural differences. For designs to be internationally accepted, they should be culturally sensitive. This was also highlighted in other studies (e.g. [18, 24], etc.).
- *UE #5 : Not speaking user's language.* Examples include "search ALL bibliographic fields" or "search SPECIFIC bibliographic fields". Users may not be familiar with bibliographic searching. This could be made clearer with a "pop-up" message.

ACMDL (ACM Digital Library)

The three subjects completed the *search* task without much difficulty. All three subjects managed to click onto the "digital library" link to start their search from the ACM Home page after looking for it. The link to the "digital library" was not obvious. Subject U4 clicked onto "publications", but failed to find "Communications of the ACM" (UE# 6). Then he remembered clicking "Search ACM" option allowed him to search the library. In the search form, he then limited his search on publication to "Communication of the ACM". Ignoring the "In-Fields" options and "Authors" field, he entered the author's name and title of the article in the "Terms" field. Subject U4 is an experienced Web user. He hoped to carry out a phrase search. When he was unsuccessful, he removed the title of the article leaving the author's name. It returned no hits. He then removed the author's name, leaving the title. The search was then successful.

Subject U5 clicked onto "ACM Journals and Magazines", followed by "Search by Subject". The list was too long and the acronyms did not make sense to him (UE# 7). He then clicked on "search library". The task was completed by Subject U5 keying in the title and limiting search to year.

Subject U6 clicked onto "Browse journals and magazines", followed by "Communications of the ACM". She clicked on "search library" and entered title and author in the relevant fields to search all journals in the year 1987. The search was successful.

Comments

Similarly, problems detected from comments/observations made by users (coded as UE) are followed with suggestion(s) on how these could be resolved.

- *UE #6: Links not obvious to users.* To address this problem, standard Web conventions should be followed. For example, links would be recognized if they are blue and underlined. Or, the cursor would turn into a 'hand' when moved over links.
- *UE #7 : List of acronyms given did not make sense to users.* Acronyms should be accompanied with full text. With DLs reaching a wider world, it would be presumptuous to expect users to be familiar with acronyms that could be based on localized knowledge.

IDEAL (IDEAL On-Line)
The three subjects completed the *search* task without much difficulty. Subject U4 had not used IDEAL before. He clicked “Search” option on the navigation bar. In the search form, he selected the category “Maths and Computer Science” and journal “International Journal of Human-Computer Studies”. He did not notice the print below the journal field that indicates “choosing a journal overrides the Category selection and applies only to searches of IDEAL”. He then entered the author’s last name and title of the article. The search was unsuccessful. The “AND” combination was “too restricting” as acknowledged by Subject U4 (UE# 8). He then removed entry of the title leaving only the author’s last name. The task was successfully completed.

Subject U5 selected “Quick Search” on the Home Page to carry out the task. He entered the title into the “Quick Search” box. There were too many articles returned (UE# 9). He was not keen to look for the articles and he exclaimed “Wow, there are too many documents returned”. He attempted a couple more times using the “Quick Search” option. He did not understand why he could not find the article. He then noticed the “Link in” feature and clicked onto it to be brought to a search form screen. It took him a couple of tries to complete the task because he entered the incorrect data.

Subject U6 began the task by clicking on the “search” option. Similar to Subject U4, she also selected the category “Maths and Computer Science” and journal “International Journal of Human-Computer Studies”, but did not notice the message warning “choosing a journal overrides the Category selection and applies only to searches of IDEAL” (UE# 10). She then selected the title option to enter the title but the title was too long to be entered (UE# 11). Next, she tried entering the author’s last name and the year of publication. Nothing was returned. She thought she should not restrict the search to “AND” but instead changed the conditions to “OR”. Again, she was not successful. Finally, she only entered the author’s last name and the search returned three hits, one of which was what she was looking for.

Comments

- *UE #8: Search options too restricting.* Designs should be catered to different levels of experience of users with the DL. For example, simple search options could be provided for novice users, but advanced search options for experienced users.
- *UE #9: Too many documents returned.* As observed in NCSTRL, perhaps a facility could be provided for users to select the number of hits to be returned.
- *UE #10 : Not speaking the users’ language*. It is not useful if warning messages are not written in simpler English for users to understand.
- *UE #11 : Field length is not long enough for users to enter title*. This could be addressed either with the field length extended or a message to inform users of the maximum length to be entered.

4 Discussion

Our pilot study suggests that the sample DLs evaluated, though providing a minimal form of information therapy, the affordance provided for information therapy, could be more explicit in guiding subjects to understand their needs, and thus help them to accomplish their goals more effectively.

In conventional libraries, the provision of professional advice is the kind of support given by a librarian. While helping users to find information by doing things for them, the librarian is also often surreptitiously teaching users how to make the best use of the library. As a result, users are able to do at least part of the task on their own. Simultaneously, the librarian learns about the interest of users. Often the support from librarians is augmented by the provision of support from user to user. More experienced users can offer informal help and advice to novice users.

In DL interfaces, a learning environment is necessary just like in conventional libraries. In order to create a learning environment, we need to provide additional facilities that help users, content providers and designers to fulfil their tasks, or even to provide intelligent intermediaries to do the tasks for them. In creating such a learning environment for end-users, we should provide suitable support features when collaboration between users is most effective. The construction of Community Memory Support Systems like Answer Garden and FAQ lists will allow users to gain an understanding of how systems can be used [24].

Awareness mechanisms have to be developed that will allow users to know when others are accessing the same resource. The use of synchronous co-operative support tools like Chat Rooms and Meeting Rooms will allow users to discuss and debate different approaches to accessing on-line resources. The core purpose of these tools is to support the co-operation and debate needed to resolve decisions.

The library's organisational principles took centuries to develop and we take for granted the very many organisational structures in libraries, such as cataloguing systems (Dewey Decimal System, etc.). Yet, organising a conventional library is difficult. Creating good DLs is even more difficult. The central organising principles in conventional libraries cannot be used in DLs, or we will end up with an electronic, conventional library.

Levy and Marshall (1995) stated that the current focus on fixed, permanent materials in DLs could be traced to a preoccupation with books as the central organising principle behind earlier libraries. They argue for DLs to be broadly-construed so that "the design of DLs must take into account a broader range of materials, technologies and practices", and they emphasize the importance of access and use of the collection by a community [12].

DLs should, therefore, be dynamic and include these three crucial aspects highlighted in [25] : (i) they house and provide access to documents including paper materials, electronic files, videotapes and audiotapes; (ii) they require technology to create and maintain documents; and (iii) they include actual work done by library users, as well as work done by library personnel in support of them.

5 Conclusion and On-Going Work

This paper reported a pilot study to investigate whether current DLs help users to understand their own needs – a form of information therapy provided by librarians in conventional libraries. Certainly, more can be done: careful analysis of data; repeat work with subject groups other than those from the Computer Science department, and control for factors such as Web skills. The paper concludes by discussing how

DLs can be more effective in providing this form of information therapy. This is initial work. Future work includes investigating relevant forms of information therapy that would be useful in making DLs more “library-like”.

Acknowledgements. I would like to thank Harold Thimbleby, Norliza Nasir and Paola Amaldi-Trillo for their feedback on earlier versions of the paper. Special thanks also go to the participants who took part in the study.

References

1. Bates, M.: The design of browsing and berrypicking techniques for the online search interface. Online Review, 13(October 1989), pp. 407 – 424. (1989).
2. Batley, S.: Visual information retrieval : Browsing strategies in pictorial databases. PhD. Thesis. (1989).
3. Blandford, A., Stelmaszewska, H. and Bryan-Kinns, N.: Use of multiple digital libraries: a case study, JCDL’01, pp. 179 – 188. (2001).
4. Borgman, C.: The user’s mental models of an information retrieval system: An experiment on a prototype online catalog. International Journal of Man-Machine Studies, Vol. 24, pp. 47-64. (1986).
5. Carmel, E., Crawford, S. and Chen, H.: Browsing in Hypertext: A Cognitive Study. IEEE Transactions on Systems, Man and Cybernetics, 22(5), pp. 865 – 884. (1992).
6. Cool, C., Park, S., Belkin, N., Koenemann, J. and Ng, K.B.: Information seeking behaviour in new searching environments. Reuters Internal Report. Available at http://scils.rutgers.edu/tipster3/colis.html (May 1, 2002).
7. Cove, J. and Walsh, B.: On-line text retrieval via browsing. Information Processing and Management, 24(1), pp. 31-37. (1988).
8. Dix, A., Finlay, J., Abowd, G. and Beale, R.: Human-Computer Interaction (2nd edition). Prentice-Hall. (1997).
9. Fox, E., Akscyn, R., Furuta, R. and Leggett, J.: Digital Libraries. Communications of the ACM, April 1995, Vol. 38, No. 4, pp. 23 – 28. (1995).
10. Gore, A.: Speech to International Telecommunications Union. March 21. (1994), http://www.goelzer.net/telecom/al-gore.html.
11. Hildreth, C.: The concept and mechanics of browsing in an online library catalogue. Proceedings of 3rd National Online meeting. (1982).
12. Levy, D. and Marshall, C.: Going digital: A look at Assumptions underlying digital libraries. Communications of the ACM. Vol. 38, No. 4, pp. 77-84. (1995).
13. Lingaard, G.: Usability testing and system evaluation: A guide for designing useful computer systems. Chapman & Hall. (1995).
14. Marchionini, G. and Shneiderman, B.: Finding facts versus browsing knowledge in hypertext systems. IEEE Computing, pp. 70-79. (1988).
15. Matson, L. D. and Bonski, D.: Do digital libraries need librarian? An Experiential Dialog. ONLINE, November 1997.
16. McAleese, R.: Navigation and browsing in hypertext. Hypertext: Theory into Practice, pp. 1-38. Intellect Books. (1993).
17. Morse, P.: Browsing and search theory. Toward a theory of Librarianship : Papers in honour of Jesse Hauk Shera. (1973).
18. Nardi, B. and O’Day, V.: Information Ecologies: Using Technology with Heart. The Mit Press. (1999).
19. Nielsen, J. and Landuer, T.: A mathematical model of the finding of usability problems. INTERCHI’93, pp. 206-213. ACM Press. (1993).

20. O'Connor, B.: Access to moving image documents: Background concepts and proposals for surrogates for film and video works. Information Retrieval Research, 41(4), pp. 209-220. (1985).
21. Reason, J.: Human error. Cambridge University Press. (1990).
22. Rowland, F.: The Librarian's Role in the Electronic Information Environment. ICSU Workshop'98. http://www.bodley.ox.ac.uk/icsu/rowlandppr.html. (1998).
23. Theng, Y.L.: Addressing the "lost in hyperspace" problem in hypertext. PhD Thesis Middlesex University. (1997).
24. Theng, Y.L., Duncker, E., Mohd-Nasir, N., Buchanan, G. & Thimbleby, H.: Design guidelines and user-centred digital libraries, Third European Conference ECDL'99, Abiteboul, S. and Vercoustre, A.(Eds.), pp. 167 - 183, Springer. (1999).
25. Theng, Y.L., Mohd-Nasir, M. and Thimbleby, H.: Purpose and Usability of Digital Libraries, Proceedings of the 5th ACM Conferences on Digital Libraries, pp. 238 – 239. (2000).

Evaluation of Task Based Digital Work Environment

Narayanan Meyyappan and Schubert Foo

Division of Information Studies, School of Communication and Information,
Nanyang Technological University, Nanyang Link, Singapore 637718
pn839200@ntu.edu.sg, assfoo@ntu.edu.sg

Abstract. A task-based Digital Work Environment (DWE) was designed and developed at the Nanyang Technological University in Singapore to support the Division of Information Studies (DIS) Masters students' information requirements for their dissertation work. This paper traces the evaluation studies of DWE to gauge the usefulness and effectiveness of its three different approaches to information organisation - alphabetical, subject category and task-based. The findings show that the task-based approach is most effective in terms of the speed of accessing information resources.

1 DWE Design and Development

A task-based DWE was designed and developed [1] to provide a one-stop information access point for both internal and external information resources that are accessed by the academic community. A job analysis was first carried out to identify all the main tasks and sub-tasks associated with dissertations (as a representative task) to create a task hierarchy and information resources required for the dissertation task through a focus group study [http://InformationR.net/ir/7-2/paper125.html]. The DWE uses a task-based approach to organise access to information resources according to different user categories (e.g. faculty, student). This was augmented with three additional approaches, namely, alphabetical approach (AR), subject category approach (SC), and a hybrid approach (HY) combining all these three different approaches.

2 DWE Evaluation and Analysis

Evaluation of the DWE. Evaluation was conducted on the DWE prototype through a questionnaire in a controlled environment on a one-to-one basis to assess the various approaches used for organising the information resources. The criteria used are time taken to access the desired information resource, the usefulness, ease of use of these approaches, and qualitative comments from the participants. Ten representative tasks according to task characteristics were designed for evaluation.

A total of 60 graduate students from DIS participated in the evaluation. They were divided into 3 groups of 20 students each. Each group was asked to evaluate two different approaches (each with 10 tasks), with the task-based approach being commonly evaluated across all 3 groups (i.e. AR-TB, SC-TB, and HY-TB). A total of 1200 tasks (200 each for AR/SC/HY and 600 for TB) were carried out in 3 months.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 465-466, 2002.

Participant Profiles – Demographic data. There were 30 male and 30 female participants. The majority (53.4% or 32) fell within the age range of 20-29 years old, followed by 40% (24) between 30-39, 5% (3) between 40-49 and 1.6% (1) who is 50 or over. 86% (or 52) had 4 or more years of computer experience. Forty-nine (81.7%) participants had experience using online databases, 78.3% (47) in e-journals, 71.7% (43) in digital libraries, 68.3% (41) in CD-ROM databases and 36.7% (22) in e-books.

The Univariate Tests carried out for the different approaches (F2, 54 =17.51 and $p < 0.0001$) and within the non-task-based approach (F 1,54=105.46 and $p < 0.001$) in terms of location time showed statistical significance. Further, stepwise multiple regression analysis was carried out on individual tasks to find out which independent variable (participant's age, computer and digital resources experience) is the best predictor. It was found that the information resources location time was greatly associated with age, gender and computer experience.

Effectiveness of Organisation of Information Resources. Based on averaged figures using the Likert scale of 1 to 5, participants gave higher ratings for HY for 6 tasks (with mean ratings of 4.35, 4.1, 4.05, 4.3, 4.3 and 4.10 for Tasks #3, #4, #6, #7, #8 and #9 respectively). The participants therefore expressed effectiveness of more than 80% for all tasks in terms of information resources organsiation..

Ease of Identifying Information Resources. For Task #1, TB was given the highest rating (mean=4.14) followed by HY (mean=3.95) on the Likert scale. For Task #2, the order was AR (4.35) followed by TB (4.33). Task #3 had an equal mean rating of 4.33 for both AR and HY. HY was preferred for Tasks #4, #5, #6, #7, #8, #9 and #10 with means of 4.05, 4.00, 4.10, 4.40, 4.30, 4.10 and 4.35 respectively. SC was less preferred for Tasks #5, #7 and #9 with means of 3.15, 4.00 and 3.7. It can be seen that that participants find it easier to locate information resources using the HY approach.

3 Conclusion

The evaluation findings showed that the task-based approach was preferred over the other approaches. Participants preferred either the task-based or hybrid model in terms of usefulness of the organization of information resources. However, the participants rated the hybrid approach as the most effective in terms of ease of identifying information resources.

References

1. Meyyappan, N., Foo,S. Design and Implementation of Digital Work Environment. (Unpublished Technical Report TR-IS (SCE) #07/2000, NTU, DIS, September 2000).
2. Meyyappan, N., Chowdhury, G.G., & Foo, S. Use of a Digital Work Environment Prototype to Create a User-Centered University Digital Library. *J. Inform. Sci. 27* (2000), 249-64.

A Framework for Flexible Information Presentation in Digital Collections

Unmil P. Karadkar[1], Jin-Cheon Na[2], and Richard Furuta[1]

[1]Center for the Study of Digital Libraries and Department of Computer Science
Texas A&M University, College Station, TX 77843-3112, USA
{unmil,furuta}@csdl.tamu.edu
[2]Division of Information Studies, School of Communication & Information,
Nanyang Technological University, 31 Nanyang Link, Singapore 637718
tjcna@ntu.edu.sg

Abstract. The caT (context-aware Trellis) architecture separates content, structure and display properties of information entities to support flexibility in information presentation. Users' browsing state is synchronized cross multiple client devices. We present a framework that is based on this architecture for creating temporary integrated environments of user devices and public access computing resources for providing information-centric services to readers who may access the system using a variety of devices.

Readers of digital collections use a wide range of devices to access a rich and varied information space. The devices range from handheld devices that are characterized by low computational power, tiny displays and slow network speeds to notebook computers, desktops, and workstations with great processing power, high resolution displays and fast networks. The information space consists of a wide variety of high quality information elements, e.g., text, audio, video, high-resolution graphics, etc.

caT extends the Trellis Petri net-based hypermedia model [2] and aims to provide optimal access to readers with respect to the computing power available to them, their context of use, and actions of other users in the environment. Trellis, and by extension caT, uses colored, timed Petri nets for document-based specification of the hypermedia structure as well as programmable browsing semantics. The graph structure of the net describes the nodes and links in the hypertext and the associated automaton semantics, when applied, specify the browsing behavior. A localized change to the document specification can be used to change its behavior. The Petri net structure incorporates parallelism, supports hierarchical net structures and can be formally analyzed for document as well as behavioral properties. The caT architecture also separates the specification of information elements to be displayed and their presentation. The Server applies user actions to the document structure and decides which information elements to display. The server-side Browser Manager routes these elements to client-based Browser coordinators, thus permitting a synchronized representation of user state on multiple clients. The coordinators supervise the display of the information content by invoking smart browsers, which retrieve the content in data formats that they can best render [1].

Most devices support few languages and scripts at the hardware level. The support for multiple scripts is usually provided via a combination of software and high quality

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 467-468, 2002.

displays. We describe a framework for temporary establishment of integrated environments consisting of user devices and publicly accessible computing resources in order to provide flexible access to rich multi-lingual digital library materials from handheld as well as computationally powerful user devices.

To support various languages effectively, caT must be aware of the capabilities of the devices in addition to those of the current browser and properties of the user as well as the environment that affect the browsing. caT stores device related information in a "Device Profile". Currently the Device Profile is a configuration file that describes device settings, both physical and operational. The Device Profile must know about the processor speed, size of memory, display size, number of colors the display can render and whether the display is graphical or textual. The configuration file also contains the operational properties, for example the access privileges (public, or restricted to a subset of users) and the sharable resources (display, printer, audio, etc.) available to each set of users. In the future, the Device Profile may incorporate a process that negotiates costs and resource availability when client devices request services from public access computing resources in various locations, for example, in hallways, hotel lobbies or airports.

Consider the following scenario in this environment. During a long layover between flights, a user browsing a digital library from a cell phone encounters a resource in Chinese that her small cell phone display cannot present coherently. She walks across to a nearby public access computing center. The caT-based cell phone extends her workspace seamlessly to a desktop computer and enables her to view the document on a larger screen. She finds the document interesting and decides to print it out for reading. The cell phone then extends her workspace to a nearby printer server that prints the document so she could read it on her next flight.

caT allows document-based creation as well as modification and easy analysis of the structure and behavior of digital information systems. We have presented a framework for the creation of temporary seamless environments that enable flexible access to digital library materials. The caT architecture facilitates the creation of such environments due to its separation of content, structure and presentation of information elements. The architecture or implementation makes no assumptions about the devices that readers may use for accessing digital collections.

Future work will consider models for negotiation and sharing of public-access resources in situations where the demand for resources may exceed the availability. Application of this architecture in a traditional library will open up exciting possibilities for transcending the physical and digital worlds. The application of this environment in a library setting could enhance user interaction with the resources and indexes, and must be investigated.

References

1. Karadkar, U. P., Na, J.-C., & Furuta, R. Employing Smart Browsers to Support Flexible Information Presentation in Petri net-based Digital Libraries. To appear in Proc. of the Sixth European Conf. on Digital Libraries, ECDL '02, LNCS, Springer-Verlag, 2002.
2. Stotts, P. D., & Furuta, R. Petri net-based Hypertext: Document Structure with Browsing Semantics. *ACM Transactions on Information Systems, 7*(1), 3-29, 1989.

Electronic Journal of the University of Malaya (EJUM): An Attempt to Provide a Truly Electronic Environment

A.N. Zainab, N.N. Edzan, and T.F. Ang

Dept. of Information Science, Faculty of Computer
Science & Information Technology, University of Malaya
{Zainab,edzan}@fsktm.um.edu.my

EJUM (Electronic Journal of the University of Malaya) is an Internet-based journal hosting system, specially designed to serve scholarly digital journal libraries. It currently hosts 13 issues (1996-2002) of the *Malaysian Journal of Computer Science* (indexed by INSPEC) and the *Malaysian Journal of Library and Information Science* (indexed by *LISAPlus* and *Library Literature*). Both journals are published semi-annually. This system aims: (a) To provide an avenue for scholarly journals to be hosted by a single system, making it easy for academic publishers to be involved in electronic publishing; (b) To provide participating academic journal publishers with the means to electronically manage article contributions; (c) To provide journal publishers with the facility to archive their older issues; (d) To provide academic publishers the means to electronically manage subscriptions; (e) To automate the editorial and refereeing process; (f) To provide users with efficient and varied search and retrieval options; (g) To provide users with a personalization service, which profiles their search interests and alerts them to new articles of interest; and (h) To provide user and access reports. A survey of e-journals in Malaysia reveals that very few publishers provide full-text access, much less a comprehensive search and retrieval function (Zainab and Edzan, 2000). Most e-journals are single journal systems and provide access to only abstracts. The first version of *EJUM* was developed in 1996 and was improved on in 2002.

Client Side. The Client side focuses on providing (a) users with varied search and retrieval options, (b) authors with a more transparent feedback system.
Registration. After registration (compulsory), a user can access the site immediately. Users are required to indicate their fields of interest and whether or not they want to receive email notification of new articles. Users can change their registration details, choose email service, and change preference categories.

User Profiling, Preferences and Alerting Service. When a registered user submits a search, the keywords submitted are stored and the list of retrieved articles kept so that he can refer to his search history when he next logs into the system. The system profiles the user by tracking the type of articles browsed and the keywords used. Uploads of new articles will automatically trigger an SQL job, which matches the new articles' keywords with keywords in the user's profile. He will be alerted via email if there is a match.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 469-470, 2002.

Search and Retrieval. The basic searches provided are by author's name, title keywords and broad subject categories. Users can limit their search to a particular journal or search all journals that *EJUM* host, limit their search to a particular year or range of years. They can also browse articles by selecting any alphabet in the index of authors' names, countries of origin and institutional affiliations. (The system automatically archives volumes published more than five years ago. The volumes in the archive are searchable and the same search options are given.) Users can sort results by title, keyword, author's name and year of publication. A new feature added in 2002 is searching the contents of full-text articles in PDF.

