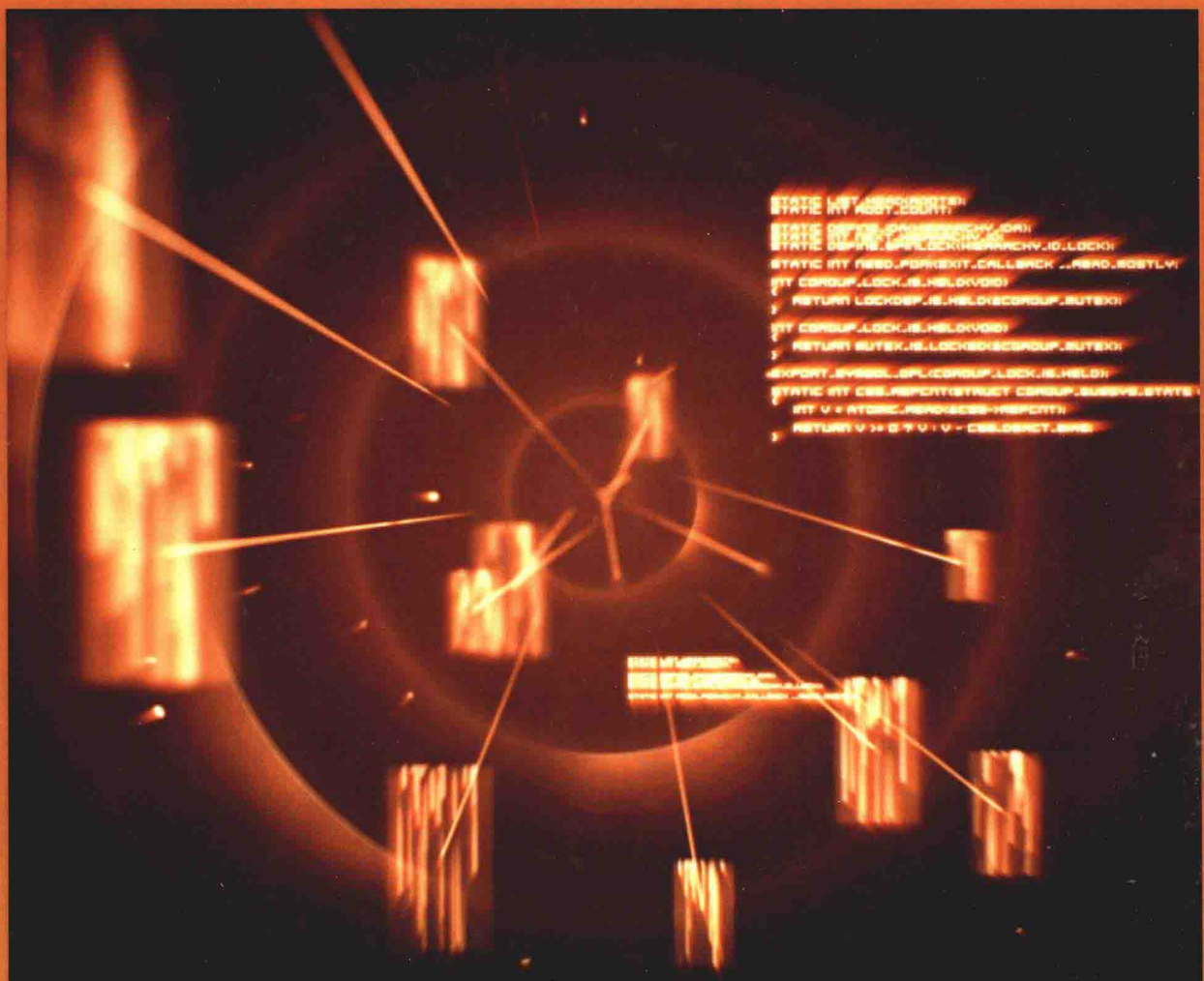


PREMIER REFERENCE SOURCE

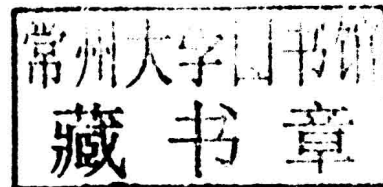
# Information Retrieval Methods for Multidisciplinary Applications