Article Contribution. An author who wishes to contribute an article has to register with the system as an author and submit his article via an online e-form (provided under the Author sub-module). The system triggers an automatic email message to editors. The author can check on the status of his article under review and for feedback by reviewers.

Administrator Side. The administrator's modules automate the editorial, refereeing, archiving and reporting processes.

Automating the Reviewing Process. Names and information about reviewers are first entered into the Review sub-module. From this list, the executive editor assigns two reviewers. *EJUM* automatically generates an email alert to the reviewers chosen. When a reviewer submits his e-evaluation form, *EJUM* automatically triggers an email alert to the executive editor. The author is automatically informed if his article is accepted, and will edit, modify and re-submit if the article requires amendments. If the article is rejected, the full-text of the article will be deleted.

Automatic Archiving. The administrator can specify the volumes to be archived automatically in order to speed up searching. (Currently, the system automatically archives volumes which are 5 years or older.)

Electronic reporting. The registration process allows *EJUM* to generate simple reports on users, who are sorted according to country as indicated in their email country extensions. A report on total users registered with the system is also available. Reports are generated via the Generate Report module.

Conclusion. The development of *EJUM* as a journal hosting system has reduced many of the mundane and tedious tasks associated with a paper-based system. It will stimulate electronic publishing of scholarly journals, making local research more accessible to students and researchers. It will contribute to the enrichment of Malaysia's info-structure, building a corpus of refereed science and technology journals available over the Internet. It has the potential to stimulate collaboration among Malaysian and Asian scholarly publishers. The challenge thus far has been to encourage authors and reviewers to fully utilize the electronic environment to submit and review articles.

References

1. Zainab, A.N. & Edzan, N.N. (2000). Malaysian Scholarly E-journals: Focus on EJUM, a Journal Management System. *Malaysian Journal of Library & Information Science*, 5(2), 69-84.

Patenting the Processes for Content-Based Retrieval in Digital Libraries

Hideyasu Sasaki[1,2] and Yasushi Kiyoki[3]

[1] Keio University, Graduate School of Media and Governance,
Zip Code 252-8520, 5322 Fujisawa, Japan
hsasaki@mdbl.sfc.keio.ac.jp
[2] Attorney-at-Law, New York State Bar
h-sasaki-7@alumni.uchicago.edu
[3] Keio University, Faculty of Environmental Information,
Zip Code 252-8520, 5322 Fujisawa, Japan
kiyoki@mdbl.sfc.keio.ac.jp
http://www.mdbl.sfc.keio.ac.jp/

Abstract. In this paper, we report formulation and case study of the conditions for patenting processes of content-based retrieval in digital libraries, especially in image libraries. Inventors and practitioners demand formulation of the conditions for patenting the processes as computer-related programs in combining prior disclosed means, and also in comprising the means for parameter settings to perform certain functions. Content-based retrieval indexes the extracted features of images and classifies the indexes to perform its retrieval function. A process for content-based retrieval often consists of a combination of prior disclosed means. That process also comprises the means for parameter settings that are adjusted to retrieve a specific kind of image at a certain narrow domain. We formulate the conditions of patentability on processes for performing content-based retrieval in combining the prior disclosed means and/or comprising the means for parameter settings from the practical standpoints of technical advancement (nonobviousness) and specification (enablement).

1 Introduction

Memory storage capacity has expanded tremendously to enable large scale storage of image data in digital libraries. The digital library community demands an automatic and scalable solution for retrieval function, especially in image libraries. After elaboration on keyword-based (or text-based) image retrieval, content-based image retrieval (CBIR) is introduced to the community of digital libraries. Keyword-based retrieval provides images in a digital library with prefixed index terms. Its search engine performs keyword-based retrieval function by pattern matching of the prefixed index terms to candidate images with a certain requested keyword.

Content-based retrieval indexes the extracted features of images and classifies the indexes to perform its retrieval function. Fig. 1 outlines the data-processing in a hypothetical example of geometric figure retrieval. The preprocessing step extracts each feature of the images, e.g., color, shape, edge, contour, etc. All images are converted

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 471–482, 2002.

to a set of several (four, in the example) binary data of extracted features from images. Images are then indexed by grouping the extracted features. The example indexes in the figure take the form of decimal data. A final processing step classifies the indexes dynamically every time users request a search. The retrieval results are ranked based on similarity at semantics that the indexes represent by computing correlation of the values constituting the indexes.

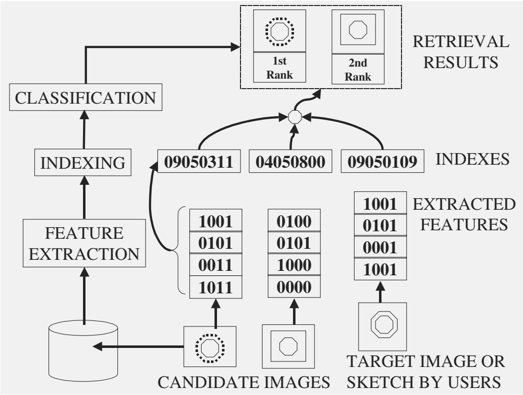


Fig. 1. A conceptual data-processing for content-based image retrieval.

CBIR has two classifications on retrieval approach at its applied domains: a domain-general approach and a domain-specific approach [1],[2]. A domain-general approach is applied at broad domains that include various kinds of features extracted from images. Its typical implementation includes Virage Image Retrieval [3] (U.S. Pat. # 5,893,095 invented by *Jain, et al.*), QBIC [4] (U.S. Pat. # 5,647,058 invented by *Agrawal, et al.*) and WebSeek [5] (also, VisualSEEK [6]). A domain-specific approach is applied at narrow domains that specify several features extracted from images. Those features represent the semantics of the target domains. Its typical implementation includes medical image retrieval systems (e.g., U.S. Pat. # 6,125,194 invented by *Yeh, et al.*) and fingerprint image retrieval systems (e.g., U.S. Pat. # 6,356,649 invented by *Harkless, et al.*).

Broad domains describe “their semantics only partially”, though specific domains restrict their “variability” on the extracted features of images in the “limited and predictable” scope [1]. Fig. 2 outlines the data-processing in the domain-specific approach using a hypothetical example of starfish image retrieval. The preprocessing step is almost same as the data-processing outlined in Fig. 1 but for the processing of restricted selection of features and classification comprising the means for “parameter settings”, i.e., the means for selecting and/or adjusting values on operating parameters. Its classification determines whether the classified indexes fall within certain predefined ranges of values on the parameter settings (e.g., for similarity computation of the indexes). The hypothetical ranges of values take ± 125 in a double-lined box as calculated from approximation of the ranges of values for each extracted feature represented in binary form.

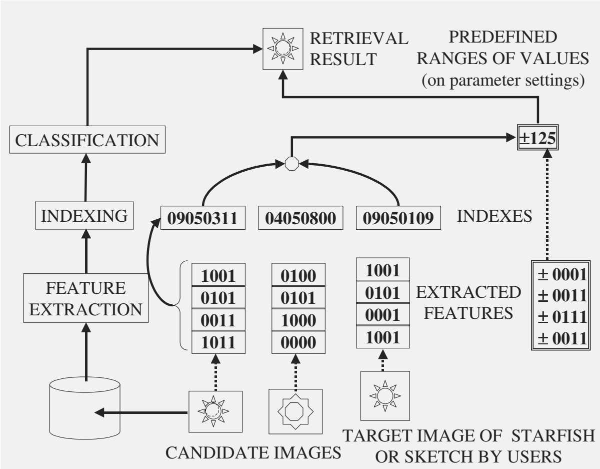


Fig. 2. A conceptual data-processing for a domain-specific approach of CBIR.

CBIR realizes its retrieval function by performing a number of “processes”, i.e., methods embodied in computer-related programs. Those processes for CBIR consist of combinations of means, some of which are prior disclosed inventions. Also, those processes comprise the means for parameter settings that are adjusted to retrieve a specific kind of image at a certain narrow domain for a domain-specific approach. A domain-specific approach implements its best mode by selecting values, such as weights, on parameter settings in order to perform specific functions for indexing and classifying extracted features of images at a certain narrow domain.

Inventors and practitioners demand formulation of the conditions for patenting those processes as computer-related programs in combining prior disclosed means, and also in comprising the means for parameter settings in order to perform certain functions. We formulate the conditions of patentability on a process as a combination of means and also on a process comprising the means for parameter settings in order to perform content-based retrieval from the practical standpoints of technical advancement (nonobviousness) and specification (enablement), respectively. The scope of this paper is restricted to nonobviousness and enablement on the processes.

The rest of the paper is organized as follows. In Section 2, we formulate the conditions on patentability. In Section 3, we provide a case study to confirm the feasibility of the formulated conditions. In Section 4, we conclude with discussions on our formulation.

2 Conditions of Patentability

Section 2 formulates the conditions of patentability on the processes for content-based retrieval of digital libraries, i.e., the processes for performing CBIR functions as computer-related programs in combining prior disclosed means, and also in comprising the means for parameter settings, as described in the block diagram in Fig. 3. The double-linedboxes represent the most critical conditions concerning combination of means and parameter settings.

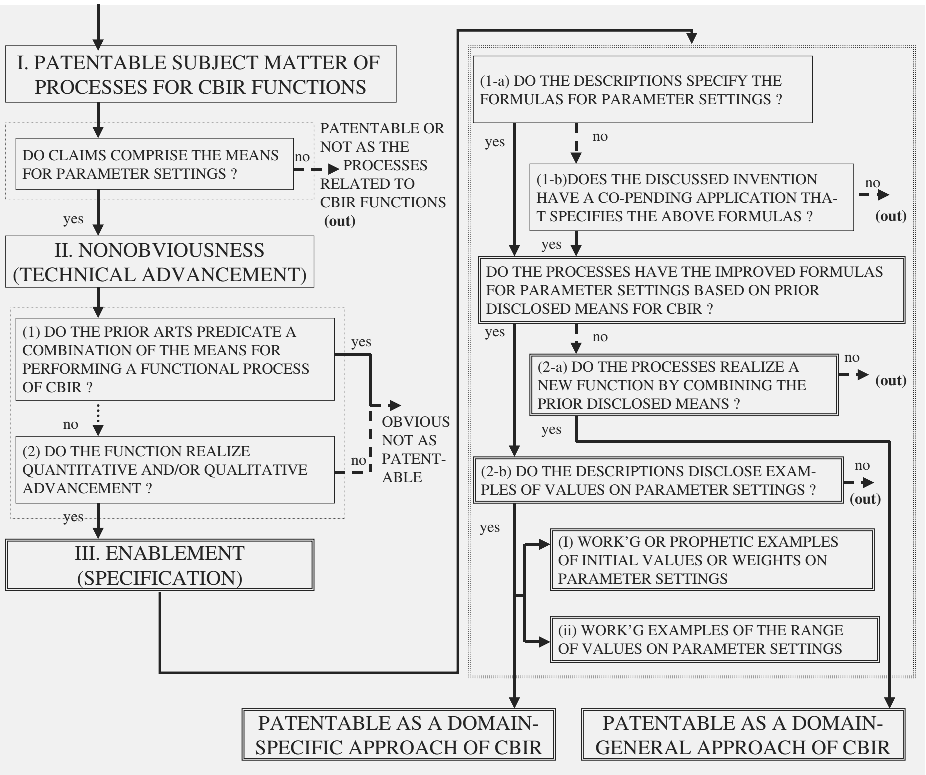

I. PATENTABLE SUBJECT MATTER OF PROCESSES FOR CBIR FUNCTIONS
DO CLAIMS COMPRISE THE MEANS FOR PARAMETER SETTINGS ?
no
PATENTABLE OR NOT AS THE PROCESSES RELATED TO CBIR FUNCTIONS
(out)
yes
II. NONOBVIOUSNESS (TECHNICAL ADVANCEMENT)
(1) DO THE PRIOR ARTS PREDICATE A COMBINATION OF THE MEANS FOR PERFORMING A FUNCTIONAL PROCESS OF CBIR ?
yes
no
OBVIOUS NOT AS PATENT-ABLE
(2) DO THE FUNCTION REALIZE QUANTITATIVE AND/OR QUALITATIVE ADVANCEMENT ?
no
yes
III. ENABLEMENT (SPECIFICATION)
(1-a) DO THE DESCRIPTIONS SPECIFY THE FORMULAS FOR PARAMETER SETTINGS ?
yes
no
(1-b)DOES THE DISCUSSED INVENTION HAVE A CO-PENDING APPLICATION THA-T SPECIFIES THE ABOVE FORMULAS ?
no
(out)
yes
DO THE PROCESSES HAVE THE IMPROVED FORMULAS FOR PARAMETER SETTINGS BASED ON PRIOR DISCLOSED MEANS FOR CBIR ?
yes
no
(2-a) DO THE PROCESSES REALIZE A NEW FUNCTION BY COMBINING THE PRIOR DISCLOSED MEANS ?
no
(out)
yes
(2-b) DO THE DESCRIPTIONS DISCLOSE EXAM-PLES OF VALUES ON PARAMETER SETTINGS ?
no
(out)
yes
(I) WORK'G OR PROPHETIC EXAMPLES OF INITIAL VALUES OR WEIGHTS ON PARAMETER SETTINGS
(ii) WORK'G EXAMPLES OF THE RANGE OF VALUES ON PARAMETER SETTINGS
PATENTABLE AS A DOMAIN-SPECIFIC APPROACH OF CBIR
PATENTABLE AS A DOMAIN-GENERAL APPROACH OF CBIR

Fig. 3. A block diagram for the formulated conditions of patentability on the processes for performing CBIR functions.

The formulated conditions consist of the following three requirements: "patentable subject matter" (entrance to patent protection), "nonobviousness" (technical advancement) and "enablement" (specification). Patentable subject matter is "the issue of which types of inventions will be considered for patent protection". Nonobviousness is "the technical advancement reflected in an invention". In general, nonobviousness is the ultimate condition of patentability, though enablement has received increasing attention as "the technical specification requirement" since the late eighties of the previous century [7]. The U.S. Patent and Trademark Office (USPTO) suggests in its *Guidelines* and *Training Materials* that a process comprising the means for parameter settings should comply with the requirement of enablement by giving examples of specific values on the parameter settings [8],[9].

2.1 Patentable Subject Matter

The processes for performing CBIR functions are patentable subject matter when their patent applications claim the means for parameter settings in order to perform certain retrieval functions.
Otherwise, the discussed processes are considered not as specific inventions of the processes for performing CBIR functions but as the inventions that are related to retrieval functions.

A process or method is patentable subject matter in the form of a computer-related program [10],[11] as "means-plus-function" [12] in the "specific machine to produce a useful, concrete, and tangible result ... for transforming ..." physical data ("*physical transformation*") [13]. A process must "perform independent physical acts (post-computer process activity)", otherwise, "manipulate data representing physical objects or activities to achieve a practical application (pre-computer process activity)". A process must not "merely manipulate abstract idea or solve a purely mathematical problem without any limitation to a practical application" [8],[9].

The processes for performing CBIR functions are patentable subject matter because they comprise the means for parameter settings and those means perform physical transformation of data, i.e., image retrieval processing. CBIR functions generate indexes as physical results on a computer and a memory, and also require pre-and-post computer process activities through data processing between feature extraction and indexing, also between indexing and classification, as indispensable procedure.

In the case of no claiming of the means for parameter settings, those processes do not realize any specific advancement of retrieval functions but facilitate certain functions that are related to data processing for CBIR.

2.2 Nonobviousness

The processes for performing CBIR functions are nonobvious from the prior arts when they
(1) comprise combinations of prior disclosed means to perform certain functions that are not predicated from any combination of the prior arts, in addition,
(2) realize quantitative and/or qualitative advancement.
Otherwise, the discussed processes are obvious so that they are not patentable as the processes for performing CBIR functions.

First, a combination of prior disclosed means should not be "suggested" from the disclosed means "with the reasonable expectation of success" [14]. Second, its asserted function must be superior to conventional functions realized in the prior disclosed or patented means. On the latter issue, several solutions for performance evaluation are applicable. Studies propose a benchmarking of CBIR functions [15],[16]. A patent application reports a suggested solution for performance evaluation by comparing the computational order of an instant process with one of a conventional disclosed process. Another general strategy is restriction of the scope of claims to a certain narrow field to which no prior arts have been applied.

2.3 Enablement

The processes for performing CBIR functions must be described clearly enough to enable those skilled in the arts to implement the best mode of those processes by satisfying the following conditions:
(1-a) The descriptions of the processes must specify the formulas for parameter settings. Otherwise, **(1-b)** the disclosed inventions have a co-pending application that describes the formulas in detail. In addition,
(2-a) the processes perform a new function by a combination of the prior disclosed means. And/or, **(2-b)** the processes have the improved formulas for parameter settings based on the prior disclosed means for performing CBIR functions and also give examples of the values on the parameter settings in the descriptions.

1-a and **1-b** determine whether the discussed processes are patentable combinations of the prior disclosed means for performing CBIR functions. **2-a** determines whether the discussed processes are patentable in the form of a domain-general approach. **2-b** determines whether the discussed processes are patentable in the form of a domain-specific approach when the processes specify the improved formulas for parameter settings based on the previous disclosed means.

For **2-b**, the processes must specify the means for parameter settings by "giving a specific example of preparing an" application [17],[18] to enable those skilled in the arts to implement their best mode of those processes without undue experiment. The processes comprising the means for parameter settings must disclose at least one of the following examples of values on the parameter settings:
(i) Working or prophetic examples of initial values or weights on the parameter settings
(ii) Working examples of the ranges of values on the parameter settings.
The "working examples" are the values that are confirmed to work at actual laboratory or prototype testing results [19]. The USPTO has also accepted in practice so-called "prophetic examples", especially in the area of biotechnology since the *In re Strahilevitz*. Those prophetic examples are given without actual work by one skilled in the art. Fig. 4 describes the hypothetical example that provides the example range of values, here weights on parameter settings, with, e.g., ±125. Its active range of values, e.g. ±100, fall within the example range of values.

It is a critical problem to define the scope of equivalent modification of process patents. The scope of patent protection extends to what is equivalent to a claimed invention as far as it is predictable from its claims and specification or its descriptions [20],[21]. Parametric values are easy to modify and adjust at application. A process for performing CBIR functions must specify its parameter settings by giving examples of the values on the parameter settings when they have the improved formulas for parameter settings based on the prior disclosed means. This specification is indispensable in the case of a domain-specific approach for performing CBIR functions. Those examples of values define the scope of the equivalent modification of a patented process within a certain specified scope as suggested from the examples of the values.

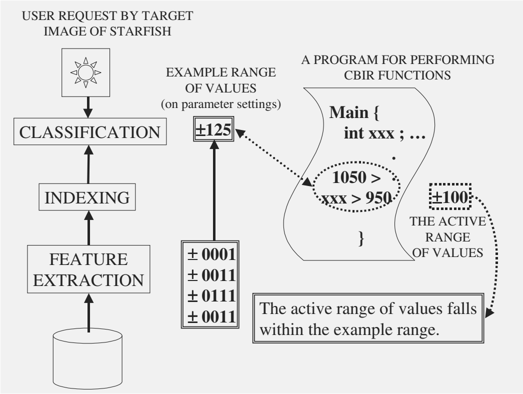


Fig. 4. A conceptual description of parameter settings.

3 Case Study

This Section confirms the feasibility of the formulated conditions of patentability on the processes for performing CBIR functions in digital libraries by applying the formulated conditions to several disclosed patents.

The "USPTO Patent Search Database" [22] responded to search terms "image(s)", "retriev(al, e, es, ed, ing)" and "database(s)" with 385 disclosed patents between September 5th, 2000 to May 28th, 2002. Those search terms are wide enough not to miss inventions performing CBIR functions but described in terms of pattern recognition. Table 1 lists eight patents that we selected from the above 385 patents for case study by screening out the other processes that do not perform CBIR functions. The first six cases take a domain-general approach while the remaining two cases take a domain-specific approach.

3.1 A Case of Patentable Subject Matter

Seven patents claim the means for parameter settings in their processes. The eighth is patentable not as a process for performing retrieval functions but as an invention related to CBIR functions. U.S. Pat. # 6,253,201 invented by *Abdel-Mottaleb, et al.* realizes a process for listing the pre-indexed image identifiers based on the prior arts of color and/or edge detection for performing fast classification. That invention is applicable to any process for CBIR as an invention related to retrieval function so that it should be determined whether it is patentable as a process related to the core CBIR functions.

3.2 A Case of Nonobviousness

A typical case of nonobviousness is the U.S. Pat. # 6,259,809 invented by *Maruo*, as described in the Table 2 . Its process realizes a new function of edge detection of the rotated, magnified or minified images of linear components, e.g., semiconductor wafers,

Table 1. A list of applied patents for case study.

Classification of inventions	**U.S. Pat. No. (Inventor(s))**	**Descriptions**
Patentable not as a process for performing retrieval functions but as an invention related to retrieval functions.	6,253,201 (*Abdel-Mottaleb, et al.*)	A process for listing the pre-indexed image identifiers based on the prior arts of color and/or edge detection for performing fast classification.
Patentable as the processes by combining the prior disclosed means to realize a domain-general approach for performing CBIR functions.	6,115,717 (*Mehrotra, et al.*)	A process for performing image recognition based on the objects and regions of the images restored in open spaces of text and image commingling multimedia.
	6,263,089 (*Otsuka, et al.*)	A process for retrieving, by the Hough transformation, the appearing and disappearing motion pictures based on texture and edge of images, e.g., images of clouds for weather forecasting.
	6,259,809 (*Maruo*)	A process for performing edge detection of the rotated, magnified or minified images of linear components, e.g., semiconductor wafers, by the Hough transformation and thresholding on similarity computation.
	6,240,423 (*Hirata*)	A process for performing object-based CBIR by a twofold processing, region detection and boundary detection, based on another process patented by the same inventor.
	6,246,804 (*Sato, et al.*)	A process for performing edge and color based region detection in the compound or over-wrapping movable regions of images by combining the conventional arts of boundary line detection.
Patentable as the processes that have the improved formulas for parameter settings based on the prior disclosed means for performing CBIR functions and that also give the examples of the values on parameter settings in the descriptions.	6,356,649 (*Harkless, et al.*)	A process for performing CBIR functions of fingerprints that rotate or dilate, translate or distort by improving the formulas disclosed in another prior patent granted to one of the instant inventors, *Thebaud, et al.*, with working examples of initial values on the thresholds and the assumed ranges of rotation and dilation.
	6,125,194 (*Yeh, et al.*)	A process for performing CBIR functions of radiological images of lung nodules by re-screening once diagnosed negative images in neural network with the working and prophetic examples of initial weights on its back-propagation.

without predetermined angles for rotation, etc. by combining the prior arts on the Hough transformation and thresholding on similarity computation. Its description asserts its superior performance by comparing the computational order of their instant method with one of a conventional disclosed method.

Table 2. U.S. Pat # 6,259,809 (Inventor; *Maruo*).

Conditions	**Determination**
Does its patent application claim the means for parameter settings ?	Yes. (e.g., means for calculating the mean value of picture element values, thresholds for providing binarization of image data, representative point calculation means for determining the coordinates of a representative point of each group, parameter space with weight)
Do the prior arts predicate the instant combination of prior disclosed means for performing a functional process ?	No. It realizes a functional process for detecting the images of flexible rotation or dilation angles. It is not predicated from any combination of the conventional means.
Does the function realize quantitative and/or qualitative advancement?	Yes. Both advancements are realized by the processing.
Do the descriptions of a process specify the formulas for parameter settings ?	Yes.
Does the process have the improved formulas for parameter settings ?	No.
Does the process realize a new function by combination of the prior disclosed means ?	Yes.

3.3 Cases of Enablement

The processes for performing CBIR functions must specify the formulas for parameter settings. Otherwise, the disclosed inventions have their co-pending applications that describe the formulas in detail. In addition, the processes must realize new functions by combination of the prior disclosed means when those processes do not have any improved formulas for parameter settings.

Its typical case is U.S. Pat. # 6,115, 717 invented by *Mehrotra, et al.*, as described in Table 3 . That invention performs image recognition based on the objects and regions of the images restored in open spaces of text and image commingling multimedia by automatic metadata indexing. First, that process does not specify any formulas for parameter settings, though their formulas are disclosed in its co-pending application of Ser. No. 08/786,932. Second, that invention realizes new functions of open space metadata indexing and region recognition by combining the prior disclosed means.

A domain-specific approach not only performs a new function by a combination of prior disclosed means but also has the improved formulas for parameter settings based on the prior disclosed means in order to perform CBIR functions. In this case, the descriptions of the process for a domain-specific approach must give the parameter settings with examples of the values that are the working or prophetic examples of initial values or weights, otherwise, the working examples of the range of values.

Its typical case is U.S. Pat. # 6,356,649 invented by *Harkless, et al.*, as described in the Table 4 . That invention performs retrieval of fingerprint images that often rotate or dilate, translate or distort. That process has the improved formulas for parameter

Table 3. U.S. Pat. # 6,115,717 (inventors; *Mehrotra, et al.*).

Conditions	**Determination**
Does its patent application claim the means for parameter settings ?	Yes. (e.g., average, variance, range of the color component pixel values, means for depicting properties of open space)
Do the prior arts predicate the instant combination of prior disclosed means for performing a functional process ?	No.
Does the function realize quantitative and/or qualitative advancement?	Yes. (Both: automatic metadata indexing realizes faster and more precise retrieval functions.)
Do the descriptions of a process specify the formulas for parameter settings ?	No.
Does the discussed patent have a copending application that specifies the above formulas ?	Yes. (Ser. 08/786,932)
Does the process have the improved formulas for parameter settings ?	No.
Does the process realize a new function by combination of the prior disclosed means ?	Yes. (Realized as image recognition based on the objects and regions of the images restored in open spaces of text-image-commingling multimedia)

settings based on the prior disclosed means as claimed in U.S. Pat. # 5,909,501 invented by *Thebaud, et al.*, who is one of the instant inventors. The process detects the rotated or dilated images of fingerprints without predetermined angles for the translation of those images. The discussed invention, *Harkless, et al.*, discloses its detailed embodiments in the appendix of its descriptions by giving the working examples of the ranges of values on parameter settings, i.e., the thresholds and the assumed ranges of rotation and dilation for correlation computation, though they are not disclosed in the prior patent application of *Thebaud, et al.*. The working examples provide the assumed ranges of rotation and dilation values that are indispensable to implementation of the best mode of the patented processes.

4 Discussion and Conclusion

In this paper, we formulate the conditions of patentability on a process for performing content-based retrieval functions as a combination of prior disclosed means and also on a process comprising the means for parameter settings. CBIR is promising for retrieving images in a large-scale digital library. In particular, its domain-specific approach has received attention and elaboration in several areas including the following: medical retrieval systems of radiological or MRI images of nodules in brains, abdomens, intestines, etc.; deficit image retrieval in artificial manufactures including semiconductor wafers; and authentication based on human skin patterns, etc. Our formulation is unique in the

Table 4. U.S. Pat. # 6,356,649 (Inventors; *Harkless, et al.*).

Conditions	**Determination**
Does its patent application claim the means for parameter settings ?	Yes. (e.g., assumed rotation and dilation, and distortion of images, means for rationing the respective ridge-spacing and orientation values, means for correlating the two transformed power spectral densities, a normalized spatial correlation value at similarity correlation)
Do the prior arts predicate the instant combination of prior disclosed means for performing a functional process ?	No.
Does the function realize quantitative and/or qualitative advancement?	Yes. (Precise qualitative performance by rotation and distortion detection.)
Do the descriptions of a process specify the formulas for parameter settings ?	Yes.
Does the process have the improved formulas for parameter settings ?	Yes.
Do the descriptions of the process give examples of the values on parameter settings in the descriptions ?	Yes. (With working examples of initial values on the thresholds and also on the assumed ranges of rotation and dilation for correlation computation.)

sense that it is based on process patents as computer-related programs which receive attention as patents of the "methods of doing business", i.e., business model patents. A combination of the prior disclosed means and parameter settings are critical components in determining patentability on processes. The case study shows the feasibility of our formulated conditions of patentability that facilitate patenting those functional processes of CBIR. We are preparing an extended version of this paper based on a larger scale case study in our paper [23].

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. on Pattern Analysis and Machine Intelligence **22(12)** (2000) 1349–1380
2. Rui, Y., Huang, T.S., Chang, S.-F.: Image Retrieval: Current Techniques, Promising Directions and Open Issues. J. of Visual Communication and Image Representation **10(4)** (1999) 39–62
3. Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Jain, R., Shu, C.F.: The Virage Image Search Engine: An Open Framework for Image Management. In: Proc. of SPIE Storage and Retrieval for Still Image and Video Databases IV. San Jose, CA, USA (1996) 76–87
4. Flickner, M., Sawhney, H., Niblack, W.: Query by Image and Video Content: The QBIC System. IEEE Computer **28(9)** (1995) 23–32
5. Smith, J.R., Chang, S.F.: Visually Searching the Web for Content. IEEE Multimedia **4(3)** (1997) 12–20

6. Smith, J.R., Chang, S.F.: Querying by Color Regions Using the VisualSEEK Content-Based Visual Query System. In: Maybury, M.T. (ed.): Intelligent Multimedia Information Retrieval. (1996)
7. Merges, R.P.: Patent Law and Policy: Cases and Materials. 2nd edn. Contemporary Legal Education Series. The Michie Company (1997)
8. U.S. Patent and Trademark Office : Examination Guidelines for Computer-Related Inventions, 61 Fed. Reg. 7478 (Feb. 28, 1996) ("*Guidelines*"). Available via WWW from *http://www.uspto.gov/web/offices/pac/dapp/oppd/patoc.htm*" (1996)
9. U.S. Patent and Trademark Office: Examination Guidelines for Computer-Related Inventions Training Materials Directed to Business, Artificial Intelligence, and Mathematical Processing Applications ("*Training Materials*"). Available via WWW from *http://www.uspto.gov/web/offices/pac/compexam/examcomp.htm* (1996)
10. Title 35 U.S.C. §§101, 103, & 112 (US Patent Act)
11. Jakes, J.M., Yoches, E.R.: Legally Speaking: Basic Principles of Patent Protection for Computer Science. Comm. of the ACM **32(8)** (1989) 922–924
12. In re Trovato 42 F.3d 1376, 33 U.S.P.Q.2d 1194(Fed. Cir. 1994), cavated & remanded, 60 F.3d 807, 33 U.S.P.Q.2d 1194(Fed. Cir. 1995)(en banc)
13. In re Alappat 33 F.3d 1526, 31 U.S.P.Q.2d 1545(Fed. Cir. 1994)(en banc)
14. In re Dow Chemical Co. 837 F.2d 469, 473, 5 U.S.P.Q.2d 1529, 1531 (Fed. Cir. 1988)
15. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Automated Benchmarking in Content-Based Image Retrieval. In: Proc. of the 2001 IEEE International Conference on Multimedia and Expo (ICME 2001) Tokyo, Japan (2001)
16. Manchester Visualization Center: CBIR Evaluation. Available via WWW at *http://www.man.ac.uk/MVC/research/CBIR/* (2000)
17. Autogiro Co. of America v. United States 384 F.2d 391, 155 U.S.P.Q. 697 (Ct. Cl. 1967)
18. Unique Concepts, Inc. v. Brown 939 F.2d 1558, 19 U.S.P.Q.2d 1500 (Fed. Cir. 1991)
19. In re Strahilevitz 682 F.2d 1229, 212 U.S.P.Q. 561 (C.C.P.A. 1982)
20. Graver Tank & Mfg. Co. v. Linde Air Products Co. 339 U.S. 605 (1950)
21. Laitran Corp. v. Rexnord, Inc. 939 F.2d 1533, 19 U.S.P.Q.2d (BNA) 1367 (Fed. Cir. 1991)
22. USPTO Patent Full-Text and Image Database. Available via WWW at *http://www.uspto.gov/* (2002)
23. Sasaki, H., Kiyoki, Y.: The Patent Application Methodology of a Multimedia Database Search Engine Comprising Parameter Settings for Optimizing Data Processing Mechanisms as a Patentable Computer-Related Invention. IPSJ Trans. on Databases **42(11)** (2001) 22–38 (*in Japanese*)

Secure Content Distribution for Digital Libraries[*]

Mariemma I. Yagüe, Antonio Maña, Javier López, Ernesto Pimentel, and
José M. Troya

Computer Science Department
University of Málaga. Spain.
{yague,amg,jlm,ernesto,troya}@lcc.uma.es

Abstract. Security is a very relevant aspect in the implementation of most digital libraries. Two important security issues in these systems are distributed access control and secure content distribution. This paper presents a system that provides these two features for digital libraries. The system is based on the combination of the use of an external authorization infrastructure, a software protection mechanism and a modular language to specify how access to the contents is controlled. The extensive use of semantic information makes possible the integration of such a complex set of subsystems.

1 Introduction

Digital Libraries (DLs) integrate a variety of information technologies that provide opportunities to assemble, organize and access large volumes of information from multiple repositories. Regarding access control, DLs usually contain objects with heterogeneous security requirements. The problems of controlling access to such objects present many similarities to those found in other distributed systems. However, some of the problems are specific, or more relevant, for DLs. Some of those problems are: (i) libraries are available to previously unknown users; (ii) payment or other activities (like the execution of copyright agreements) must be bound to access to the objects; (iii) the originator or owner of the object must retain control over it even after it is accessed by users; (iv) a high degree of flexibility is required because of the heterogeneous nature of the objects; (v) the ability to change the access control parameters dynamically and transparently is also essential in most DLs; and finally (vi) due to the large amount of objects, it is important to establish access conditions automatically, based on information about objects.

This paper presents XSCD-DL (XML-based Secure Content Distribution for Digital Libraries), a particular application of the XSCD infrastructure [1]. XSCD-DL provides distributed access control management and enforcement, as well as secure content distribution in digital libraries. To achieve our goals we combine an external authorization infrastructure, a modular language called *Semantic Policy Language* (SPL) to specify how access to the contents is controlled, and a software protection

[*] Work partially supported by Spanish Ministerio de Ciencia y Tecnología. Project TIC2002-04500-C02-02

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 483-494, 2002.

mechanism (*SmartProt*). The extensive use of semantic information makes possible the integration of such a complex set of subsystems into XSCD-DL.

The rest of the paper is organized as follows. Section 2 summarizes some related work. Section 3 describes the fundamentals and global structure of XSCD-DL. Finally, section 4 summarizes the conclusions and presents ongoing and future work.

2 Related Work

Different projects, such as the ADEPT Digital Library [2] and the Alexandria Digital Library [3], have focused on handling various structural and semantic issues, while providing users with a coherent view of a massive amount of information. The use of metadata is essential in these systems, but the application to security issues is not considered. The Stanford Digital Library Project [4] covers most of the different issues involved in this field. One important outcome is the FIRM architecture [5] that proposes the separation of objects that implement control from objects that are controlled, enhancing the flexibility of the system.

Regarding access control, several proposals have been introduced for distributed heterogeneous resources from multiple sources [6][7]. Unfortunately, these proposals do not address the specific problems of distributed access control in DLs. Traditional access control schemes such as *mandatory access control* (MAC), *discretionary access control* (DAC) or even *role based access control* (RBAC) are not appropriate for complex distributed systems such as digital libraries. It has been shown that an approach based on attribute certificates represents a more general solution that fits more naturally in these scenarios [8]. In fact, MAC, DAC and RBAC schemes can be specified using the attribute-based approach.

Because of the specific requirements imposed on the access control systems of digital libraries, the most widespread architecture is that of a federated set of sources, each with a centralized access control enforcement point. This architecture has important drawbacks such as the reduced system performance produced because the centralized access control enforcement point becomes a bottleneck for request handling. Other drawbacks are that (a) the control point represents a weak spot for security attacks and fault tolerance, (b) it does not facilitate the deployment of owner retained control mechanisms, and (c) it usually enforces homogeneous access control schemes that do not fit naturally in heterogeneous user groups and organizations.

On the other hand, distributed access control solutions proposed so far do not provide the flexibility and manageability required. An interesting approach based on the concept of mobile policies [9] has been proposed to solve some of the limitations of RBAC schemes [10]. This is a limited improvement because of the requirement to execute the access control policies in trusted computers (object servers in this case). Furthermore, when access to an object is granted, this object has to be sent to the client computer where control over it is lost. Finally, because object and policy are compiled in a package, any single change in the policy that controls an object requires that the object-policy package be recompiled and distributed to all trusted servers.

Several XML based languages have been developed for access control, digital rights management, authentication and authorization. These languages do not support powerful features such as policy modularisation, parameterisation and composition.

Furthermore, some of their features are not necessary in DLs [11]. Two relevant proposals are the Author-X system [12] and the FASTER project [13][14], which propose two similar systems for access control to XML documents. Both systems define hierarchic access control schemes based on the structure of the document. The FASTER system does not support any content protection mechanism. FASTER access control is based on user groups and physical locations following the well-known technique of defining a subject hierarchy. In scenarios such as digital libraries, this approach is not adequate because a fixed hierarchy cannot represent the security requirements of all the different contents, users and access criteria. On the other hand, content protection in Author-X is founded on the concept of (passive) secure container, which introduces disadvantages from the point of view of security and access control system management. Author-X is based on credentials that are issued by the access control administrator. Therefore, in practice, each credential will be useful only for a single source, limiting interoperability. A direct consequence of this approach is that users must subscribe to sources before they can access their contents.

3 XSCD-DL Fundamentals and Global Structure

New solutions are required to address the need of some of the new distributed applications such as DLs, web services or grid computing. Some of the problems found in existing access control systems are:

- The security administration in current DLs is very complex and error prone. A flexible and powerful policy language that incorporates features to manage the complexity inherent in DL environments represents a step towards the solution of this problem. Automated management tools also serve this objective.
- The explicit static allocation of policies to objects is not adequate in highly dynamic environments with heterogeneous contents, where new resources are often added to the system and security requirements change frequently. In order to solve this problem, dynamic allocation of policies to resources and the definition of an access policy language designed to ease the management of the system must be considered.
- The access control criteria are usually defined either explicitly or on the basis of the structure of the contents. These approaches present severe drawbacks as we will show later.
- In new environments we deal with a large number of (possibly anonymous) users. Existing schemes, based on user identity, need to collect some profile information in advance. Therefore, a registration phase is used to collect information about the user and issue the corresponding local credentials. The semantic integration of an external PMI represents a step towards the solution of the interoperability of different DLs with heterogeneous access control systems.
- The access policy depends on the administrator of the server where the object is stored. In DL environments, it would be desirable that originators of the contents are able to define the access policy to apply in a dynamic and transparent way, regardless of the object's storage location.
- Finally, no secure content distribution mechanisms are used.

Because our system includes different elements that are combined to solve some of these problems, in the following subsections we will focus on each of the problems.

3.1 Modular Language for Flexible Security Administration

XML is widely considered a suitable candidate for a policy specification language [7]. Existing XML-based languages for access control, authorization and digital rights management are not based on a modular approach and do not provide some important features such as policy composition and parameterisation. These ones play an essential role in the flexibility of management of the access control system [11].

The definition of access control policies is a complex and error prone activity that presents many similarities with computer programming. Therefore, we have included some of the mechanisms used to reduce the complexity in programming languages such as modularity, parameterisation and abstraction. In order to provide the simplicity and flexibility required in complex systems such as digital libraries, our solution is based on the modular definition of policies. Modularity in our solution implies: (a) the separation of specification in three parts; that is, access control criteria, allocation of policies to resources and semantic information (properties about resources and context); (b) the abstraction of access control components; (c) the ability to reuse these access control components; and (d) the reduction of the complexity of management due to previous properties. Moreover, the use of semantic information about the context allows the administrator to include contextual considerations in a transparent manner, also helping the (semantic) validation task.

Usual components of access policies include the target resource, the conditions under which access is granted/denied and, sometimes, access restrictions. Opposed to other languages, SPL policy specifications do not include references to the target object. Instead, a separate specification called *Policy Applicability Specification* (PAS) is used to relate policies to objects dynamically when a request is received. Both SPL policies and PAS use semantic information about resources included in *Secured Resource Representation* (SRRs) and other contextual information documents, which is an original contribution. SPL policies and PAS can be parameterised allowing the definition of flexible and general policies and reducing the number of different policies to manage. Parameters, which can refer to complex XML elements, are instantiated dynamically from semantic and contextual information. Finally, policies can be composed importing components of other policies without ambiguity. This compositional approach allows us to define the abstract meaning of the elements of the policies, providing a mechanism to achieve abstraction, which also helps in reducing the complexity of management. Tools to graphically manage the relations among policies and with other components are also essential for a simple and flexible management.

The schema for SPL specifications is represented as a set of XML-Schema [15] templates that facilitate the creation of these specifications, allowing their automatic syntactic validation. Figure 1 shows the structure of the SPL language.

SPL policies can include locally defined components as well as imported elements. The ability to import elements enables the modular composition of policies based on the XPath standard [16]. An SPL Policy is composed of a set of *access_Rule* elements, each one defining a particular combination of attribute certificates required to gain access, associated with an optional set of actions (such as *Notify_To*, *Payment* and *Online_Permission*) to be performed before access is granted. In this way provisional authorization is enabled in SPL.

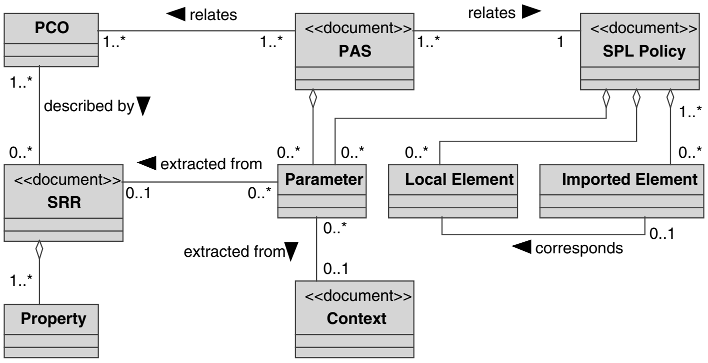


Fig. 1. Conceptual model of the SPL Language

The *Policy Applicability Specification* provides an expressive way to relate policies to resources, either explicitly or based on the metadata about the objects (e.g. type of content, owner, price, etc.). PAS documents include three main elements: policy, objects and instantiation. The policy element indicates which policy is applicable to the specified objects. Objects are defined by their location and conditions to be fulfilled by the semantics of these objects (SRRs). Optionally, operation elements can be used to define which operations of the target object are controlled by the declared policy, allowing a finer grained access control. In case no operation element is included, the policy is applicable to all of the object operations. The instantiation element describes the mechanism to instantiate parameters in the policies. Figure 3 shows an example of applicability rules for SPL policies to objects.

The *Secured Resource Representation* is a simple and powerful mechanism to describe properties about resources. Properties described in SRRs are used for the instantiation of policies and PAS, and to locate the applicable policies. An example of an SRR is also included in figure 3. The SRR is designed specifically for the process of dynamic allocation of policies to resources. Dynamic allocation is a very flexible and useful mechanism that solves the problem of associating policies to newly created objects. The use of dynamic policy allocation needs a rich set of metadata about the resources. This semantic meta-model is used to locate the right policy for each resource, based on its relevant properties.

3.2 Local Credentials versus External PMI

Most of current access control schemes base their authorization approaches on locally issued credentials that are linked to user identities. This type of credentials present many drawbacks. Among them we highlight: (a) they are not interoperable; (b) the same credentials are issued many times for each user, what introduces management and inconsistency problems; (c) credentials are issued by the site administrator; however, in most cases, the administrator does not have enough information or resources to establish trustworthy credentials; and (d) they are tight to user identity. In practice, it is frequent that the identity of the user is not relevant for the access decision. Sometimes it is even desirable that the identity is not considered or revealed. Furthermore, in systems based on identity, the lack of a global authentication infrastructure (a global PKI) forces the use of local authentication schemes. In these cases, subscription is required and users have to authenticate themselves to every accessed source. To solve the aforementioned problems, single-sign-on mechanisms have been used in last years. These mechanisms are based on federation of sources that represent a limited improvement because credentials remain local (not to a site, but to a set of them). Moreover, all federated sources must agree on a homogeneous access control scheme.

On the other hand, digital certificates can securely convey authorizations or credentials. Attribute certificates bind attributes to keys providing means for the deployment of scalable access control systems in the scenarios that we have depicted. These authorizations are interoperable and represent a general and trustworthy solution that can be shared by different systems. Taking into account security, scalability and interoperability, the separation of the certification of attributes and access control management responsibilities is widely accepted as a scalable and flexible solution. In this case, the access control system needs to be complemented by an external component: the *Privilege Management Infrastructure* (PMI)[17]. The main entities of a PMI, known as *Source of Authorizations* (SOAs), issue attribute certificates. Usually, each SOA certifies a small number of semantically related attributes. This scheme scales well in the number of users and also in the number of different factors (attributes) used by the access control system.