Zhongyu (Joan) Lu

# Information Retrieval Methods for Multidisciplinary Applications

Zhongyu (Joan) Lu  
*University of Huddersfield, UK*



Information Science  
**REFERENCE**

|                                |                   |
|--------------------------------|-------------------|
| Managing Director:             | Lindsay Johnston  |
| Editorial Director:            | Joel Gamon        |
| Book Production Manager:       | Jennifer Yoder    |
| Publishing Systems Analyst:    | Adrienne Freeland |
| Development Editor:            | Joel Gamon        |
| Assistant Acquisitions Editor: | Kayla Wolfe       |
| Typesetter:                    | Lisandro Gonzalez |
| Cover Design:                  | Jason Mull        |

Published in the United States of America by  
Information Science Reference (an imprint of IGI Global)  
701 E. Chocolate Avenue  
Hershey PA 17033  
Tel: 717-533-8845  
Fax: 717-533-8661  
E-mail: [cust@igi-global.com](mailto:cust@igi-global.com)  
Web site: <http://www.igi-global.com>

Copyright © 2013 by IGI Global. All rights reserved. No part of this publication may be reproduced, stored or distributed in any form or by any means, electronic or mechanical, including photocopying, without written permission from the publisher. Product or company names used in this set are for identification purposes only. Inclusion of the names of the products or companies does not indicate a claim of ownership by IGI Global of the trademark or registered trademark.

#### Library of Congress Cataloging-in-Publication Data

Information retrieval methods for multidisciplinary applications / Zhongyu (Joan) Lu, editor.  
pages cm

Includes bibliographical references and index.

Summary: "This book provides innovative research on information gathering, web data mining, and automation systems, addressing multidisciplinary applications and focusing on theories and methods with an enterprise-wide perspective"-- Provided by publisher.

ISBN 978-1-4666-3898-3 (hardcover) -- ISBN 978-1-4666-3899-0 (ebook) -- ISBN 978-1-4666-3900-3 (print & perpetual access) 1. Data mining. 2. Querying (Computer science) 3. Information storage and retrieval systems. I. Lu, Zhongyu, 1955-

QA76.9.D343I546 2013

006.3'12--dc23

2012048660

#### British Cataloguing in Publication Data

A Cataloguing in Publication record for this book is available from the British Library.

The views expressed in this book are those of the authors, but not necessarily of the publisher.

## Associate Editors

Hamid R. Arabnia, *The University of Georgia, USA*  
Andrew Ball, *University of Huddersfield, UK*  
Takashi Hasuike, *Osaka University, Japan*  
Sabah Jassim, *Bacukingham University, UK*  
Andy Marsh, *HOIP, UK*  
Peter Slood, *University of Amsterdam, The Netherlands*  
Qiang Xu, *Teesside University, UK*  
Shaoyan Zhu, *Tsinghua University, China*

## List of Reviewers

Chuming Chen, *University of Delaware, USA*  
Carmen Costilla, *Technical University of Madrid, Spain*  
Roger E. Eggen, *University of North Florida, USA*  
Hiroaki Fukuda, *Keio University, Japan*  
Jaafar Gaber, *Université de Technologie de Belfort-Montbéliard, France*  
Shanmugasundaram Hariharan, *J. J. College of Engineering and Technology, India*  
Sean He, *University of Technology, Australia*  
Nadine Jessel, *University Paul Sabatier, France*  
Ni Jun, *The University of Iowa, USA*  
Qu Junfeng, *Clayton State University, USA*  
Michel Kamel, *University of Nice, France*  
Elena Kozerenko, *The Russian Academy of Sciences, Russia*  
Stefania Marrara, *University of Milan, Italy*  
Andrea Morgera, *Università degli Studi di Cagliari, Italy*  
Kia Ng, *University of Leeds, UK*  
Wolfgang Renz, *Hamburg University, Germany*  
Weiming Sheng, *National Research Council of Canada, Canada*  
Yunchun Shi, *Tsinghua University, China*  
John B. Stav, *Norwegian University for Technology and Science, Norway*

Hamish Taylor, *Heriot-Watt University, UK*  
Bezboruah Tulshi, *Gauhati University, India*  
Lizhen Wang, *Yunnan University, China*  
Yuxin Wang, *The University of Tokyo, Japan*  
Weidong Yang, *Fudan University, China*  
Weizhong Yang, *GE, USA*  
Shaowen Yao, *Yunnan University, China*

# Preface

## AN OVERVIEW OF THE SUBJECT MATTER

Information retrieval is ubiquitous. It is directly linked to multidisciplinary and interdisciplinary applications. Discovery of findings from scientific research requires it. Analysis of experiment results in engineering and technology requires it. Obtaining arguments and opinions in social science requires it. Dealing with large dataset in healthcare requires it. Improving learning and teaching in education requires it. Observing the market trends in business requires it. Predicting financial profits requires it. It follows that as long as information exists, information retrieval is pervasive.