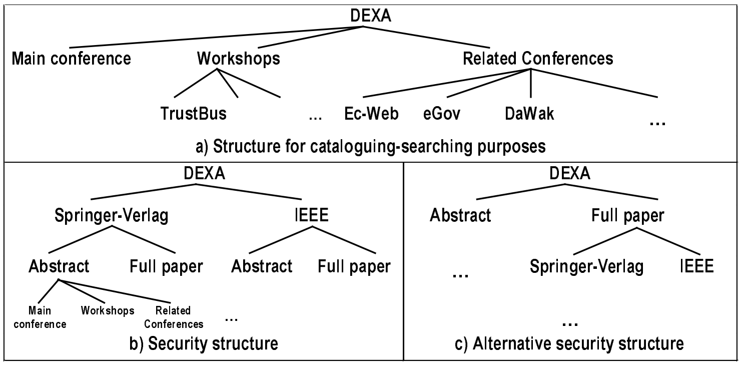
With this approach, each access control system will select which SOAs to trust and which combination of attributes to use. Because they are separate systems, a mechanism to establish the trust between the access control and the PMI is required. Metadata described about the PMI, represented as *Source Of Authorization Description* (SOAD) documents, is the key to achieve the necessary interoperability. SOADs are RDF [18] documents protected by digital signatures [19] that express the different attributes certified by each SOA, including their names, descriptions and relations. These descriptions state a series of facts about the environment of the system using metadata to represent the semantics of the different attributes that are certified by the SOA, including names, descriptions and relations among attributes.

In our scheme SOADs represent the semantic description mechanism that establishes trust between the PMI and the access control system. The semantic information about the certificates issued by each SOA is also used to assist the security administrators in the creation of access control policies. Additionally this semantic information allows the detection of possible inconsistencies in our SPL policies, during the semantic validation process.

3.3 Structure versus Semantics as the Basis of the Access Control Scheme

Usually, conditions and restrictions of access depend on the semantic properties of the target object that are neglected in structure-based approaches. Because the security requirements for each object in the DL depend on different properties about the object, an approach based on semantic descriptions of the contents is much more flexible and natural. Moreover, it is easy to incorporate structure-based requirements in the semantic model. Finally, the structure is much more volatile than the semantics. In particular, when the contents are XML documents, the structure is dynamic and flexible, introducing serious problems in the security administration. Some works have appeared to automate the process of translating the authorizations for the transformed XML document, although the proposed solutions are very limited [20].

In order to illustrate the advantages of content-semantics over content-structure as the basis for the policy specification, lets consider the case of a digital library of proceedings of scientific conferences. Usually, titles, authors and abstracts of all papers are public while full papers have different access control requirements depending on the publisher and the type of paper. Figure 2 shows how the ideal structuring of the contents for cataloguing and searching purposes does not match the one for security requirements. Moreover, two different possibilities appear in the latter case, illustrating that the structure-based approach is not adequate for the case of digital libraries, where contents have heterogeneous security requirements. The incompatibility between the structure required for the application domain and the ones that match the security requirements confirms that structure-based approaches are not able to represent these situations in a natural way. Furthermore, if we were to use the structuring in figure 2a then multiple authorizations of opposite sign (grant/deny) would be required to match the security requirements of each piece of content.



a) Structure for cataloguing-searching purposes

b) Security structure

c) Alternative security structure

Fig. 2. Different structuring of the contents of a digital library

Another drawback of structure-based approaches is that the number of policies becomes very large. In fact, these approaches usually imply the definition of several policies for each resource. Positive and negative authorizations are used in these cases to facilitate the definition of simple policies and to reduce the number of policies. The price to pay is the introduction of ambiguities, which in turn requires the definition of conflict resolution rules. Consequently, the administration of the system becomes complex and difficult to understand increasing the chance of producing incorrect policies. The semantic-based and modular approach adopted in XSCD-DL facilitates the definition and management of policies avoiding the use of positive and negative authorizations. Tools provided to support the policy specification, composition and validation also serve this objective.

The semantic-based approach adopted in XSCD-DL is illustrated in figure 3. It shows how the use of semantic information, the modularisation, parameterisation and dynamic instantiation of policies results in simple and flexible policies reducing the complexity of management of the system. The specifications also demonstrate how the system can manage dynamic changes in a transparent way. No modifications are necessary in the policies when requirements or properties of resources change.

```
<?xml version="1.0" encoding="UTF-8"?>
<spl:PAS xmlns:spl="http://www.lcc.uma.es/ICADL"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.lcc.uma.es/ICADL pas.xsd">
      <spl:parameter>proceeding_Publisher</spl:parameter>
      <!--   PAS parameters are instantiated directly from the target SRR,
therefore no instantiation tag is required -->
      <spl:policy>FullPaper.xml</spl:policy>
      <spl:object>
          <spl:object_Location>
                  http://www.dexa.org/2002/ecweb/
          </spl:object_Location>
          <spl:conditions>
              <spl:condition>
                  <spl:property_Name>type</spl:property_Name>
                  <spl:property_Value>full_Paper</spl:property_Value>
              </spl:condition>
              <spl:condition>
                  <spl:property_Name>Publisher</spl:property_Name>
                  <spl:property_Value>*paper_Publisher</spl:property_Value>
              </spl:condition>
          </spl:conditions>
      </spl:object>
      <spl:instantation>
          <spl:formal_Parameter>Publisher</spl:formal_Parameter>
          <spl:actual_Parameter
      path="//Publisher_Info[@name='*proceeding_Publisher']/parent::">
          Publishers_Context_Info.xml</spl:actual_Parameter>
      </spl:instantation>
<!--  Publisher_Context_Info contains details about each Publisher -->
</spl:PAS>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<spl:policy xmlns:spl="http://www.lcc.uma.es/ICADL"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.lcc.uma.es/ICADL Policy.xsd">
<spl:parameter>Publisher</spl:parameter>
<spl:access_Rules>
      <spl:access_Rule>
              <spl:attribute_Set>
                  <spl:attribute>
              <spl:attribute_Name>Registered</spl:attribute_Name>
              spl:attribute_Value>*Publisher[@name]</spl:attribute_Value>
              <spl:SOA_ID>*Publisher[@SOA]</spl:SOA_ID>
                  </spl:attribute>
              </spl:attribute_Set>
          </spl:access_Rule>
      </spl:access_Rules>
</spl:policy>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<spl:SRR xmlns:spl="http://www.lcc.uma.es/ICADL"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.lcc.uma.es/ICADL SRR.xsd"
resource="http://www.dexa.org/2002/ecweb/Lopez_acdacs.pdf">
<!-- This SRR states access control about the paper Lopez_acdacs.pdf  -->
   <spl:property>
       <spl:property_Name>type</spl:property_Name>
       <spl:property_Value>full_Paper</spl:property_Value>
   </spl:property>
   <spl:property>
       <spl:property_Name>proceeding_Publisher</spl:property_Name>
       <spl:property_Value>Springer_Verlag</spl:property_Value>
   </spl:property>
   <!-- e-mail of the responsible -->
   <spl:property>
       <spl:property_Name>responsible</spl:property_Name>
       <spl:property_Value>manager@ecweb.com</spl:property_Value>
   </spl:property>
</spl:SRR>
```

Fig. 3. Example Policy, its corresponding PAS and the SRR for a 'full paper'

3.4 Content Protection

Two important issues arise when considering content protection in digital libraries: the content distribution mechanism itself and the owner-retained-control issue. The first one must ensure that contents are protected so that only the intended recipients can access them. In the case of digital libraries it also entails other requirements such as the need to bind the execution of digital rights agreements, payment or other actions to the access to the contents. This is known as provisional authorization or *provision-based access control* (PBAC) [21]. The second one deals with enabling that owners of the contents retain control over them even when contents are stored in external untrusted servers.

Our solution to the previous problems is based on the use of secure active containers. A *secure active container* [22] is a piece of protected mobile software that conveys the contents and forces the user to fulfil the applicable policy before access is granted. By "protected software" we mean that it is neither possible to discover nor to alter the function that the software performs and it is also impossible to impersonate the software. In our scheme, this is achieved using a variant of the *SmartProt* system [1]. *SmartProt* partitions the software into functions that are executed by two collaborating processors. One of those processors is a trusted computing device that enforces the correct execution of the functions and avoids that these functions are identified or reverse engineered. We are currently using smart cards for this purpose although other alternatives are possible. Our secure active containers are implemented as Java™ applets that we call *Protected Content Objects* (PCOs). They include the contents to be accessed (which are encrypted), the access control enforcement mechanism, and a cryptographic link to the *Mobile Policy* (MP) required to gain access to the contents. We extend the concept of mobile policy described in [9] by allowing their execution in untrusted systems. Moreover, in our solution policies are bound to the object but not integrated with. This modification makes possible that policies are dynamically changed in a transparent manner. The definition of the MP structure allows a high degree of flexibility.

The PCO generation process is independent of the customer card and will be performed just once for each piece of content. PCOs can be distributed and copied freely. One important constraint to the free distribution of protected contents in our system is that originators of those contents must be able to dynamically change the applicable access control policy regardless of the storage location of the PCO. In order to fulfil this requirement, MP and PCO must be separated. In this way, the MP is retrieved from the originator DL during the execution of the PCO. Requesting the MP at access time from the originator slightly reduces the performance of the system but, in return, it allows a high degree of flexibility and gives the originator more control over the application of the policies. To improve the efficiency and flexibility we have included validity constraints in MPs that can be used to control the need for an online access to the originator server. As a result, originators can define certain validity constraints for each MP (based on number of accesses, time, etc. depending on the smart card features). Hence, MPs can be cached by clients and used directly while they are still valid. The generation of MPs is a fast process while the generation of PCOs is slower. Furthermore, PCOs are much more stable than policies. Finally, opposed to PCOs, each MP is specific for a smart card. As each PCO has its own key, we can manage them individually, which is not possible in other software protection proposals where all applications are protected using the same key.

When the client requests some data object from a DL server, it receives the PCO containing it. Before the PCO can execute the protected sections of its code it has to retrieve the corresponding MP sending a request containing the certificate of the public key of the client smart card. In case the server from where the PCO was retrieved is the originator of the PCO, it produces the MP for that PCO. Otherwise the server just forwards this request to the PCO originator. When the MP is received by the client smart card, it is decrypted, verified and stored inside the card until it expires or the user explicitly decides to extract it. Once the MP is correctly installed in the card the protected sections of the PCO can be executed, which requires the cooperation of the card containing the MP. The protected sections of the software do not reside in the cards. Instead, during the execution of the PCO, these sections are transmitted dynamically as necessary to the card, where they are decrypted using the installed MP and executed. When finished, the card may send back some results. Some other partial results will be kept in the card in order to obtain a better protection against function analysis and other attacks.

3.5 Global Structure

A general overview of the main components of the system and their relation is depicted in figure 4. The first component is the *SmartProt* protection system. This component transforms unprotected content objects in the originator DL server into PCOs as described in section 3.4.

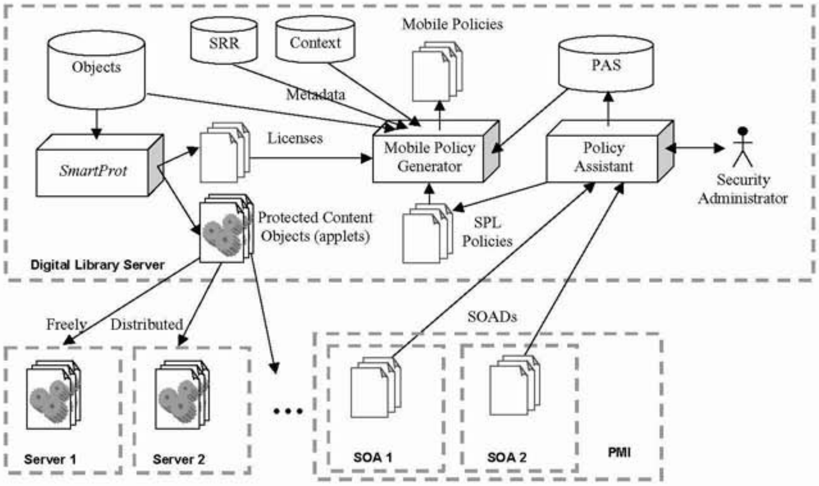


Fig. 4. XSCD-DL Infrastructure for Digital Libraries

The second component, called *Policy Assistant*, is designed to help security administrators in the specification, management and validation of access control policies. This component uses the SOADs as a basis for the specification of SPL policies and PAS. It is also responsible for the automated validation of policies at different levels. SPL policies are validated syntactically using XML-Schema. Semantic validation is made possible by the use of a specific *Semantic Policy Validator* (included in the Policy Assistant) that uses the DOM API to parse the SPL specification validating it. Finally, as an extension of the semantic validation, policies can also be validated in the context where they will be applied. Policy context validation uses the semantic information contained in the SOADs for the detection of possible inconsistencies in the SPL policies. Therefore, the Policy Assistant integrates all the tools to facilitate the administration of the access control system.

The third component, called *Mobile Policy Generator*, attends requests from end users producing MPs dynamically. To determine the set of applicable policies for a given PCO, the generator uses different sources of metadata. After receiving a request the Mobile Policy Generator analyses the semantic metadata available for the target PCO, which is contained in SRRs, finds the appropriate PAS and retrieves the necessary SOADs. Using this information, the Mobile Policy Generator is able to find the applicable SPL policies. These policies are then analysed and instantiated using the metadata about the resource (SRRs) and the context. Finally, these policies are combined. The combination of policies helps reducing the complexity of administration while enabling more expressive and flexible policies to be considered in the access decision.

4 Conclusions and Future Work

The system presented in this paper, XSCD-DL, combines an external PMI, a modular language (*Semantic Policy Language*, SPL) and a software protection mechanism (*SmartProt*) for the specification of the access control policies in order to provide distributed access control management and enforcement and secure content distribution in digital libraries. XSCD-DL allows policies to be dynamically changed by the owner or originator of the resource in a transparent manner.

An important feature is the extensive use of XML metadata technologies to facilitate the security administration in digital libraries and other complex environments. It also enables interesting functionalities of the system such as the contextual validation of policies. In our system, XML metadata technologies are applied at different levels to express the semantic information. On one hand, metadata is used for the creation and semantic and contextual validation of access control policies. Likewise, metadata about the objects included on the DL enables dynamic policy allocation and parameter instantiation. On the other hand, metadata is an essential tool for the integration of the external PMI in the access control system.

To summarize, XSCD-DL represents a solution applicable in different distributed scenarios, is flexible, solves the originator-retained-control problem, can be applied regardless of the attribute certification scheme, implements distributed access control management and enforcement mechanisms, incorporates secure content distribution and allows the dynamic modification of policies transparently and efficiently.

A prototype of this system has been implemented for a Digital Library scenario. In such an environment, PCOs are implemented using Java applets. The *e-gate Cyberflex*™ USB Java smart cards are used as secure coprocessors. The high capacity and the transfer speed of these cards helps ensure that the performance of the PCO is very good. A set of techniques, such as temporary authorizations, is used to improve the performance. We are currently working on the formalization of SPL specifications. A fair payment mechanism has been designed to complement the system and is currently being implemented.

References

1. López, J., Maña, A., Pimentel, E., Troya, J.M., Yagüe, M.I. *An Infrastructure for Secure Content Distribution*. To appear in Proceedings of ICICS'02. Springer-Verlag. 2002.
2. Janée, G., Frew, J. *The ADEPT Digital Library Architecture*. Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '02). 2002.
3. Coleman, A. *Metadata Rules the World: What Else Must We Know About Metadata?*. http://www.alexandria.ucsb.edu/~acoleman/mrworld.html

4. Baldonado, M., Chang, K., Gravano, L.P, Paepcke, A. *The Standford Digital Library Metadata Architecture*. International Journal of Digital Libraries, 1(2), February 1997.
5. Ketchpel, S., Garcia-Molina, H., Paepcke, A. *Shopping Models: A Flexible Architecture for Information Commerce*. Proceedings of the Second ACM International Conference on Digital Libraries. 1996.
6. Thompson, M., et al., *Certificate-based Access Control for Widely Distributed Resources*. Proceedings of the 8th USENIX Security Symposium. pp. 215-227. 1999.
7. Chadwick, D. W. *An X.509 Role-based Privilege Management Infrastructure*. Business Briefing. Global Infosecurity 2002. http://www.permis.org/
8. López, J., Maña, A., Yagüe, M.I. *XML-based Distributed Access Control System*. In Proceedings of EC-Web'02. Springer-Verlag, LNCS 2455. 2002.
9. Fayad, A., Jajodia, S. *Going Beyond MAC and DAC Using Mobile Policies*. In Proceedings of IFIP SEC'01. Kluwer Academic Publishers. 2001.
10. McCollum, C.J.; Messing, J.R.; Notargiacomo, L. *Beyond the pale of MAC and DAC - Defining new forms of access control*. Proceedings of the IEEE Symposium on Security and Privacy, pp. 190-200. 1990.
11. Yagüe, M. I. *On the suitability of existing access control and DRM languages for mobile policies.* University of Málaga. Department of Computer Science Technical Report nb. LCC-ITI-2002/10. 2002.
12. Bertino, E., Castano, S., Ferrari, E. *Securing XML documents with Author-X*. IEEE Internet Computing, 5(3):21-31, May/June 2001.
13. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarati, P. *Controlling Access to XML Documents.* In IEEE Internet Computing, vol. 5, n. 6, November/December 2001, pp. 18-28.
14. Damiani, E., De Capitani di Vimercati, S., Paraboschi, S., Samarati, P. *A Fine-Grained Access Control System for XML Documents.* In ACM Transactions on Information and System Security (TISSEC), vol. 5, n. 2, May 2002, pp. 169-202.
15. W3C. *XML-Schema*. http://www.w3.org/XML/Schema
16. W3C. *XML Path Language*. http://www.w3.org/TR/xpath
17. ITU-T Recommendation X.509. *Information Technology – Open systems interconnection – The Directory: Public-key and attribute certificate frameworks*. 2000.
18. W3C. *Resource Description Framework* (RDF). http://www.w3c.org/RDF/
19. W3C. *XML-Signature Syntax and Processing.* http://www.w3.org/TR/xmldsig-core/. 2002.
20. Chatvichienchai, S., Iwaihara, M., Kambayashi, Y. *Translating Access Authorizations for Transformed XML Documents.* In Proceedings of Ec-Web'02. Springer-Verlag, LNCS 2455. 2002.
21. Kudo, M., Hada, S. *XML Document Security based on Provisional Authorization*. In Proceedings of the 7th ACM Conference on Computer and Communications Security. 2000.
22. Maña, A., Pimentel, E. *An Efficient Software Protection Scheme*. In Proceedings of IFIP SEC'01. Kluwer Academic Publishers. 2001.

A Strategic Level for Scientific Digital Libraries

Ling Feng[1], Manfred Jeusfeld[2], and Jeroen Hoppenbrouwers[2]

[1] Dept. of Computer Science, University of Twente, PO Box 217
7500 AE Enschede, The Netherlands. `ling@cs.utwente.nl`*
[2] Infolab, Tilburg University, PO Box 90153
5000 LE Tilburg, The Netherlands.
`{jeusfeld,hoppie}@kub.nl`

Abstract. Digital libraries (DLs) are a resource for answering complex questions. Up to now, such systems mainly support keyword-based searching and browsing. The mapping from a research question to keywords and the assessment whether a retrieved article is relevant to the research question are completely the responsibility of the user. In this paper, we present a two-layered DL model. The aim is to enhance current DLs to support different levels of human cognitive acts, thus enabling new kinds of knowledge exchange among library users. The lower layer of the model, namely, the *tactical cognition support layer*, is intended to provide users with requested relevant documents, just as in searching and browsing. The upper layer of the model, namely, the *strategic cognition support layer*, not only provides users with relevant documents but also directly and intelligently answers users' cognitive questions. On the basis of the proposed model, we divide the DL information space into two subspaces, i.e., a *knowledge subspace* and a *document subspace*, where documents in the document subspace serve as the justification for the corresponding knowledge in the knowledge subspace. A detailed description of the knowledge subspace is discussed.

1 Introduction

Users' information retrieval activities have traditionally been categorized into *searching* and *browsing*. Searching implies that users know exactly what to look for, while browsing should assist users navigating among correlated searchable terms to look for something new or interesting. So far, most of the major work on DLs has focused on supporting these two kinds of information requirements. To support efficient searching activities, efforts have been made to develop retrieval models, build document and index spaces, extend and refine queries for DLs [9]. In [6], index terms are automatically extracted from documents and a vector - space paradigm is exploited to measure the matching degrees between queries and documents. Indexes and metadata can also be manually created from which semantic relationships are captured [2]. Furthermore, the information space consisting of a large collection of documents can be semantically partitioned into different clusters, so that queries can

* This work was completed while the author was in the Infolab of Tilburg University in the Netherlands.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 495-506, 2002.

be evaluated against relevant clusters [21]. According to topic areas, a distributed semantic framework is proposed in [17] to contextualize the entire collection of documents for efficient large-scale searching. To improve query recall and precision, several query expansion and refinement techniques based on relational lexicons/thesauri or relevance feedback have been explored [20].

Since one DL usually contains many distributed and heterogeneous repositories which may be autonomously managed by different organizations, in order to facilitate users' browsing activities across diverse sources easily, much effort has been expended in handling various structural and semantic variations and providing users with a coherent view of a massive amount of information. Nowadays, the interoperability problem has sparked vigorous discussion in the DL community. The concept extraction, mapping and switching techniques enable users in a certain area to easily search the specialized terminology of another area. A dynamic mediator infrastructure [13] allows mediators to be composed from a set of modules, each implementing a particular mediation function, such as protocol translation, query translation, or result merging [15]. [18] presents an extensible digital object and repository architecture FEDORA, which can support the aggregation of mixed distributed data into complex objects, and associate multiple content disseminations with these objects. [16] employs the distributed object technology to cope with interoperability among heterogeneous resources. With XML becoming the Web data exchange standard, considerable work on modeling, querying and managing semistructured data and non-standard data formats are conducted to enable the integration of heterogeneous resources [3,7].

Despite many fruitful achievements in the DL area, from the standpoint of satisfying a user's information needs, the current DL systems suffer from the following two shortcomings.

Inadequate High-Level Cognition Support. Traditional searching in DLs is keyword-based. Users ask for information by entering some keywords, and DL systems return matching documents. But users expect more than this. Typically, users have some pre-conceived hypotheses or domain-specific knowledge. They may desire the library to confirm/deny their existing hypotheses, or to check whether there is some exceptional/contradictory documented evidence against their pre-existing notions, or to provide some predictive information so that they can take effective action. For example, a user working in a flood-precaution office is concerned about whether there will be floods in the coming summer. According to his previous experience, it seems that "*A wet winter may cause floods in summer*". In this situation, instead of using *disperse keywords* to ask for *documents*, the user would prefer to pose a *direct question* to the DL like : "*Does a wet winter cause floods in summer*?", expecting a confirmed/denied *intelligent answer* as well as a series of *supporting documents* to justify the answer. He does not want a list of articles lacking explanatory semantics, which require further checking.

Inadequate Knowledge Sharing and Exchange Channel. Traditional libraries are public places where a great deal of mutual learning, knowledge sharing and exchange can happen. A user may ask a librarian for searching assistance. Librarians may collaborate in the process of managing, organizing and disseminating information. Users themselves may communicate and help each other to use library resources. When we progress from physical libraries to virtual DLs, these valuable features must be retained. Future DLs should not just be simple storage and archival systems. To be successful, DLs must become a *knowledge place* for a wide spread of knowledge acquisition, sharing and propagation. In the above example, if the DL could make readily available knowledge and expertise to users, who might otherwise have to carry out time-consuming searching and consultation with librarians and/or experts, we can improve users' working effectiveness and efficiency. Moreover, since machine knowledge does not deteriorate over time as human knowledge does, DLs provide an ideal repository for long-term retention of knowledge.

In this paper, we propose a two-layered DL model to support users' tactical and strategic level information requirements. The model moves beyond simple searching and browsing across multiple correlated repositories to acquisition of knowledge. On the basis of the proposed function model, we further divide the DL information space into two subspaces, i.e., a *knowledge subspace* and a *document subspace*. Documents in the document subspace serve as the justification for the corresponding knowledge in the knowledge subspace.

The remainder of the paper is organized as follows. In Section 2, we outline a two-layered DL model. A formal description of the DL knowledge subspace is presented in Section 3. Section 4 concludes the paper.

2 A Two-Layered DL Function Model

We propose a two-layered DL model, consisting of a *tactical cognition support layer* and a *strategic cognition support layer*, to address users' information needs, as shown in Figure 1.

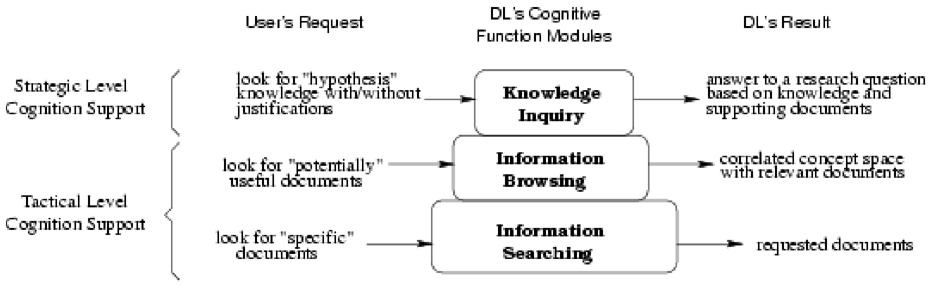

User's Request
DL's Cognitive Function Modules
DL's Result
Strategic Level Cognition Support
Tactical Level Cognition Support
look for "hypothesis" knowledge with/without justifications
Knowledge Inquiry
answer to a research question based on knowledge and supporting documents
look for "potentially" useful documents
Information Browsing
correlated concept space with relevant documents
look for "specific" documents
Information Searching
requested documents

Fig. 1. A two-layered DL model

2.1 Tactical Level vs. Strategic Level Cognition Support

Tactical level cognition support. We view traditional DL searching and browsing as tactical level cognitive acts. The target of searching is towards certain specific documents. One searching example is "*Look for the article written by John Brown in the proceedings of VLDB88.*" As the user's request can be precisely stated beforehand, identifying the target repository where the requested document is located is relatively easy. In comparison to searching whose objective is well-defined, browsing aims to provide users with a conceptual map, so that users can navigate among correlated items to hopefully find some potentially useful documents, or to formulate a more precise retrieval request. For instance, *a user reads an article talking about a water reservoir construction plan in a certain region. S/he wants to know the possible influence on ecological balance. By following semantic links for the water reservoir plan in the DL, s/he navigates to the related "ecological protection" theme, under which a set of searchable terms with relevant documents are listed for selection*. As the techniques of searching and browsing have been extensively studied and published in the literature, we will not discuss these any further here.

Strategic level cognition support. In contrast to tactical level cognition support which is intended to provide users with requested documents, strategic level cognition support not only provides relevant documents but also intelligently answers high-order cognitive questions, while providing justification and evidence. For instance, instead of retrieving documents with dispersed keywords like *wet winter* and *summer flood*, the user would prefer to pose a direct question like "*Tell me whether a wet winter will cause summer flood*", and receive a direct confirmed/denied answer from the DL system rather than a list of articles lacking explanatory semantics, which require his assessment. The provision of strategic level cognition support adds values to DLs beyond simply providing document access. It reinforces the exploration and utilization of information in DLs, and advocates a closer and more powerful interaction between users and DL systems.

2.2 An Enlarged DL Information Space

In order to support the two kinds of cognitive acts, we divide the DL information space into two subspaces, i.e., a *knowledge subspace* and a *document subspace*, as shown in Figure 2. Here, we illustrate only the organization of documents for knowledge justification purposes. Documents in a DL are in fact also indexed and clustered based on the ontology and thesauri.

The knowledge subspace. The basic constituent of the *knowledge subspace* is knowledge, such as hypotheses, rules, beliefs, etc. In this initial study, we focus on hypothesis knowledge in the empirical sciences. Each hypothesis describes a certain relationship among a set of concepts. For example, the hypothesis "H_2: *wet winter causes summer flooding*" explicates a causal relationship between a cause *wet winter* and the effect *summer flooding*. Considering that the DL knowledge subspace is for users to retrieve strategic level knowledge, it will inevitably be subject to classical information retrieval's vocabulary (synonymy and polysemy) problem. Previous research [1,5] demonstrated that different users tend to use different terms to look for identical information. To enable knowledge exchange and reusability, we build various relationships, including *equivalence, specification/generalization* and *opposition*, over the hypotheses knowledge, expanding a single user's hypothesis into a network of related hypotheses. Later, if one user's inquiry has the form of a hypothesis, the above relationships can be explored to find matching hypotheses in the knowledge subspace. The hypotheses, together with supporting documents serving as the justification of the hypotheses, are returned to the user as a part of the answer to his/her strategic request. For example, a more general hypothesis with respect to H_2 is "H_1: *wet winter is related with river behavior*".

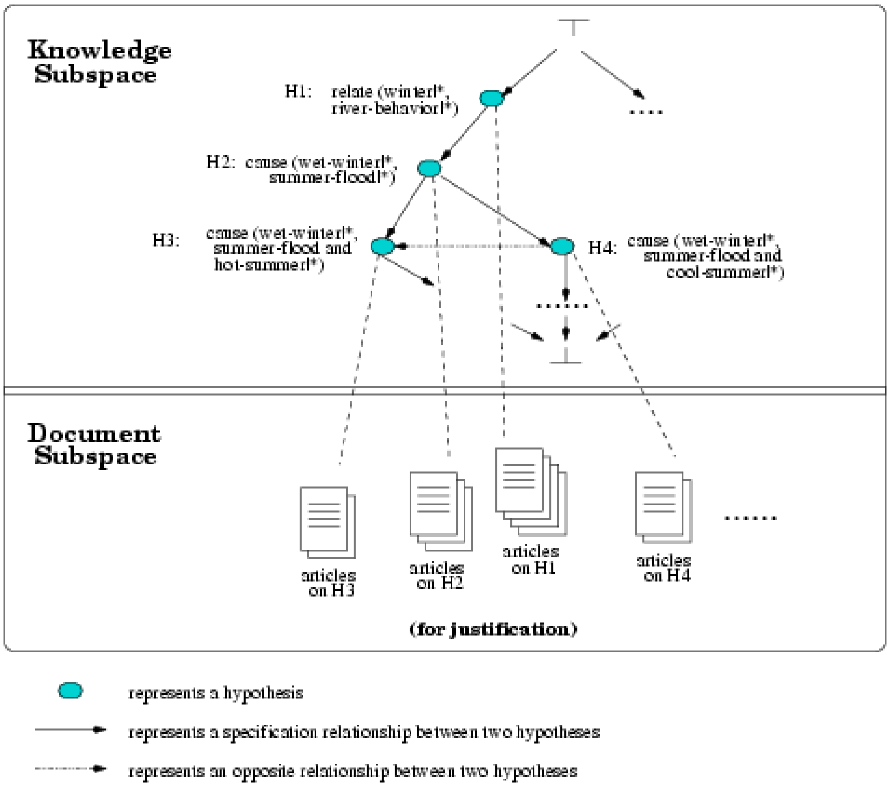

Knowledge Subspace
⊤
H1: relate (winter|*, river-behavior|*)
....
H2: cause (wet-winter|*, summer-flood|*)
H3: cause (wet-winter|*, summer-flood and hot-summer|*)
H4: cause (wet-winter|*, summer-flood and cool-summer|*)
......
⊥
Document Subspace
articles on H3
articles on H2
articles on H1
articles on H4
......
(for justification)
represents a hypothesis
represents a specification relationship between two hypotheses
represents an opposite relationship between two hypotheses

Fig. 2. An enlarged DL information space

The document subspace. Under each hypothesis is a justification set, giving reasons and evidence for the hypothesis. This justification, made up of articles, reports, etc., constitute the document subspace of the DL information space. Taking the above hypothesis H_2 as an example, articles saying exactly that "*wet winter is an indicator of summer flooding*" constitute the justification for that hypothesis. It is worth noting here that the document subspace challenges traditional DLs on literature organization, classification, and management. For belief justifications, we must extend the classical *keyword*-based index schema, which is mainly used for information searching and browsing purposes, to *knowledge*-based index schema, in order that the information in DLs can be easily retrieved by both keywords and knowledge.

3 A Formal Description of the DL Knowledge Subspace

In this section, we define the basic constituent of the DL knowledge subspace - *hypothesis*, starting with its two constructional elements, i.e., *concept terms* and *relation terms*. Throughout the discussion, we use the following notation :

- A finite set of entity concepts $EConcept=\{e_1, e_2, \ldots, e_t\}$.
- A finite set of relation concepts $RConcept=\{r_1, r_2, \ldots, r_s\}$.
- A finite set of concepts *Concept*, where $Concept = EConcept \cup RConcept$.
- A finite set of contextual attributes $Att=\{a_1, a_2, \ldots, a_m\}$. The domain of $a_i \in Att$ is denoted as $Dom(a_i)$.
- A finite set of concept terms $CTerm = \{n_1, n_2, \ldots, n_u\}$.
- A finite set of relation terms $RTerm = \{m_1, m_2, \ldots, m_v\}$.
- A finite set of hypotheses $Hypo = \{H_1, H_2, \ldots, H_w\}$.

3.1 Concepts

Concepts represent real-world entities, relations, states and events. We classify concepts into two categories, i.e., *entity concepts* and *relation concepts*. Entity concepts are used to describe concept terms, while relation concepts are mainly for presenting various relationships among concept terms. Based on the substantial work on lexicography and ontology [19,12,8,14,10,11], three typical primitive relationships (*Is-A, Synonym*, and *Opposite*) between concepts can be established.

3.2 Concept Terms

Each entity concept can be associated with a conceptual context, denoting the circumstance under which the entity concept is considered. Basically, a conceptual context can be described using a set of attributes called **contextual attributes**, each of which represents a dimension in the real world. Typical contextual attributes include *time, space, temperature*, and so on. For example, the context for the entity concepts *wet-winter* and *summer-flood* could be constructed by two contextual attributes, *Year* and *Region*. Such a context can be further instantiated by assigning concrete values to its constructional attributes. For example, we can restrict the contextual *Region* and *Year* of *wet-winter* to the *north* and *south* areas in *2000* by setting *Region:=*{*"north", "south"*} and *Year:=*{*"2000"*}. *Region:=Dom*(*Region*) assigns all applicable regions ("*south*", "*north*", "*east*", and "*west*") to contextual attribute *Region*. In this paper, we name an entity concept with an associated context as a **concept term**.

Definition 1. A **concept term** n is of the form $n = e|_{AV}$, where $e \in EConcept$ and $AV= \{a:= V_a \mid (a \in Att) \wedge (V_a \subseteq Dom(a))\}$.
Let $Att = \{a_1, a_2, \ldots, a_m\}$ be a set of contextual attributes which constitute the context under consideration. The default setting for attribute $a_i \in Att$ is the whole set of applicable values in the domain of a_i, i.e., $a_i{:=}Dom(a_i)$. $AV=\{a_1{:=}Dom(a_1), a_2{:=}Dom(a_2), \ldots, a_m{:=}Dom(a_m)\}$ depicts a universe context. A simple and equivalent representation of the universe context is $AV = *$.

Example 1. Suppose the context is comprised of two contextual attributes *Region* and *Year*. $wet\text{-}winter|_{\{Region:=\{\text{"north"}, \text{"south"}\}, Year:=\{\text{"2000"}\}\}}$ is a concept term, denoting a *wet-winter* entity concept in the *north* or *south* in *2000*.
$wet\text{-}winter|_{\{Region:=Dom(Region), Year:=Dom(Year)\}}$ is equivalent to $wet\text{-}winter|_{*}$ according to Definition 1.

Different relationships of concept terms can be identified based on their entity concepts and associated contexts. Before giving the formal definitions, we first define three relationships between two instantiated contexts.

Definition 2. Let AV_1 and AV_2 be two instantiated contexts.

- $AV_1 \leq_a AV_2$ (or $AV_2 \geq_a AV_1$), iff $\forall(a{:=}\, V_2) \in AV_2\, \exists(a{:=}\, V_1) \in AV_1 \;\; (V_1 \subseteq V_2)$.
- $AV_1 =_a AV_2$, iff $(AV_1 \leq_a AV_2) \wedge (AV_2 \leq_a AV_1)$.
- $AV_1 <_a AV_2$ (or $AV_2 >_a AV_1$), iff $(AV_1 \leq_a AV_2) \wedge (AV_1 \neq_a AV_2)$.

According to Definition 1, for any instantiated context AV, $AV \leq_a *$.

The $=_a$ relationship between two instantiated contexts AV_1 and AV_2 indicates that they both have exactly the same contextual attributes with the same attribute values. $AV_1 \leq_a AV_2$ states that AV_2 is broader than AV_1, covering the contextual scope of AV_1.

Example 2. Assume we have four instantiated contexts:
$AV_1 = *$, $AV_2 = \{Year{:=}\{\text{“}1999\text{”}, \text{“}2000\text{”}, \text{“}2001\text{”}\}\}$,
$AV_3 = \{Region{:=}\{\text{“}north\text{”}\}, Year{:=}\{\text{“}2000\text{”}\}\}$,and
$AV_4 = \{Region{:=}\{\text{“}south\text{”}\}, Year{:=}\{\text{“}2000\text{”}\}\}$.
Following Definition 2, we have $AV_2 <_a AV_1$, $AV_3 <_a AV_1$, $AV_4 <_a AV_1$, $AV_3 <_a AV_2$, and $AV_4 <_a AV_2$.

Based on the primitive relationships of concepts (*Is-A, Synonym, Opposite*), as well as their contexts ($\leq_a$,$=_a$,$<_a$), we can formulate the following three concept-term-based relationships, i.e., equivalence EQ_{ct}, specification $SPEC_{ct}$ and opposition $OPSI_{ct}$. Assume $n_1 = e_1|_{AV1}$ and $n_2 = e_2|_{AV2}$ are two concept terms in the following definitions, where $n_1, n_2 \in CTerm$.

Definition 3. **Equivalence** EQ_{ct} (n_1, n_2). n_1 is **equivalent** to n_2, iff the following two conditions hold: 1) $(e_1 = e_2) \vee$ Synonym (e_1, e_2); 2) $(AV_1 =_a AV_2)$.

Example 3. Given two concept terms: n_1 = $\textit{wet-winter}|_{\{Region:=\{\text{“}north\text{”}\}\}}$ and $n_2 = \textit{high-rainfall-winter}|_{\{Region:=\{\text{“}north\text{”}\}\}}$, EQ_{ct} (n_1, n_2) since Synonym (*wet-winter, high-rainfall-winter*).

Definition 4. **Specification** $SPEC_{ct}$ (n_1, n_2). n_1 is a **specification** of n_2 (conversely, n_2 is a **generalization** of n_1), iff the following two conditions hold:
1) $(e_1 = e_2) \vee$ Is-A $(e_1, e_2) \vee$ Synonym (e_1, e_2); 2) $(AV_1 \leq_a AV_2)$.

Example 4. Let $n_1 = \textit{wet-winter}|_{\{Region:=\{\text{“}north\text{”}\}, Year:=\{\text{“}2000\text{”}\}\}}$,
$n_2 = \textit{wet-winter}|_{\{Year:=\{\text{“}2000\text{”}\}\}}$ be two concept terms. $SPEC_{ct}$ (n_1, n_2) since
$\{Region{:=}\{\text{“}north\text{”}\}, Year{:=}\{\text{“}2000\text{”}\}\} <_a \{Year{:=}\{\text{“}2000\text{”}\}\}$.

Definition 5. **Opposition** $OPSI_{ct}$ (n_1, n_2). n_1 is **opposite** to n_2, iff the following two conditions hold : 1) Opposite (e_1, e_2); 2) $(AV_1 =_a AV_2)$.

Example 5. Let $n_1 = \textit{wet-winter}|_{\{Region:=\{\text{“}north\text{”}\}\}}$, $n_2 = \textit{dry-winter}|_{\{Region:=\{\text{“}north\text{”}\}\}}$ be two concept terms. $OPSI_{ct}$ (n_1, n_2) since Opposite(*wet-winter, dry-winter*).

To facilitate the description of hypothesis-based inter-relationships in Subsection 3.4, we further extend the three relationships (i.e., EQ_{ct}, $SPEC_{ct}$ and $OPSI_{ct}$) defined over a pair of concept terms to the ones (i.e., EQ_{CT}, $SPEC_{CT}$ and $OPSI_{CT}$) over a pair of concept term sets. Let N_1 and N_2 be two concept term sets.

Definition 6. **Equivalence** EQ_{CT} (N_1, N_2). N_1 is **equivalent** to N_2, iff $\forall n_1 \in N_1 \exists n_2 \in N_2$ EQ_{ct} $(n_1, n_2) \wedge \forall n_2 \in N_2 \exists n_1 \in N_1$ EQ_{ct} (n_2, n_1).

Definition 7. **Specification** $SPEC_{CT}(N_1,N_2)$. N_1 is a **specification** of N_2, iff $\forall n_2 \in N_2 \exists n_1 \in N_1$ $SPEC_{ct}$ (n_2, n_1).

Definition 8. **Opposition** $OPSI_{CT}$ (N_1,N_2). N_1 is **opposite** to N_2, iff $\exists n_1 \in N_1$ $\exists n_2 \in N_2$ $OPSI_{ct}(n_1,n_2) \vee \exists n_2 \in N_2$ $\exists n_1 \in N_1$ $OPSI_{ct}(n_2,n_1)$.

As long as there exists a pair of opposite concept terms in the two concept term sets, we declare they are opposite.

3.3 Relation Terms

A relation concept explicates a certain correlation among a set of conceptual terms. Unlike entity concepts, relation concepts can be affiliated with different kinds of modals like *necessity, possibility, permission*, etc. to qualify the truth of the relationships. In this paper, we apply well-established modal logic [4] to our relation concept study. By prefixing a relation concept *r* with the symbol □ or ◊, we can achieve different levels of a certain ability regarding relation *r*. For example, □ *cause* implies a *necessarily causal* relation, while ◊ *cause* implies a *possibly causal* relation.

Definition 9. A **relation term** *m* is of the form $m = \delta r$, where $r \in RConcept$, and δ could be □, ◊, or an empty modal.

Definition 10. According to modal logic, we define the order " □" < " " < "◊" for symbols □, empty modal, and ◊.

The three relation-term-based relationships can be defined using the same names (i.e., *EQ, SPEC* and *OPSI*) as concept-term-based relationships but with a different subscript flag "$_{rt}$" to make the difference.

Definition 11. **Equivalence** EQ_{rt} $(\delta_1 r_1, \delta_2 r_2)$. $\delta_1 r_1$ is **equivalent** to $\delta_2 r_2$, iff the following two conditions hold: 1) $(r_1 = r_2) \vee$ Synonym (r_1, r_2); 2) $(\delta_1 = \delta_2)$.

Definition 12. **Specification** $SPEC_{rt}$ $(\delta_1 r_1, \delta_2 r_2)$. $\delta_1 r_1$ is a **specification** of $\delta_2 r_2$ (conversely, $\delta_2 r_2$ is a **generalization** of $\delta_1 r_1$), iff the following two conditions hold:
1) $(r_1 = r_2) \vee$ Is-A $(r_1, r_2) \vee$ Synonym (r_1, r_2); 2) $(\delta_1 = \delta_2) \vee (\delta_1 < \delta_2)$.

Definition 13. **Opposition** $OPSI_{rt}$ $(\delta_1 r_1, \delta_2 r_2)$. $\delta_1 r_1$ is **opposite** to $\delta_2 r_2$, iff the following two conditions hold: 1) Opposite (r_1, r_2); 2) ($\delta_1 = \delta_2 \neq$ "◊").

Example 6. EQ_{rt} ($\Diamond cause$, $\Diamond lead\text{-}to$), $SPEC_{rt}$ ($\Box cause$, $\Diamond cause$), $SPEC_{rt}$ ($cause$, $relate$), and $OPSI_{rt}$ ($\Box relate$, $\Box unrelate$).

3.4 Hypotheses

A hypothesis communicates a user's cognitive idea or thinking about things in existence, such as the causal connection of situations, the sequential occurrence of events, etc. Here, we describe each piece of hypothesis using a relation concept which correlates a set of input concept terms with a set of output concept terms. For example, the hypothesis "*wet winter in the north causes summer flooding in the south and hot summer in the east*" causally relates concept term $wet\text{-}winter|_{\{Region:=\{\text{“north”}\}\}}$ to $summer\text{-}flood|_{\{Region:=\{\text{“south”}\}\}}$ and $hot\text{-}summer|_{\{Region:=\{\text{“east”}\}\}}$.

Definition 14. A hypothesis H is of the form $H=\delta r(I_N, O_N)$, where δr is a relation term, I_N and O_N are concept term sets.

Various hypothesis-based inter-relationships can be established based on the relationships of their components, i.e., concept term sets and relation terms. Assume that $H_1=\delta_1 r_1$ (I_{N1}, O_{N1}) and $H_2=\delta_2 r_2$ (I_{N2}, O_{N2}) are two hypotheses.

Definition 15. H_1 is **equivalent** to H_2, written as $H_1 =_h H_2$, iff EQ_{rt} $(\delta_1 r_1, \delta_2 r_2) \wedge EQ_{CT}$ $(I_{N1}, I_{N2}) \wedge EQ_{CT}$ (O_{N1}, O_{N2}).

For two equivalent hypotheses, they must have equivalent relation terms, as well as equivalent input and output concept term sets.