The methodologies of information retrieval are moving from traditional document engineering, static data, text, and images to dynamic multimedia audio video systems, from simple database management system to complex spatial datasets, from PC desktop to Smartphone technologies. It integrates emerging technologies to speed up the modernization of information retrieval.

Information retrieval involves fields beyond the word “retrieval” itself. Data search-ability, security, integrity, confidentiality, and accessibility are the same important as efficiency and accuracy. Advanced hardware improves retrieval efficiency and storage capacity for large datasets. Advanced software system services like web pages, emails, music, and videos add additional meanings and contents in today’s information retrieval. Service oriented framework such as cloud computing enhances the facilities and environments for the information retrieval but brings new challenges to the subject area. It follows that the topic involves in this book significantly fits the demanding in the world today.

Emerging technologies, such as XML technologies, has strong impact on information retrieval. XML related research on the data integrity, confidentiality, authentication and other security issues becomes an immediate task for the scientists and researchers in the subject area. XML related database mapping delivered a convenient tool for users who are familiar with traditional applications but willing to accept new challenges. XML related e-resources, e-reading, e-journals or other e-platforms are next moves in next generations’ information retrieval.

In the light of above, this book will open a conversation to the audience about what happened in today’s information retrieval.

## THE TARGET AUDIENCE

Immediate audiences for this book are from the area of information retrieval communities around world. The book targets at the readers from learner/trainers in educational institutes, industrial, commercial or non-commercial associations. Researchers in science and social science can be also benefited from the book. Professors, lecturers, and teachers from a wide range of subject areas can be benefited from the book if they are interested in information retrieval. The book can be an inspiration for research initiatives, and reading materials for educationists and students, and a library collection for this fast developing subject area with emerging cutting age technologies.

## THE IMPORTANCE OF EACH OF THE CHAPTER

This book is organized in 17 chapters with different topics in the subject area. The importance of each chapter are introduced as follows.

Chapter 1 is to introduce a popular search technology: XML keyword filter, especially used on XML Stream. The authors argued that keyword search technology in XML stream is user friendly in comparison with most existing XML stream processing systems that are difficult for ordinary users to apply. This chapter claims that the system XKFilter is innovative as it is the first system for supporting keyword search over XML stream. In XKFilter, the concepts of XLCA (eXclusive Lowest Common Ancestor) and XLCA Connecting Tree (XLC ACT) are used to define the search semantic and results of keywords, and present an approach to filter XML stream according to keywords. The prototype XKFilter is implemented in the experiments.

Chapter 2 describes a study based on XML related client server applications, particularly for the interactive web applications. The work discussed the key methods and technologies used in the research. Meanwhile, authors also discussed the state of the art technology, implementation for the developed system, and finally, contributions to the management of dynamic information management systems.

Chapter 3 addresses a topic related to the technology in document engineering in information retrieval research. The authors argued that keyword search did have inherent disadvantages. Traditional clustering techniques are inadequate in search efficiency for web documents. Finally the research proposes an alternative method based on the phrase based clustering algorithm. The research demonstrates that experimental results verify the method's feasibility and effectiveness.

Chapter 4 introduces a new query method that is based on clustering processes in which groups of semantically similar queries are detected. The clustering process uses the content of historical preferences of users registered in the query log of the search engine. This facility provides queries that are related to the ones submitted by users in order to direct them toward their required information. This method not only discovers the related queries but also ranks them according to a similarity measure. The method has been evaluated using real data sets from the search engine query log.

Chapter 5 presents XML documents normalization using GN-DTD. The research approached a fundamental problem in XML technology, i.e. designing well structured XML documents. The techniques of graphic notation for Document type definition are used in the investigation. Case study demonstrated the developed normal forms and normalization algorithm.