Definition 16. H_1 is a **specification** of H_2 (conversely, H_2 is a **generalization** of H_1), written as $H_1 \leq_h H_2$ $(H_2 \geq_h H_1)$, iff $SPEC_{rt}$ $(\delta_1 r_1, \delta_2 r_2) \wedge SPEC_{CT}$ $(I_{N1}, I_{N2}) \wedge SPEC_{CT}$ (O_{N1}, O_{N2}).
We call H_1 a **strict specification** of H_2 (conversely, H_2 a **strict generalization** of H_1), written as $H_1 <_h H_2$ $(H_2 >_h H_1)$, iff $(H_1 \leq_h H_2) \wedge (H_1 \neq_h H_2)$.
If H_1 is a specification of H_2 and a specification of H_3, then H_1 is a **common specification** of H_2 and H_3. Conversely, if H_1 is a generalization of H_2 and a generalization of H_3, then H_1 is a **common generalization** of H_2 and H_3.

Example 7. Given the following three hypotheses:
$H_1 = \Box cause$ $(\{wet\text{-}winter|_{\{Region:=\{\text{“north”}\}\}}, warm\text{-}winter|_{\{Region:=\{\text{“north”}\}\}}\}, \{summer\text{-}flood|_{\{Region:=\{\text{“south”}\}\}}\})$,
$H_2 = \Diamond cause$ $(\{wet\text{-}winter|_{\{Region:=\{\text{“north”}\}\}}\}, \{summer\text{-}flood|_{\{Region:=\{\text{“south”}\}\}}\})$,
$H_3 = \Diamond cause$ $(\{wet\text{-}winter|_{\{Region:=\{\text{“north”}\}\}}\}, \{river\text{-}behavior|_{*}\})$,
H_1 is more specific than H_2 and H_3, and H_2 is also more specific than H_3 (i.e., $H_1 \leq_h H_2$, $H_1 \leq_h H_3$ and $H_2 \leq_h H_3$). All of them are strict specifications. Besides, H_1 is a common specification of H_2 and H_3, and H_3 is a common generalization of H_1 and H_2.

Definition 17. H_1 is **opposite to** H_2, written as $H_1 \propto_h H_2$, iff either of the following conditions holds:

1) $OPSI_{rt}\ (\delta_1 r_1,\ \delta_2 r_2) \wedge EQ_{CT}\ (I_{N1},\ I_{N2}) \wedge EQ_{CT}\ (O_{N1},\ O_{N2})$; or
2) $EQ_{rt}\ (\delta_1 r_1,\ \delta_2 r_2) \wedge EQ_{CT}\ (I_{N1},\ I_{N2}) \wedge OPSI_{CT}\ (O_{N1},\ O_{N2})$.

For two opposite hypotheses, they may have equivalent input/output concept term sets but with opposite relation terms (Case 1 of the definition), or they may have equivalent relation terms and input concept term sets, but with at least one opposite output concept term pair (Case 2 of the definition).

Example 8. Given the following two hypotheses:
H_1 = □*relate*({*wet-winter*$|_*$}, {*summer-flood*$|_*$, *hot-summer*$|_*$}),
H_2 = □*unrelate*({*wet-winter*$|_*$}, {*summer-flood*$|_*$, *hot-summer*$|_*$}),
$H_1 \propto_{\mathrm{h}} H_2$ since $OPSI_{rt}$ (□*relate,* □*unrelate*).

3.5 The Knowledge Subspace and Its Linkage to the Document Subspace

Hypotheses and their inter-relationships constitute a DL knowledge subspace. At an abstract level, a knowledge subspace can be viewed as an oriented diagram consisting of a series of nodes (each representing a hypothesis) that are connected to each other through directed labeled edges (representing various relationships between hypotheses), as shown in the upper part of Figure 2. To make the diagram connected, we introduce two special hypotheses: the universal hypothesis ~ that is a generalization of all other hypotheses, and the absurd hypothesis ⊥ that is a specification of all other hypotheses.

Definition 18. A DL **knowledge subspace** is composed of a set of nodes representing hypotheses, and a set of directed edges representing relationships of hypotheses.

The DL document subspace accommodates all the documents in the DL. They are the sources for answers to users' information searching and browsing requests. In addition, for the enhanced DL system proposed in this paper, documents in the DL document subspace also provide justification for the corresponding knowledge in the knowledge subspace - that is, under each hypothesis is a set of justification documents, giving reasons and evidence for the hypothesis. Let *Doc* denote the whole set of documents in a DL. All the documents in the DL that support a hypothesis constitute the referent for that hypothesis.

Definition 19. Let *H* be a hypothesis in the knowledge subspace. The **referent** of *H*, written as φ*H*, is the set of documents in the DL that support *H*. For the absurd and universal hypotheses (⊥ and ~), we assume that $\varphi\perp = \varnothing$ and $\varphi\sim\ = Doc$.

The defined equivalence, (strict) specification and opposition relationships of hypotheses lead us to the following axiom and theorem.

Axiom 1. Let H_1, H_2 be two hypotheses.

- If H_1 is a (strict) specification of H_2, i.e., $H_1 \leq_{\mathrm{h}} (<_{\mathrm{h}})\ H_2$, then $\varphi H_1 \subseteq \varphi H_2$.
- If H_1 is equivalent to H_2, i.e., $H_1 =_{\mathrm{h}} H_2$, then $\varphi H_1 = \varphi H_2$.
- If H_1 is opposite to H_2, i.e., $H_1 \propto_{\mathrm{h}} H_2$, then $\varphi H_1 \cap \varphi H_2 = \varnothing$.

Using Definition 16 of specialization/generality between hypotheses, we can be sure that if a hypothesis is consistent with a set of documents, any generalization of it will also be consistent with this document set. In contrast, if a document does not conform to a hypothesis, it cannot conform to any specialization of that hypothesis either. For any hypothesis $H \in$ *Hypo* where $(\perp \leq_h H \leq_h \sim)$, it is obvious that $\emptyset = \varphi\perp \subseteq \varphi H \subseteq \varphi\sim = Doc$.

Theorem 1. Let H_1, H_2, H_3 be three hypotheses, where H_1 is a common generalization of H_2 and H_3, i.e., $(H_2 \leq_h H_1)$ and $(H_3 \leq_h H_1)$. We have $(\varphi H_2 \cup \varphi H_3) \subseteq \varphi H_1$.

Proof. Since $(H_2 \leq_h H_1)$, according to Axiom 1, $(\varphi H_2 \subseteq \varphi H_1)$. Similarly, $(\varphi H_3 \subseteq \varphi H_1)$ because of $(H_3 \leq_h H_1)$. Thus, $(\varphi H_2 \cup \varphi H_3) \subseteq \varphi H_1$.

4 Conclusion

In this paper, we presentes a two-layered DL model to address users' different information requirements. On the basis of the proposed model, we divide the DL information space into a knowledge subspace and a document subspace. A detailed description of the knowledge subspace and its construction mechanisms, as well as query facilities against the enhanced DL, are discussed. Currently, we are researching practical methods of knowledge acquisition to fill in the knowledge subspace.

References

1. M.J.Bates. Subject access in online catalogs: A design model. Journal of the American Society for Information Science, 37(6):357-376, 1986.
2. K.Beard and V.Sharman. Multidimensional ranking in digital spatial libraries. Proc. of the 2nd Intl. Conf. On the Theory and Practice of Digital Libraries, June 1995.
3. P.Buneman, A.Deutsch, and W.C.Tan. A deterministic model for semi-structured data. Proc. of the 1999 Intl. Workshop on Query Processing for Semi-Structured Data and Non-Standard Data Formats, January 1999.
4. B.F.Chellas. Modal Logic:An Introduction. Cambridge University Press, 1980.
5. H.Chen, K.J.Lynch, K.Basu, and T.Ng. Generating, integrating, and activating thesauri for concept-based document retrieval. IEEE Expert, 8(2):25-34, 1993.
6. M.Dunlop and C. van Rijsbergen. Hypermedia and free text retrieval. Journal of Information Processing and Management, 29(3), 1993.
7. C.E.Dyreson, M.H.Bohlen, and C.S.Jensen. Capturing and querying multiple aspects of semistructured data. Proc. of the 1999 Intl. Conf. VLDB, pages 290-301, 1999.
8. M.W.Evens and R.N.Smith. A lexicon for a computer question-answering system. American Journal of Computational Linguistics, 83:1-93,1979.
9. W.B.Frakes and R.Baeza-Yates. Information Retrieval: Data Structures and Algorithms, Prentice Hall, 1992.
10. T.Gruber. A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993.
11. N.Guarino and C.Welty. Ontological analysis of taxonomic relationships. Proc. of the 19th Intl. Conf. Conceptual Modeling, 2000.

12. F.Kiefer editor. Studies in Syntax and Semantics, chapter Semantics and Lexicography: Towards a New Type of Anilingual Dictionary, 1969.
13. S.Melnik, H.Garcia-Molina, and A.Paepcke. A mediation infrastructure for digital library services. Technical report, Stanford University, 2000.
14. J.T.Nutter, E.A.Fox, and M.W.Evens. Building a lexicon from machine-readable dictionaries for improved information retrieval. Journal of Literary and Linguistic Computing, 5(2):129-137,1990.
15. A.Paepcke, R.Brandriff, G.Janee, R.Larson, B.Ludaescher, S.Melnik, and S.Raghavan. Search middleware and the simple digital library interoperability protocol. D-Lib Magazine, 6(3), 2000.
16. A.Paepcke, s.b.cousins, H.Garcia-Molina, S.W.Hassan, s.k.Ketchpel, M.Roscheisen, and T.Winograd. Using distributed objects for digital library interoperability. Technical report, Stanford University, 1998.
17. M.Papazoglou and J.Hoppenbrouwers. Contextualizing the information space in federated digital libraries. SIGMOD Record, 28(1):40-46, 1999.
18. S.Payette and C.Lagoze. Flexible and extensible digital object and repository architecture (FEDORA). Proc. of the 2nd European Conf. On Research and Advanced Technology for Digital Libraries, pages 41-59, 1999.
19. J.F.Sowa. Conceptual Structures: Information Processing in Mind and Machine. Addison-Wesley, 1984.
20. B.Velez, R.Weiss, M.A. Sheldon, D.K.Gifford. Fast and effective query refinement. Proc. of the 20th Intl. ACM SIGIR Conf. On Research and Development in Information Retrieval, pages 6-15, 1997.
21. P.Willett. Recent trends in hierarchical document clustering: a critical review. Information Processing and management, 24(5), 1998.

Bridging the Gap between Information Resource Design and Enterprise Content Management

Sue Fowell

Faculty of Information Technology
University of Technology, Sydney
City campus, PO Box 123, Broadway,
NSW 2007, Australia
sfowell@it.uts.edu.au

Abstract. The enterprise information landscape is now more complex than ever and includes a diverse range of internally created information resources and services such as: intranets, extranets, digital libraries, content, knowledge, and document management systems. Alongside information resources created internally, there are also those obtained from external content providers, business partners, suppliers and customers, as well as information products generated by the organisation and sold to external customers. The information environment is continually evolving in response to changes in business requirements, user needs and the affordances of technology. The increasingly complex, evolutionary nature of the enterprise information environment and the perspectives of different stakeholder groups present a number of challenges for the designers and managers of information resources and services. In this paper, the impact of these challenges for the design and management of useful and usable information resources is explored and new agendas for both research and practice identified.

1 Introduction

The enterprise information environment is a dynamic system, continually evolving through the addition of resources and services in response to changes in business goals, work practices and technologies. Most organisations support a range of internal information resources and services, including a digital library, intranets, extranets, document management systems, e-learning systems and data warehouses, to name but a few. Together with the information created internally, there are also information resources obtained from external content providers, business partners, suppliers and customers. The diversity in the nature of the different information resources and services and their location, management and coordination within the complex and changing information landscape present a number of challenges to the designers and managers of enterprise information services. It is not yet clear to what extent traditional methods and methodologies for information resource design and service provision are appropriate in this dynamic and changing environment.

As well as the diversity in resources and services we now find an increasingly diverse and distributed user population. In many organisations there is now greater

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 507-515, 2002.

emphasis on end-users creating, managing and sharing information. Whilst this is a desirable situation it also leads to information management issues such as duplication of effort, excessive redundancy of information, integrity and provenance problems and confusion over information management and coordination responsibilities. Whilst there is the desire for greater interoperability between information systems and the sharing of information between workgroups[1], the distributed and devolved creation and management of information raise the potential for fragmentation of information resources and inefficiency.

2 Information Design and Management Activity in Organisations

Organisations are constituted through interrelated networks of communities of work or communities of practice [3,4,5]. These are groups of people bound together by shared expertise or activity. They may be co-located in a work team or drawn together through their expertise and interests in a professional community, special interest group, etc. Information design and information management activity are distributed across many different communities. For some communities of work such as information professionals, it is the primary focus of their day-to-day activity. For others, it is a necessary part of the work they do but not their primary focus. For example, business managers are often involved in maintaining sections of a corporate intranet and managing the publication of their department's reports alongside their core activities managing a department or project team. In order to understand how information design and management activities occur in organisations, we need to understand how the activities of the various communities of work (both formal and informal) interleave to form an information design and management infrastructure. That is, to understand how information design and management work are actually achieved in practice.

2.1 The Work of Information Professionals

Within most organisations there are specialists whose main role is the management of corporate information resources, such as librarians, records managers, publication managers and other information professionals. Changes in the enterprise information environment have led to changes in the work of these information professionals within organisations [6,7,8,9,10,11,12,13,14]. The recent literature on the role of the information professional in the changing enterprise information environment emphasises a number of areas where existing knowledge is being enhanced and new skills developed. These areas can be summarised into three levels of activity:

[1] "Optimising enterprise-wide IS services" was the top issue reported by IS Managers in the CSC Critical Issues for IS Management Surveys in both 2000 and 2001. "Optimising organisational effectiveness" was ranked as the second most important issue in 2001, up from fourth place in 2000 [1,2].

information resource design, information service provision and *enterprise content management*. Table 1 contains examples of activities included at each level.

Table 1. Examples of information resource design and management activity

Information resource design
Developing information storage and retrieval systems
Digitisation of paper based information assets and OCR
Document markup - selecting and applying document markup languages
Metadata - selecting, developing and applying of appropriate schemas and tools
Information labelling - classifying, cataloguing, indexing
Information architecture - navigation/search design
Content delivery
Information service provision
Understanding information use and user needs analysis
Enquiry and reference services work
Establishing and maintaining end user relationships
Providing training and professional development
Evaluating and improving performance of existing information services
Usability and user interface development
Enterprise content management
Buying information content for the enterprise
Identifying opportunities to create and sell information products
Developing and managing overall content solutions
Liaising with business units and IT department to coordinate system development
Negotiation of contracts with content providers
Content evaluation, selection and acquisition

Information resource design is concerned with the design and management of information artefacts. Attention is focused on the design of scalable, standards-based resources and reliable and efficient systems for their storage, location and retrieval.

Information service provision covers the role of the traditional information intermediary. The focus here is on information in use and the needs of information users. That is, ensuring that information resources and services are accessible and usable and support the day-to-day work practices of users (both internal and external).

At the level of enterprise content management the focus is on coordination of information resources and services and understanding their contribution to business efficiency and effectiveness.

The three levels of information resource design and management activity represent three different communities of practice. Whilst these three communities of practice have a declared interest in the design of usable information resources and services, there are major differences between the intent, discourse and activity undertaken by each community (Fig. 1). At the information resource design level research and development activity tends to focus on designing *usable* information artefacts and systems for information storage, discovery and retrieval [15,16,17,18]. The discourse around technical design issues is primarily that of looking inward; taking a structural and infrastructural view of the information resource.

At the information service design level the focus is on the design and management of *useful* information resources and services. Research at this level focuses on the user-experience and the design of resources and services that support the day-to-day work practices of users [19,20,21,22]. Information service design is initially inward looking supporting the organisation's internal business processes and users. However, through the development of internal resources opportunities for new information products are identified. These can then be made available as goods or services to external customers giving an outward looking view. Information service design tends to take both a business and a technology view, matching business and users information needs with technological solutions.

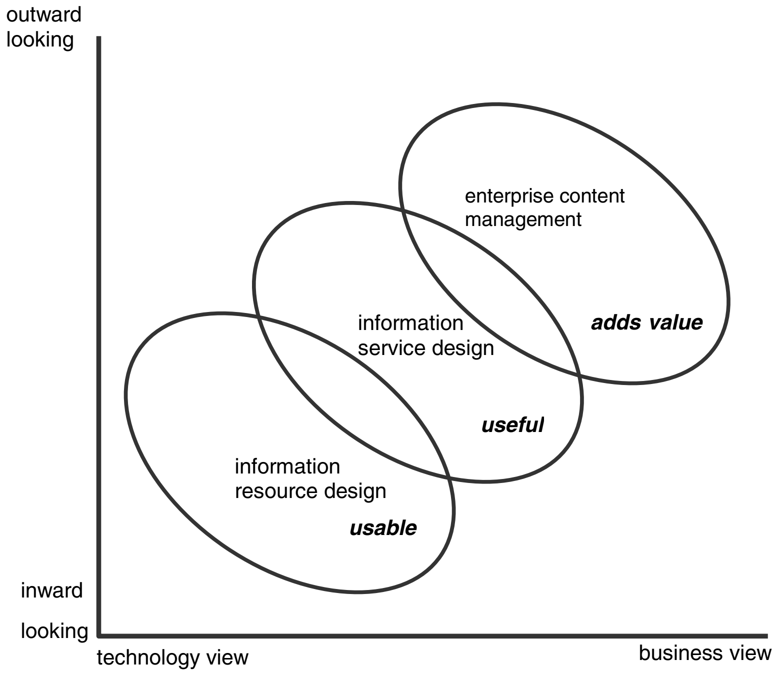

outward looking
enterprise content management
adds value
information service design
useful
information resource design
usable
inward looking
technology view
business view

Fig. 1. Three levels of information resource and service design

At the enterprise content management level attention is focused on information resources and services in the context of the extended enterprise. This group focus on the coordination of systems and services and on their contribution to organisational value in terms of both improved business efficiency and adding business value [23,24,25,26]. Business value can be achieved by using information resources and services to improve business processes or by identifying new revenue generating information products and services. The view at this level is mainly strategic with the aim of seeking to improve the organisation's competitive strengths and advantages.

A major challenge for organisations is to bridge the gap between the three communities of practice in order to create an efficient and usable information environment. That is, to carry out what could be called *deep information design*. That is, to adopt a more holistic approach to information design that brings together these different dimensions. That is, to integrates the three design views of useful, useable and adds value and that also take into account the inward/outward looking dimension and the business/technology dimension. There is currently little theoretical or practitioner literature available to guide this holistic approach to designing the enterprise information landscape.

2.2 Other Communities of Work

In the previous section the work of the information professional is considered and three interrelated areas of information work are identified. Those concerned with information resource design at a structural/infrastructural level, those concerned with the delivery of information services to users, and those who take an enterprise-level view of information and content management. Overlapping with these three communities of information practice are other communities of interest such as end users of all kinds, general business managers and strategic managers.

Through developments in information technology and changes in organisational structure, end users are now able (or required) to create, publish and manage their own information resources. For instance, to create digital documents such as forms and departmental reports and make them available on the organisational intranet, or to archive information artefacts such as email messages, business and project reports to the organisation's content management system.

General business managers are expected to have *information responsibility* [27]. Marchand notes that "[w]hile general managers may not be expected to be information specialists in their business, they are expected to create the conditions in a business for effective information use" [24:p5]. This includes not only understanding what information is being used and how it contributes to achieving business results, but also how such information is to be sourced and organised. Michael Earl identifies this *information literacy* of the workforce as being particularly important as one of the critical success factors for organisations participating in e-business [28].

The emergence of business opportunities from the development of new information goods and information based intermediary services has raised the profile of information resource design and management for strategic managers. The discourse in this community is around the concepts of information assets and information goods and their value [29,30,31] and the virtual value chain and value nets [32,33].

Thus, within an organisation there may be numerous communities of work with a direct or indirect interest in information resource design and management. The activities and intent of each of these communities varies. Some communities focus at the micro level around the day-to-day information resource creation and use and others focus at the macro level on the coordination of enterprise information resources and services and the gaining of competitive advantage. The means by which the activities of these various communities of work are interleaved to form an information design and management infrastructure is not well understood.

3 Bridging the Gap

The previous sections have identified a number of changes associated with information resource design and management in organisations. Firstly, the enterprise information environment is continually evolving in response to changes in business requirements and the affordances of technology. This leads us to consider if traditional forms of analysing organisational information use, and expressing information policy meet the current needs of organisations? For instance, are traditional information audit methods [25,34] well suited to rapidly evolving information landscapes?

This is compounded by the extended nature of the information environment. We find increasing levels of information intermediation and disintermediation. With digital information being sourced from external providers and internal information resources being developed as goods and services to external customers [6]. This raises interesting questions for the planning and coordination of information management across the organisation. Where, when, why and how are intermediation and disintermediation occurring in the information lifecycle; and what impact does this have on the provision of information services? How adequate are current conceptions of information lifecycle models when mapped onto virtual value chains?

In addition to the changing information environment, the location and nature of information work is changing and becoming even more distributed. This becomes problematic when it leads to information management issues such as duplication of effort, excessive redundancy of information, integrity and provenance problems and confusion over information management responsibilities.

Interoperability becomes more important; yet the distributed nature of information work makes the setting and implementing of information standards to ensure interoperability more difficult. End users and business managers are expected to be information literate but how far does information literacy extend? Are they required to become experts in document structuring and the various metadata languages and schema associated with their domain of expertise? If not, how are standards implemented to assure interoperability of information resources within and beyond the organisation? In order to develop information design and management strategies appropriate to the new information landscape and to achieve deep information design we need to understand how information work is actually achieved in organisations. That means making information work visible.

These challenges point to a new role for the information intermediary (and others) around the coordination of information resource and service development; the dissemination of standards; and the provision of education and training to communities of work where enhanced information literacy skills are seen as mandatory. Information intermediaries translate between the business focused and technology focused communities of work, bridging the gap between information resource design and enterprise content management.

The emerging challenges also present new opportunities for research activity for the information science and digital library communities. There is a need to make information work more visible by clearly identifying the work being carried out by information specialists and other communities of work such as end users and business managers. Making sense of the relationships between the various communities of work and the way these relationships are mediated in practice will enable us to take a more holistic view of information resource design and management. In doing so we may conceive of new approaches and strategies for information resource design and information management appropriate to the increasingly complex and changing enterprise information environment.

4 Concluding Remarks

This position paper outlines a number of challenges associated with information resource design and enterprise information management. Three levels of information work (information resource design, information service provision and enterprise content management) have been identified, each taking a different perspective on information resource design ('usable', 'useful', 'adds value', respectively). In addition to the different levels of activity and perspectives, we find a range of communities of work associated with the development and management of information resources.