Chapter 6 provides mining Product Reviews in Web Forums. The research revealed information retrieval research on social impact. Opinion mining techniques have been used to analyze user-reviews. Finally the research has provided a recommendation to the products available on the web by analyzing the context to score the sentences for each review by identifying the opinion and feature words using a novel algorithm.

Chapter 7 talks about Schema Independent XML Compressor. XML has become the standard way for representing and transforming data over the World Wide Web. The problem with XML documents is that they have a very high ratio of redundancy, which makes these documents demanding a large storage capacity and large network band-width for transmission. This study designs a system for compressing and querying XML documents (XMLCQ) which compresses the XML document without the need to its schema or DTD to minimize the amount of technologies associated with these documents. XMLCQ first compressed the XML document by separating its data into containers according to the path of these data from the root to the leaf, and then it compressed these containers using a back-end compression technique. The compressed file then could be retrieved with any kind of queries applied. Only the required information is decompressed and submitted to the user. Depending on several experiments, the query processor part of the system showed the ability to answer different kinds of queries ranging from simple exact match queries to complex ones. Furthermore, this paper introduced the idea of retrieving information from more than one compressed XML documents.

Chapter 8 describes predictive modeling techniques in Determination of Algorithms Making Balance between Accuracy and Comprehensibility in Churn Prediction Setting. The research provides some detailed comparisons of rule based classifiers in churn prediction context. Key techniques in logistic regression (LR) and additive logistic regression (ALR) are used in the investigation. The research has developed eight distinctive algorithms, namely C4.5, C4.5 CP, RIPPER, RIPPER CP, PART, PART CP, LR, and ALR.

Chapter 9 presents a new model in Virtual Community of Practice Ontocop: Towards a New Model of Information Science Ontology (ISO). Information Science (IS) is an ambiguous field as its boundaries overlap with other domains such as Archive Science, Library Science and Computing Science which requires defined clear definition. This study creates a systematic and comprehensive ontology targeted to explore IS boundaries and foundations. This paper uses Mereotopology theory to describe classes, instances, and their relations. The classes are created based on taxonomy of IS to create an asserted model of Information Science Ontology (ISO) that can be as a skeletal foundation for knowledge base. The main classes are actors, method, practice, studies, mediate, kinds, domains, resources, legislation, philosophy & theories, societal, time, and space. The design is based on Methontology to create ISO from scratch. Its framework facilitates the construction of ontology at the knowledge levels. It is found that identifying the IS boundaries through implementation ontology workflow is encoded using Protégé and Web Ontology Language (OWL) for formalizing and representation of the ISO. ISO is an effective way to represent knowledge and overcome semantic heterogeneity, ISO is a fundamental integration between semantic that realizes the interoperability information of the domain.

Chapter 10 studied Improved Parameterless K-Means: Auto-Generation Centroids and Distance Data Point Clusters. The research presents a new approach and an improved method for effective and efficient clustering process evidenced by an improved version of K-means algorithm with auto-generate an initial number of clusters (k) and a new approach of defining initial Centroid for effective and efficient clustering process. The efficiency and effectiveness are analyzed, evaluated and discussed within the context.



Chapter 11 presents a research into the Ranking Tagged Resources Using Social Semantic Relevance. The research addresses the information retrieval research in web information environment. Ontology and web technology are discussed in the context. A questionnaire was designed to assess the crawled web pages for their graded relevance on a topic. Experiment studies have been conducted during investigation.

Chapter 12 introduces a Model of E-Reading Process for E-School Book in Libya. E-resources are factors that provide significant insights into actual reading behaviors and cognitive processes of readers. Two different samples of students, who study in Libyan primary schools, aged 9 to 12, were selected to investigate how students use and interact with both print and digital school books, identify the e-reading process, outline the aims of using the internet and technology, and define what students like and dislike in both versions. Furthermore, students found using the e-textbook to be more difficult than paper book and a significant difference is found in the reading process between paper books and electronic books. In addition, two reading strategies were used to read school book in both versions (electronic and paper): (1) view the text then answer the questions, or (2) view the questions than search for the correct answers.

Chapter 13 describes The Effect of Stemming on Arabic Text Classification: An Empirical Study. The research applies techniques of text classification for Arabic text documents. The results achieved an accuracy using the test modes up to 87.79% and 88.54%. Experiments and evaluation have been discussed in the chapter.