These challenges raise a number of new research questions, as outlined in the previous section. There is currently little in the way of theoretical or practical work to guide our understanding in this area. This leads to a call for a research agenda that seeks to understand and theorise about the way communities of information work are interleaved and the ways in which deep information design is achieved within organisations. Perhaps equally profoundly, it points to envisioning new roles for the information professional and new types of information work.

References

1. CSC, 14th Annual Survey of IS Management Issues, 2001 downloaded from: http://www.csc.com/aboutus/content/publications.shtml
2. CSC, 14th Annual Survey of IS Management Issues, 2000 downloaded from: http://www.csc.com/survey/ci_web/overview.html
3. Taylor, J.R, Groleau, C., Heaton, L. and Van Every, E. *The computerization of work: a communication perspective*. Thousand Oaks, CA: Sage Publications Inc., 2001.

4. Wenger, E. and Snyder, W.M. Communities of practice: the organisational frontier. *Harvard Business Review*, January-February, 2000, 139-145.

5. Wenger, E., McDermott, R., and Snyder, W.M. *Cultivating communities of practice*. Boston: Harvard Business School Press, 2002.

6. Fowell. S.P. Designing for enterprise use: a response to the changing information environment. *The New Review of Information and Library Research*. 2001, 93-107.

7. Marfleet, J. and Kelly, C. Leading the field: the role of the information professional in the next century. *The Electronic Library*, *17*(6), 1999, 359-364.

8. Rowbotham, J. Librarians – architects of the future? *ASLIB Proceedings, 51*(2), 1999, 59-63.

9. Ehrlich, K. and Cash, D. The invisible world of intermediaries: a cautionary tale. *Computer Supported Cooperative Work*, 8, 1999, 147-167.

10. Fourie, I. Should we take disintermediation seriously? *The Electronic Library, 17*(1), 1999, 9-14.

11. Sreenivasulu, V. The role of the digital librarian in the management of digital information systems (DIS). *The Electronic Library, 18*(1), 2000, 12-20.

12. Corcoran, M., Dagar, L., and Stratigos, A. Changing roles*.* *Online 24*(2), 2000, 29-34.

13. Corcoran M. Changing roles of information professionals: choices and implications. *Online 24*(2), 2000, 29-34.

14. Fichter, D. Search master: a new role for information professionals. *Online, 24*(2), 2000, 76-78.

15. Maly, K., Zubair, M. and Hesham, A. An automated classification system and associated digital library services. *Proceedings of the 1st International Workshop on New Developments in Digital Libraries*, 2001, 113-126.

16. Leung, H.K.N. Quality metrics for intranet applications. *Information and Management,* 38, 2001, 137-152.

17. Witten, I.H., McNab, R.J., Jones, S., Apperley, M., Bainbridge, D. and Cunningham, S.J. Managing complexity in a distributed digital library. *IEEE Computer*, February, 1999, 74-79.

18. Bates, M.J. Indexing and access for digital libraries and the Internet: Human, Database, and Domain Factors. *Journal of the American Society for Information Science*. 49(13), 1998, 1185-1205.

19. Kling, R and Elliott, M. Digital library design for usability. *Conference Proceedings Digital Libraries 94*, 1994. Downloaded from:
http://www.csdl.tamu.edu/DL94/paper/kling.html

20. Fowell, S.P., Alsmeyer, D. and Owston, F. Organisational usability and the BT digital library. *Proceedings of the 1st International Workshop on New Developments in Digital Libraries*, 2001, 34-46.

21. Bishop, A.P. and Star, S.L., Social informatics of digital library use and infrastructure. In; M.E. Williams, ed,. *Annual Review of Information Science and Technology*. Medford, NJ: Information Today, Inc. 1996, 301-401.

22. White, M. Enterprise information portals. *The Electronic Library*, 18(5), 2000, 354-362.

23. Davenport, T.A. *Information ecology: mastering the information and knowledge environment*. New York: Oxford University Press, 1997.

24. Marchand, D.A. (ed.) *Competing with information.* Chichester: John Wily and Sons Ltd, 2000.

25. Orna, E. *Practical information policies (2nd ed.).* Aldershot: Gower Publishing Ltd., 1999.

26. Orna, E. Information products revisited. *International Journal of Information Management*, 21, 2001, 301-316.
27. Drucker, P. F. *Management challenges for the twenty-first century*. Oxford; Butterworth-Heinemann, 1999.
28. Earl, M.J. Evolving the e-business. *Business Strategy Review*, 11(2), 2000, 33-38.
29. Shapiro, C. and Varian, H.R. *Information rules: a strategic guide to the networked economy*. Boston: Harvard Business School Press, 1999.
30. Meyer, M.H. and Zack, M.H. The design and development of information products. *Sloan Management Review*, Spring, 1996, 43-59.
31. Horne, N.W. Putting information assets on the board agenda. *Long Range Planning*, 31(1), 1998, 10-17.
32. Rayport, J.F. and Sviokla, J.J. Exploiting the virtual value chain. *Harvard Business Review,* November-December, 1995, 75-85.
33. Evans, P.B. and Wurster, TS. Strategy and the new economics of information. *Harvard Business Review*, September-October, 1997, 71-82.
34. Buchanan, S. and Gibb, F. The Information audit; an integrated strategic approach. *International Journal of Information Management*, 18(1), 199, 29-47.

A Digital Content Management Model for Making Profits in Digital Content Sites[*]

Hyeonjeong Mun[1], Sooho Ok[2], and Yongtae Woo[1]

[1] Changwon National University, Changwon Kyungnam 641-773, Republic of Korea,
{mun,ytwoo}@sarim.changwon.ac.kr
[2] Kosin University, Yeongdo-gu Pusan 606-701, Republic of Korea,
shok@kosin.ac.kr

Abstract. In this paper, we present a digital content management model to detect users who attempt to make an illegal copy of fee-charging contents. We define a set of rules to identify abnormal access behavior patterns. Our model can be applied to protect contents in fee-charging content sites.

1 Introduction

Fee-charging digital content sites, such as digital libraries and E-commerce areas, have increased in number and have become a common business proposition recently. It is important to protect content against illegal copying in these sites. However, existing content management models are insufficient to protect fee-charging contents[1,2]. In this paper, we present a digital content management model to detect 'abnormal' users who attempt to make an illegal copy of fee-charging contents. We define a set of four detection rules to monitor access patterns so as to differentiate between 'normal' and 'abnormal' users.

2 Detection Rules and Content Management Model

Reading Time Rule. This rule detects a user when the reading time of contents is abnormally longer or shorter than that of normal users.
Visit Frequency Rule. This rule detects a user when the number of visits is abnormally higher or lower than that of normal users.
Session Time Rule. This rule detects a user when the session time is abnormally longer or shorter than that of normal users.
Content Access Rule. This rule detects a user when the amount of the content accessed per session is abnormally larger or smaller than that of normal users.

Fig. 1 shows the conceptual diagram for the proposed model. The detection rule engine is a subsystem that executes the above detection rules to analyze the audit records. The user behavior monitoring engine is to identify whether a user is normal or abnormal by comparing his behavior with the behavior patterns.

[*] This research was financially supported by Changwon National University in 2001

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, pp. 516–517, 2002.

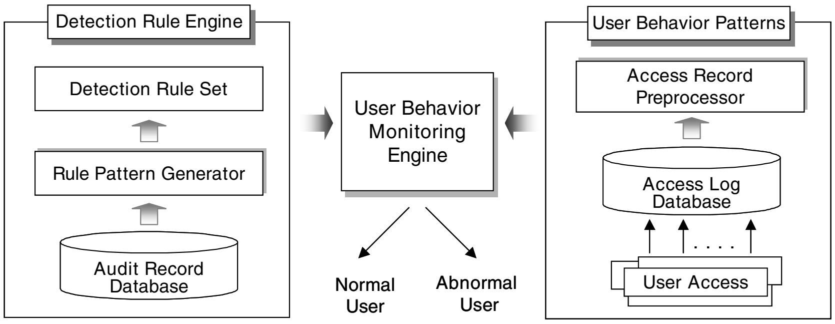


Fig. 1. User Monitoring Model in a Fee-charging Content Site

3 Experimental Results and Conclusion

The efficiency of the proposed model was tested on a recruit site, covering 16,661 users in one month. Table 1 gives the experimental results.

Table 1. Results of Behavior Patterns by Detection Rules

Detection Rule(Unit)	Average	Normal Range	Abnormal Range	
			Lower 10%	Upper 10%
Reading Time(min.)	0.91	0.37<r·t<1.57	0.02≦r·t≦0.37	1.57≦r·t≦20
Visit Frequency(times)	8	2<v·f<20	0≦v·f≦1	20≦v·f≦117
Session Time(min.)	11	1.17<l·t<22.6	0≦l·t≦1.17	22.6≦l·t≦832
Content Usage(number)	6	2.16<c·u<17.5	1≦c·u≦2.16	17.6≦c·u≦152

We have assumed that the abnormal range is 10% excess from average value. 142 users are in abnormal range for all detection rules. However, abnormal users detected in our model may not be illegal users, and normal users can act as abnormal users. Also, it is required application-dependent semantic rules such as detecting inconsistent behavior patterns that are mismatched with own's profile information to detect illegal users more precisely.

In conclusion, we propose a set of four detection rules to identify abnormal access behavior patterns. Our model can reduce the range of required monitoring considerably. Hence, our model can be applied to help protect contents in fee-charging content sites.

References

1. Ilgun, K., Kemmerer, R., Porras, P.: State Transition Analysis:A Rule-Based Intrusion Detection Approach. IEEE Tran. on Software Engineering (1995) 181-199
2. Petitcolas, F., Anderson, R., Kuhn, M.: Information hiding–A survey. In Proc. IEEE (1999) 1062–1078

The Idea of a Digital Library: Issues of Today

Swapan K. Dasgupta[1] and Asoke P. Chattopadhyay[2]

[1] Internet Centre and Central Library, and
[2] Department of Chemistry, University of Kalyani,
Kalyani 741235, India
Dasgupta_swapan@yahoo.com, asoke@klyuniv.ernet.in

Internet and the Digital Library (DL)

With the internet, libraries are being transformed from centralised collections to distributed service providers. There have been initiatives in this regard, e.g. the Stanford Digital Library Project. Significant advances have been made, such as the Data Documentation Initiative standards.[1] Archival formats are being hammered out, from images to historical documents. Some of the problems that remain, are:

Bridging the digital divide: the best academic library in India has a budget 25% of the lowest ranked (of 100 surveyed in 2001) of ALA member libraries.

Integrated service provider: provides not just data but knowledge, and service

Adoption of universal standards: a lot has been achieved, but more needs to be done, especially involving multimedia/not-text materials, protocols etc.

Addressing IPR issues: essential for library personnel and knowledge workers [2].

Staff development: most often, in these changing environments, they feel inadequate; hence, periodic and need-based training is needed for them [3].

Inspiring initiatives: the Human Genome Project, the GNU project and numerous scientific, technological, business and other collaborations across the globe. If these can succeed, why not one for digital libraries [4], for sorting out issues outlined above.

References

1. Witten, I. and Bainbridge, D.: How to build a Digital Library, Morgan Kaufmann,.(2002).
2. Litman, J.: Digital Copyright: Protecting Intellectual Property on the Internet. Prometheus. (2001).
3. Bertot, J.C. and McClure, C.R.: The 1997 National Survey of Public Libraries and the Internet: Costs, Connectivity, and Services. National Commission on Libraries and Information Science, Washington, DC. (1997).
4. BBC News Online, Thursday, 14 February, 2002.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 518, 2002.

US-Korea Collaboration on Digital Libraries: An Overview and Generalization for Pacific Rim Collaboration

Edward A. Fox[1], Reagan W. Moore[2], Ronald L. Larsen[3], Sung Hyon Myaeng[4], and Kim Sung-Hyuk[5]

[1] Virginia Tech, Dept. of Comp. Science, M/C 0106, Blacksburg, VA 24061 USA, fox@vt.edu
[2] San Diego Supercomputer Center, La Jolla, CA 92093-0505 USA, moore@sdsc.edu
[3] U. Pittsburgh, School of Info. Sciences, Pittsburgh, PA 15260 USA,
rlarsen@mail.sis.pitt.edu
[4] Chungnam National U., Div. Electr. & Comp. Eng., Taejon, Korea,
shmyaeng@cs.cnu.ac.kr
[5] Sookmyung Women's U., Div. of Information Science, Seoul, Korea,
ksh@sookmyung.ac.kr

Abstract. We report on recommendations to remove barriers to worldwide development of digital libraries, drawing upon an Aug. 2000 workshop involving researchers from the US and Korea who met at the San Diego Supercomputer Center. We developed a summary table identifying application domains (e.g., education), institutions engaged in those applications, example activities, technical challenges, and possible benefits. Since education has high priority in Korea, an important opportunity that should be explored is involvement of Koreans in the initiative led by the US NSF to develop the National STEM (Science, Technology, Engineering, and Mathematics) education Digital Library, NSDL.

There are many barriers to worldwide development of digital libraries (DLs). These are of particular concern in the context of DL support of collaboration on research and education between pairs of nations with very different languages and cultures. Recommendations to remove such barriers were developed at a US-Korea workshop involving DL researchers who met August 10-11, 2000 at the San Diego Supercomputer Center (see http://fox.cs.vt.edu/UKJWDL). A brief summary list of these includes:

1. Real / potentially significant collaboration opportunities should be nurtured.
2. An important opportunity to explore is involvement of Koreans in the US NSF led National Science education Digital Library, NSDL (www.nsdl.org).
3. Teams should apply to NSF's International Digital Libraries Collaborative Research and Applications Testbeds Program and similar programs in Korea.
4. Further focused programs might relate to one or more of:
   a) Korea Culture and Heritage Digital Libraries;
   b) ontologies and the areas dealing with human languages (e.g., machine translation, cross-language information retrieval, text summarization);
   c) application areas of particular importance, such as health / medical care;
   d) important DL problems, such as architecture (including managing data, information, and knowledge), interoperability, and user-centered design.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 519, 2002.

ETDs at HKU: A Spearhead for Digital Library Growth

David T. Palmer

The University of Hong Kong, Pokfulam Road, Hong Kong

Background

- HKU, 1912 established, 12,000 students
- Approximately 8,000 theses

A New Database – HKU Online Theses (HKUTO), from 1999

- Highlights important research source
- Bibliographic data extracted from catalogue, loaded into Oracle
- Enhanced with scanned abstract, TOC, and Chinese character names
- Indexes made for degree and program
- Searching in Roman or Chinese
- http://sunzi.lib.hku.hk/hkuto

Inter-Departmental Cooperation

- HKU Senate, HKU Registry Research Unit, & HKU Graduate School
- November 2000 new requirement approved for submission of full-text

Format

- Now PDF, change to XML?
- Porting HKUTO into Tamino XML Server

Interoperability

- UMI's Dissertation Abstracts?
- Networked Digital Library of Theses & Dissertations (NDLTD)
  - Federated Search
- Open Archive Initiative (OAI)
  - HKUTO is OAI compliant and searchable in OAI Service Providers
- NDLTD's ETD OAI Union Catalog
- VTLS' Networked Digital Library of Theses & Dissertations

Lessons Learned, Impetus for Other Databases

- Hong Kong Table of Contents Database (HKTOC)
  - Scanning of TOC, OCR on English & Chinese
- Web lists of HKUL databases, e-journals, e-books and e-news
  - Scripted (Expect) extraction from catalogue
- E-Reserves
  - Scanning, PDF creation & copyright issues
- HKU Research & Scholarship Database
  - Cooperation with other departments, copyright issues
- Sun Yat-sen in Hong Kong database
  - XML

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 520, 2002.
© Springer-Verlag Berlin Heidelberg 2002

MiMedicalLibrary: A Digital Health Library for Michigan

Harvey R. Brenneise

Michigan Community Health Electronic Library
Michigan Public Health Institute
2436 Woodlake Circle
Okemos, MI 48864 USA
hbrenne@mphi.org

MiMedicaLibrary is the joint digital library project being developed by members of the Michigan Health Sciences Library Association (MHSLA). The first step in its development was a 2-year planning process (1998-2000) called the AccessMichigan Community Health Electronic Library (AMECHII), building on the foundation of the statewide general digital library, AccessMichigan, and a call for action from the report of the Michigan Technology Commission Report.

When outside funding for implementation did not materialize as envisioned by that report, MHSLA developed a second model for development, which depends entirely on the project being self-funded. In October 2001, a pilot project was implemented with pooled access to the Stat!Ref suite of digital medical textbooks from Teton Data Systems. With a single license for 40 simultaneous users, participating libraries use a “Priceline” model of funding in which they voluntarily choose how many of the total number of simultaneous users they will pay for. This pilot has been judged successful and is just beginning its second year. Because of its success, member libraries are seeking additional digital resources to acquire, using the same model or another as appropriate.

The newest project is determining if this model can be applied to publisher contracts for electronic serials, loosely following the OhioLINK model as well as nation-wide projects already in place in Belgium, Sweden and Iceland using SWETS Blackwell as serials vendor and publisher negotiator. Negotiations are under way at this time, and additional (non-medical) Michigan libraries may be invited to join the project if it is deemed mutually beneficial.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 521, 2002.

Reference Services in a Digital Library of Historical Artifacts[1]

Roy Chan, Dion Goh, and Schubert Foo

Division of Information Studies
School of Communication and Information
Nanyang Technological University
Singapore 637718
royclу@yahoo.com, {ashlgoh, assfoo}@ntu.edu.sg

The National Archives of Singapore (NAS) is a Singapore government organization tasked with the preservation of national records. The NAS offers a reference service where NAS staff respond to public enquiries on archival resource matters. Over time, it was noticed that certain enquiries were being asked more frequently than others. Hence, physical folders were maintained to hold information on subjects that were identified as popular enquiries. When swarmed with enquiries, however, this paper-based approach was found to be inadequate. The Enquiry Database (ED) was thus conceived to automate the reference service as well as to eliminate the necessity for the public to enquire in person at the NAS' premises for common enquires.

The ED uses a subject hierarchy in which broad terms branch out to narrower terms that are more precise or specific. Leaf subjects within this hierarchical organization are represented as virtual folders that contain the information sought by users. Functions supported in the ED include the ability to browse the subject hierarchy, search for folders, add/create new folders, modify existing folders, etc., through a web-based interface.

The ED is part of the NAS digital library project which aims to build a unified platform upon which all of the NAS' digitized resources may be accessed through various applications. The ED maintains an application-specific repository that stores the electronic equivalent of the physical folders. Folder contents are references to resources (e.g. photographs, maps and their associated metadata) found within other repositories of the NAS digital library. This approach is in line with the digital library's philosophy of reusability, where resources are stored in centralized repositories and reused across multiple applications.

The ED is expected to help free up time and resources for NAS staff, as repetitive and routine tasks are automated. In addition, the ED will be available online to users even when the physical reference helpdesk closes for the day, thereby offering a value-added service that enhances the image of the NAS reference service as an easily reached public information resource center.

[1] Resources used in this project were obtained in collaboration with the National Archives of Singapore.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 522, 2002.

An Integrative User-Centered Purchase Request Service in the Age of Digital Library Development

Ching-Fan Wu[1] and Hui-Chen Hsieh[2]

[1] National Taitung Teachers College Library,
684 Chunghua Rd., Sec. 1, Taitung 950, Taiwan
frank@cc.ntttc.edu.tw

[2] Wenzao Ursuline College of Languages Library,
900 Mintsu 1st Rd., Kaohsiung 807, Taiwan
hchsieh@mail.wtuc.edu.tw

Abstract. The Integrative Purchase Request Service System designed by Taiwan's National Taitung Teachers College Library aims to provide a proactive and user-friendly purchase request service to affiliated users. They may search in the bibliographic database and click on items requested for purchase rather than type in bibliographic information, thus preventing typos and other errors. It also benefits both acquisitions and cataloger librarians – in expediting orders and copy cataloging.

To avoid typos and other errors in print or web-based purchase request forms, the National Taitung Teachers College Library in Taiwan, in early July 2002, started developing a system to offer a proactive and user-friendly purchase request service to affiliated users.

The Integrative Purchase Request System integrates the local library automation system and imports the bibliographic database with authorization from ISBNnet, which issues ISBN and creates CIP records for all publications since 1989 in Taiwan. Hence, the system provides users with comprehensive bibliographic information. Users search for items by title keywords, ISBNs, publishers, or classification numbers. Once logged in the system, affiliated users are authorized to submit purchase requests to acquisitions librarians through the system. Furthermore, users may determine if items they request for purchase are available in local holdings and their circulation status.

This system (accessible at http://acq.lib.ntttc.edu.tw) has already been announced for use. To date, it allows affiliated users to recommend only books published in Taiwan for library purchase. It also expedites the acquisition process and improves library service quality. The ultimate goal is to provide users with SDI services on up-to-date publication information, to integrate bibliographic databases of major online bookstores worldwide, and to expedite the availability of user requested materials for better library services.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 523, 2002.

Vitalising Library and Information Science Education: A Challenge in the Digital Information Environment

K. Padmini

Associate Prof., Dept.of LIS, S.V.University, Tirupati, A.P.

In the context of the emerging information society, it has become very important to the LIS schools that information professionals are properly trained, enabling them to adapt to the ever changing information environment. Areas to be vitalized are infrastructure, manpower, curriculum, mechanisms, and evaluation.

WHO SHOULD DO WHAT?

1. LIS Schools

- *Standardisation of curriculum*
- *Make use of expertise available in other departments.*
- *Measures to improve communication skills of the students*
- *Establish close rapport with R & D, industry, and higher learning institutions.*
- *Internship for students.*
- *Acquiring multi media products*
- *Re-orientation and re-education of the faculty*
- *Adequate infrastructure.*

2. LIS Educators

- *Psychological readiness*
- *Updating their knowledge*
- *Periodic self-appraisal*
- *Only really interested persons should enter the profession*

3. Administration

- *Recognizing the performance of educators*
- *Facilities for continuous competence development.*
- *Should provide minimum requirement of infrastructure.*

4. Associations

- *Compile database of faculty members*
- *Prepare research in-progress bulletins*
- *Sincere efforts towards implementation of the recommendations made at various national and international forums*

5. Governments

- *Much attention is required.*
- *Establish a national accreditation mechanism for LIS.*
- *While restructuring the curricula, symbiosis must be maintained.*
- *Should recognize the need for information professionals and provide job opportunities.*

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 524, 2002.

Developing a Dialogue Library System

Ivan Kopecek [1] and Miroslav Bartosek [2]

[1] Faculty of Informatics and [2] ICS, Masaryk University, Botanicka 68a, Brno, Czech Rep.
kopecek@fi.muni.cz, bartosek@ics.muni.cz

The concept of using a dialogue system as an interface to digital libraries is supported by the library dialogue system, which is being developed at Masaryk University. The basic idea is to combine the possibilities of standard library systems (TINLIB, ALEPH, VOYAGER, etc.) and digital libraries of any kind (ACM DL, Idealibrary, ScienceDirect, pre-print archives, etc.) with the user comfort that can assure a dialogue system.

To integrate these systems, there are basically two ways to achieve the level of interoperability required by a specific community of users. The first one is to use a strong standard, like the Z39.50 search and retrieval protocol, which provides a system independent interface enabling searching and manipulation of the data items. The drawback of this approach is that not all library and DL systems support this kind of protocols. The second possibility is to use a web-based interface. Practically all library systems and digital libraries available today offer the possibility of sending queries to the system encoded in a URL and then extracting the data from the corresponding HTML document. An increasing number of DLs are already OpenURL enabled, providing a reasonable level of standardization for a URL-based approach to integration. Both methods may be appropriately combined. Via a web-based interface, the dialogue system can utilize the basic library modules, like Online Public Access Catalogue, Cataloguing module, Circulation module, Interlibrary Loan Service module, as well as search or browse any other user service modules.

Maximal comfort of the user can be achieved by an appropriate combination of the strategies with system initiative, user initiative and mixed initiative strategies [3]. Error handling, clarifying dialogues and meta-communication are typically processed via the strategies with system initiative. Dialogue strategies are implemented in VoiceXML [2]. In order to have full control over the interpreter capability, compatibility and speed, the research team has developed the VoiceXML interpreter ELVIRA [1], which serves at the same time as a module of the library dialogue system. (ELVIRA can be downloaded from [4]).

References

1. Cenek, P.: Dialogue Interfaces for Library Systems; FI MU Report Series, FIMU-RS-2001-04, June 2002.
2. VoiceXML Forum – http://www.voicexml.org/specs/VoiceXML-100.pdf
3. Kopecek, I.: Active and Passive Strategies in Dialogue Program Generation. In Proceeding of TSD 2000, LNAI 1902, Springer Verlag, pp. 427–432, 2000.
4. http://www.fi.muni.cz/lsd/elvira/

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 525, 2002.

WebClipper: A Personal/Community Link Library Builder Based on Web Link Management Technique

Young S. Kim, Jae H. Lim, Soon J. Hyun, and Dongman Lee

School of Engineering, Information and Communications University(ICU)
P.O.Box 77, Yuseong, Daejeon, 305-600, Korea
{seung92,theagape,shyun,dlee}@icu.ac.kr

Information finding on the Web is well served by many high-tech search engines. The information found can be collected for sharing and reuse, by individuals as well as the community at large. This will effectively narrow the search domain for web surfers, leading to a reduction in the cognitive load placed on them. This paper reports the development of a personal/community link library building system, called WebClipper, which enables individual and community users to collect index data from Web documents. WebClipper manages the various links of a Web document and extracts its index data to form a virtual digital library. Using WebClipper, a user can collect links and index data from useful Web documents to form his own link library. When searching, he can first search his own link library before going out to navigate the Web. In the community, users can pool their own link libraries for all to use. In this paper, we show an implementation of Web Clipper. It comes with a Web link database management scheme, a Web information clipping technique, and a user interface design. The proposed Web link database management technique allows the system to manipulate link information without archiving physical data, thus eliminating the problem of storage and copyright. The clipping technique allows users to conveniently choose the granularity of the data to be collected; it also carries out automatic keyword extraction and indexing. Database management of link data enables user-defined link descriptions and makes deadlink management easy.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 526, 2002.

Hiding a Logo Watermark in an Image for Its Copyright Protection

Jun Zhang[1], Feng Xiong [2], and Nengchao Wang[1]

[1] School of Computer Science & Technology, Huazhong University of Science & Technology, Wuhan 510320, P.R.China
[2] School of computer, Guangdong Commercial College, Guangzhou 510320, P.R.China

Abstract. Due to the urgent need for protecting the copyright of digital products in a digital library, this paper proposes a novel public watermarking scheme for a still image. Firstly, the image is decomposed to the multiwavelet domain, in which there are four subblocks in the coarsest resolution level. Then a logo watermark is embedded into the low frequency band by quantizing the difference value between corresponding coefficients in a pair of subblocks in the coarsest resolution level. Experimental results show that the proposed scheme is superior to the conventional quantization based approaches in terms of robustness.

1 Idea of the Proposed Method and Experimental Results

Kundur et al. [1] propose a quantization-based public watermarking scheme, in which they divide a real number axis in the wavelet domain into intervals with equal size and assign watermark bit to each interval periodically. In this paper, we embed a logo watermark bit into a new transformation domain---multiwavelet domain using the difference quantization-based scheme. Owing to the fact that corresponding coefficients in two subblocks in the coarsest resolution level are both approximations of the same section of the original image, the difference value is more stable than the single coefficient. The correct ratios of our method and Kundur's are listed in Table 1. It is clearly obvious that our method has superior performance.

Table 1. Comparison results between our method and Kundur's method

	JPEG	Median	Noise	Sharpening
Kundur's method	69.36	89.33	76.66	71.19
Our method	90.53	96.56	98.66	84.28

References

1. D.Kundur, D.Hatzinakos: Digital Watermarking for Telltale Tamper Proofing and Authentication. *Proceedings of the IEEE*, Vol 87 (1999) 116-1180.

This work is supported by the Nature Science Foundation of Guangdong Province. (020199)

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 527, 2002.

Searching Video Segments through Transcript, Metadata, and SVG Objects

M. Karthikeyan[1], D. Uzagare[1], S. Krishnan[1], Meng Yang[2], and
Gary Marchionini[2]

[1]Information Division,
National Chemical Laboratory, Pune 411 008, India
{karthi,krish}@ems.ncl.res.in
[2]Interaction Design Lab, School of Information and Library Science
University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3360
{yangm,march}@ils.unc.edu

Abstract. The objective of this paper is to highlight our attempt in to search video content based on transcript text, metadata, SVG objects, etc.. Scalable Vector Graphics (SVG) components extracted from video frames can be used for multiple purposes, in addition to shape-based search of video frames. It is also proposed to integrate text and SVG data along with video data as a searchable component.

We want to search video for content to synchronize with transcript text generated from audio, other metadata, and SVG objects, in a timely and efficient manner. At present, the user needs to preview the video, downloading and streaming the entire video, which is a laborious and time-consuming process. There is a need to develop a mechanism to synchronize the video themes with text-based searching. Transforming video images to scalable vector graphics (SVG) is also a major challenge. There is a need to reduce unwanted information from SVG to get object outline information as shown in Figure 1. This process reduces the newly created file manifold.

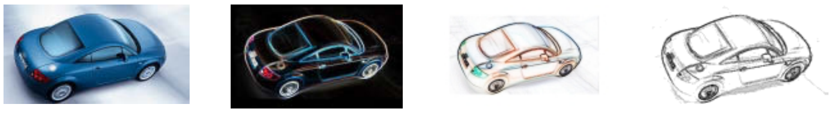


Fig. 1. Transformation of Video frame into SVG object

To demonstrate the advantages of the proposed technique, an IDLVideoSearch[1] program was developed using Java programming language. This program is an attempt to provide a search tool to find and view a particular segment in an entire video based on the search query. The block diagram is described in Figure 2.

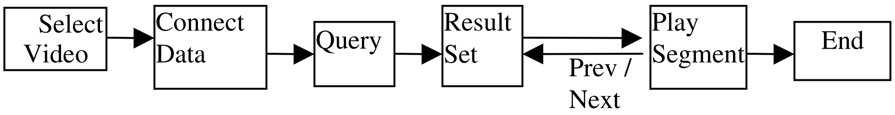


Fig. 2. Block diagram of IDLVideoSearch Program

Reference

1. IDLVideoSearch Program developed in Java environment: For details visit URL: http://www.ncbi.org.in/cili/video.htm

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 528, 2002.

Similar Sub-trajectory Retrieval Based on k-Warping Distance Algorithm for Moving Objects in Video Databases

Choon-Bo Shim and Jae-Woo Chang

Dept. of Computer Eng., Chonbuk National Univ., Chonju, Chonbuk 561-756, Korea
{cbsim,jwchang}@dblab.chonbuk.ac.kr

1 Introduction

Recently, a lot of research has been done on content-based video retrieval using the location information of mobile objects. To support similar sub-trajectory retrieval on moving objects in video databases, we present three considerations as follows. First, similar sub-trajectory retrieval should allow the replication of only a query trajectory. Secondly, it should allow the replication of up to the fixed number (k) of motions, so called k-warping distance. Finally, it should support a distance property as well as the angle property. Therefore, we propose a new k-warping distance (D_{kw}) algorithm for similar sub-trajectory retrieval. The proposed algorithm calculates a k-warping distance between a given query trajectory and a data trajectory by permitting up to k replications for an arbitrary motion of a query trajectory. $d_{df}(s[i], q[j])$ means distance function between i-th motion of data trajectory *s* and j-th motion of query trajectory *q*.

$$D_{kw}(0,0) = 0,\ D_{kw}(s,0) = D_{kw}(0,q) = \infty$$

$$D_{kw}(s,q) = d_{df}\ (s[1],q[1])+\min\{D_{kw}((s[2+i:-], q),\ 0\leq i<k),\ D_{kw}(s[2:-], q[2:-])\}$$

2 Performance Analysis

To verify the usefulness of our algorithm, we performed analysis on 350 real soccer video data. We compared our k-warping distance algorithm with Li [1] and Shan [2] in terms of retrieval effectiveness, that is, average precision and recall. In the case when the weight of the angle is about two times that of distance (w_a =0.7 and w_d =0.3), it is shown that our scheme achieves about 15-20% higher precision than those of Li and Shan, while it has about the same recall.

References

[1] J. Z. Li, M. T. Ozsu, and D. Szafron, "Modeling Video Temporal Relationships in an Object Database Management System," MMCN'97, pp. 80-91, 1997.

[2] M. K. Shan and S. Y. Lee, "Content-based Video Retrieval via Motion Trajectories," SPIE Electronic Imaging and Multimedia System II, Vol. 3561, pp. 52-61, 1998.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 529, 2002.

A Keyword Spotting System of Korean Document Images

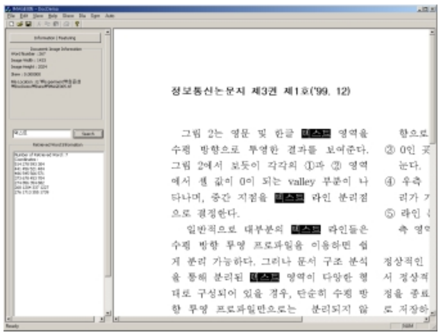
Il-Seok Oh[1], Yoon-Sung Choi[1], Jin-Ho Yang[1], and Soo-Hyung Kim[2]

[1]Department of Computer Science, Chonbuk National University, Korea
[2]Department of Computer Science, Chonnam National University, Korea

As a result of the Korean NDL (National Digital Library) project, an enormous amount of old paper documents are in the database in a digital image format. However, it appears that advanced services like full-text retrieval were not developed for the document images.

This paper presents a keyword spotting system for these document images. A two-stage *coarse-to-fine* retrieval scheme was devised. A word image is segmented into character images by analyzing the vertical projection. After size-normalizing the character image into 32*32 array, two kinds of features are extracted. First, four profiles are computed from the top, bottom, left, and right sides. Each profile is represented as a one-dimensional array with 32 values. By averaging the 32 values for each of four sides, a coarse 4-dimensional feature set is obtained. For the second feature set, the character image is wavelet-transformed using Harr operator and the 30 largest coefficients are taken as the fine feature set.

In the first stage, coarse matching is performed using the 4-dimensional profile feature set. This stage aims at selecting candidate words at a very high speed. The matching is accomplished using the Euclidean distance between the query and target word images. The matching for the second stage is accomplished in a similar manner. The parameters in the matching conditions for the first stage were set to get a high recall rate while allowing low precision. On the contrary, high rates both in recall and precision were produced by the second stage.



정보통신논문지 제3권 제1호('99. 12)

그림 2는 영문 및 한글 텍스트 영역을 수평 방향으로 투영한 결과를 보여준다. 그림 2에서 보듯이 각각의 ①과 ② 영역에서 셀 값이 0이 되는 valley 부분이 나타나며, 중간 지점을 텍스트 라인 분리점으로 결정한다.

일반적으로 대부분의 텍스트 라인들은 수평 방향 투영 프로파일을 이용하면 쉽게 분리 가능하다. 그러나 문서 구조 분석을 통해 분리된 텍스트 영역이 다양한 형태로 구성되어 있을 경우, 단순히 수평 방향 투영 프로파일만으로는 분리되지 않

향으로
② 0인 곳
는다.
④ 우측
리가
⑤ 라인
측 영
정상적인
시 정상적
점을 종료
로 저장하

A prototype system was implemented using C++ language on the IBM-PC platform. The page image layout analysis module developed by Chonnam National University was installed into the system. The Figure illustrates the spotted words.

An evaluation was performed using 238 page images scanned from technical journals in the Korean language. The page images are in fairly good printing quality and scanned in 300 dpi binary tone. The speed was 199,918 words/second and accuracy was 99.6% recall and 97.5% precision. When one stage with only the wavelet feature set was applied, similar accuracy was obtained and the speed was 33,063 words/second. A conclusion can be drawn that the two-stage scheme accelerates the retrieval module by about 6 times without losing any accuracy. Another evaluation was performed on the medium-quality document images in 200 dpi from the *Journal of Korean Information Science Society*. The recall and precision were 91.0% and 88.0%, respectively.

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 530, 2002.

An Efficient Strategy for Adding Bulky Data into B+-Tree Indices in Information Retrieval Systems*

Jin-Ho Kim[1], Ju-Young Kim[1], Sang-Wook Kim[2], Soo-Ho Ok[3], and Hi-Young Roh[1]

[1] Dept. of Computer Science, Kangwon Nat'l University,
[2] Dept. of Inf. & Telecomm. Eng., Kangwon Nat'l University,Chunchon, Republic of Korea
[3] Dept. of Computer Science, Kosin University, Busan, Republic of Korea
{jhkim,yjkim,wook,young}@kangwon.ac.kr, shok@koshin.ac.kr

In Internet-based information retrieval systems, a large amount of data are periodically collected from the Internet by robot agents and stored into a database. For fast retrieval, their key values are also added into an existing index, widely constructed as an inverted file using a B+-tree. Obviously, the simplest method of adding multiple key values into a B+-tree is to apply the insertion algorithm repeatedly. However, this method inserts new key values into a B+-tree in a random order without considering adjacency between them. This makes each page within the B+-tree accessed a lot of times, and thus requires large processing time. To solve this problem, we present a new method called Bulk_Add that effectively adds a bulky set of new key values into an existing B+-tree index.

Bulk_Add employs the basic idea of the *batch-construction* of a B+-tree [1], which efficiently builds a B+-tree by placing a set of key values on each B+-tree node with only one disk access. We simply sketch our Bulk_Add as follows:

1. Sort new key values to place these new leaf entries of the B+-tree with similar key values adjacent to one another.
2. Merge the result of step 1 with the leaf entries of the existing B+-tree.
3. Apply the batch-construction algorithm[2] to the result of step 2 to build a new B+-tree that contains both old and new key values.

In case of adding *N* key values into an existing B+-tree with *M* key values, the cost of Bulk_Add in terms of the number of disk accesses is analyzed as follows:

$$TotalCost = (N/B) * (\log_k (N/B) - 1) * 2 + (M/B) + \sum_{k=1}^{\log_k (N+M)} ((N+M)/B^k)$$

where *B* is a blocking factor of a B+-tree node and *k* is the number of input buffers for *k*-way merging. With extensive experiments for performance evaluation, we observed that Bulk_Add outperforms the method using the insertion algorithm significantly. More specifically, Bulk_Add achieves 3 times speedup when *N* is 1% of *M*, and 15 times speedup when *N* is 5% of *M*. The results also show that the performance gain gets much higher as *N* increases. We conclude that Bulk-Add enhances the system performance considerably when a large volume of data are periodically added to existing databases.

1. Kim, S. W., "Batch-Construction of B+-trees: Algorithm and Its Performance Analysis," Journal of the Korea Information Science Society(B), Vol. 23, No. 11, pp. 1113-1121, 1996.

* This work was supported by the KOSEF through Advanced Information Technology Research Center, the Research Project(Grant KRF-2000-041-E00258) of Korea Research Foundation, and the Basic Research Project(Grant R05-2002-000-01085-0) of KOSEF

E.-P. Lim et al. (Eds.): ICADL 2002, LNCS 2555, p. 531, 2002.

Author Index

Abdoulaye, Kaba 366
Acharya, Raj 179
Adams, Anne 392
Allen, R.B. 111
Anagnostopoulos, D. 256
Anastasiou, D. 256
Anderson, Terry 440
Ang, T.F. 469
Anutariya, Chutiporn 153
Azcarraga, Arnulfo P. 218

Bacchin, Michela 117
Bainbridge, David 129
Bakar, Abu 366
Baker, Thomas 25
Bakeri, Ahmad 366
Bartosek, Miroslav 525
Basile, T.M.A. 182
Blandford, Ann 392
Borgman, Christine L. 1
Braddick, Hywel 275
Brenneise, Harvey R. 521

Chakilam, Venu 165
Chan, Roy 522
Chang, Jae-Woo 529
Chattopadhyay, Asoke P. 518
Chen, Chao-chen 425
Chen, Hsinchun 36
Chen, Hsueh-hua 425
Cheng, Jun 352
Choi, Key-Sun 315
Choi, Yoon-Sung 530
Chotmanee, Ano 153
Chowdhury, Gobinda G. 379
Chua, Lian-Heong 427
Chua, Tat-seng 218, 241
Cunningham, Sally Jo 177, 301

Dasgupta, Swapan K. 518
Deng, Zhi-Hong 88
Dewsnip, Michael 129

Edzan, N.N. 469
Esposito, F. 182

Faloutsos, Christos 194
Fanizzi, N. 182
Feng, Ling 495
Ferilli, S. 182
Ferro, Nicola 117
Foo, Schubert 260, 465, 522
Fowell, Sue 507
Fox, Edward A. 14, 299, 519
Fu, Lin 260
Furuta, Richard 467

Gao, Wen 352
Goh, Dion Hoe-Lian 99, 260, 287, 427, 522
Goudar, I.R.N. 303
Gross, Benjamin M. 258

Harinarayana, N.S. 305
Hatzopoulos, M. 256
He, Zhenying 421
Heng, Junius Soh Hock 141
Hoffman, Ellen 299, 404
Hoppenbrouwers, Jeroen 495
Hsieh, Hui-Chen 523
Huang, Tie-jun 352
Hussam, Ali 440
Hwang, San-Yih 229
Hyun, Soon J. 526

Jacobs, Nathan 440
Jeusfeld, Manfred 495
Jose, Joemon M. 275

Kageura, Kyo 315
Karadkar, Unmil P. 467
Karthikeyan, M. 528
Ke, Hao-Ren 76
Kei, Jeremy R. 241
Khoo, Christopher S.G. 99
Kim, Do-Wan 315
Kim, Go U. 165
Kim, Jin-Ho 531
Kim, Ju-Young 531
Kim, Sang-Wook 531
Kim, Soo-Hyung 530
Kim, Sung H. 165
Kim, Young S. 526

Kiyoki, Yasushi 471
Kopecek, Ivan 525
Krishnan, S. 528
Krottmaier, Harald 254
Kumbar, Mallinath 305

Larsen, Ronald L. 519
Lee, Dongman 526
Lee, Kyung-Soon 315
Li, Chao 423
Li, Jianzhong 421
Li, Kar Wing 328
Lim, Ee-Peng 229, 287, 427
Lim, Jae H. 526
Liu, Bin 352
Liu, Jimin 241
Liu, Zehua 287, 427
López, Javier 483
Luo, Ming 14

MacPherson, Greg 275
Maña, Antonio 483
Marchionini, Gary 528
Mardis, Marcia 404
Martin, Innes 275
Mauro, N. Di 182
Melucci, Massimo 117
Menon, Raghu 179
Meyyappan, Narayanan 465
Moore, Reagan W. 519
Mun, Hyeonjeong 516
Myaeng, Sung Hyon 519

Na, Jin-Cheon 467
Narayana, Poornima 303
Nicholas, David 406
Nikolaidou, M. 256

Oh, Il-Seok 530
Ok, Soo-Ho 516, 531
Ou, Shiyan 99

Padmini, K. 524
Palmer, David T. 520
Pan, Jia-Yu 194
Pang, Natalie Lee-San 427
Pendleton, Michael 406
Pimentel, Ernesto 483
Plummer, Bill 440
Polydoratou, Panayiota 406

Prey, Jane C. 53

Razdan, Anshuman 419
Renda, M. Elena 262
Robertson, Brian 275
Roh, Hi-Young 531
Rowe, Jeremy 419

Sasaki, Hideyasu 471
Semeraro, G. 182
Shim, Choon-Bo 529
Shrestha, Jagan Nath 354
Sinthupinyo, Sukree 340
Srikant, Ramakrishnan 75
Straccia, Umberto 262
Sugimoto, Shigeo 25
Suleman, Hussein 14
Sung-Hyuk, Kim 519
Sweitzer, Wilson 419

Tan, Jonathan 218
Tang, Shi-Wei 88
Theeramunkong, Thanaruk 340
Theng, Yin-Leng 427, 452
Tocheri, Matthew 419
Troya, José M. 483

Urs, Shalini R. 305
Uzagare, D. 528

Vaidya, Binod 354

Walker, Craig 275
Wang, Hongzhi 421
Wang, Nengchao 527
Weibel, Stuart L. 25
Witten, Ian H. 67, 129
Wong, Patricia Bao-Bao 427
Wongtapan, Chainat 340
Woo, Yongtae 516
Wu, Ching-Fan 523
Wu, Xiao-Bin 88
Wu, Yejun 111
Wuwongse, Vilas 153

Xing, Chunxiao 206, 252, 423
Xiong, Feng 527

Yagüe, Mariemma I. 483
Yang, Christopher C. 328
Yang, Dong-Qing 88

Yang, Jin-Ho 530
Yang, Meng 88, 528
Yang, Wei-Pang 76
Ye, Shiren 241
Yeh, Jen-Yuan 76
Ying, Phoon Moi 141
Yoon, Jong P. 165

Zainab, A.N. 469
Zeng, Chun 206, 252
Zhang, Jun 527
Zhang, Ling 352
Zhang, Ming 88
Zhang, Zhiqiang 206
Zheng, Xiaohui 252
Zhou, Lizhu 206, 252, 423
Zhou, Xiaodan 206
Zia, Lee L. 53