Chapter 14 presents a research in Electronic Resources Management in Jaykar Library. The research involves in the electronic document transformation through Internet. The research provides some suggestions on ICT enhanced management of electronic resources and speeding up the use of online journals among scientific department in the university.

Chapter 15 introduces a research in the MapReduce Based Information Retrieval Algorithms for Efficient Ranking of Webpages. The authors discuss the MapReduce implementation of crawler, indexer, and ranking algorithms in search engines. The algorithms are used in search engines to retrieve results from the World Wide Web. Tools like the crawler and an indexer in a MapReduce environment are used to improve the speed of crawling and indexing. Categorization is used to retrieve and order the results according to the user choice to personalize the search. A new score is introduced in this investigation. The experiments are conducted on Web graph datasets and the results are compared with the serial versions of crawler, indexer and ranking algorithms.

Chapter 16 presents a study in SAR: An Algorithm for Selecting a Partition Attribute in Categorical-Valued Information System Using Soft Set Theory. A soft-set based technique for decision making in categorical-valued information system is investigated, tested and evaluated. The results of this research will provide useful information for decision makers to handle categorical datasets.

Chapter 17 introduces a research in the area of XML and database: XRecursive: Connecting XML with Relational Databases. The research proposes an alternative method named *Xrecursive* for mapping XML (eXtensible Markup Language) documents to RDB (Relational Databases). The authors described the algorithms developed in the investigation and experiment results have been analyzed and evaluated.

## CONCLUSION

In conclusion, this book presents a general picture for the latest concepts and the state of the art technology in information retrieval research. In particular, it addresses that:

1. Information retrieval is moving from traditional concepts, theories, and technologies in document engineering, text stream and language translation into an advanced level in both hardware facilities and software systems.
2. Information retrieval is pervasively working with interdisciplinary and multidisciplinary applications. The driving force is the emerging technologies, such as XML that is machine readable and human readable language to enable the applications ubiquitously on the common ground.
3. New challenges in the modernization of information retrieval are discussed throughout this classical field.

*Zhongyu (Joan) Lu*  
*University of Huddersfield, UK*



# Table of Contents

|                      |    |
|----------------------|----|
| <b>Preface</b> ..... | xv |
|----------------------|----|

## **Chapter 1**

|  |   |
|--|---|
| XKFilter: A Keyword Filter on XML Stream .....           | 1 |
| <i>Weidong Yang, Fudan University, China</i>             |   |
| <i>Fei Fang, Fudan University, China</i>                 |   |
| <i>Nan Li, Fudan University, China</i>                   |   |
| <i>Zhongyu (Joan) Lu, University of Huddersfield, UK</i> |   |

## **Chapter 2**

|  |    |
|--|----|
| On the Design and Implementation of Interactive XML Applications ..... | 19 |
| <i>Jeff Brown, University of North Carolina Wilmington, USA</i>        |    |
| <i>Rebecca Brown, University of North Carolina Chapel Hill, USA</i>    |    |
| <i>Chris Velado, University of North Carolina Wilmington, USA</i>      |    |
| <i>Ron Vetter, University of North Carolina Wilmington, USA</i>        |    |

## **Chapter 3**

|   |    |
|---|----|
| A Roadmap to Integrate Document Clustering in Information Retrieval ..... | 31 |
| <i>R. Subhashini, Sathyabama University, India</i>                        |    |
| <i>V.Jawahar Senthil Kumar, Anna University, India</i>                    |    |

## **Chapter 4**

|  |    |
|--|----|
| Query Recommendation for Improving Search Engine Results ..... | 46 |
| <i>Hamada M. Zahera, Menoufiya University, Egypt</i>           |    |
| <i>Gamal F. El-Hady, Menoufiya University, Egypt</i>           |    |
| <i>W. F. Abd El-Wahed, Menoufiya University, Egypt</i>         |    |

## **Chapter 5**

|  |    |
|--|----|
| XML Documents Normalization Using GN-DTD .....   | 54 |
| <i>Zurinahni Zainol, University of Hull, UK, &amp; Universiti Sains Malaysia, Malaysia</i> |    |
| <i>Bing Wang, University of Hull, UK</i>   |    |

## **Chapter 6**

|  |    |
|--|----|
| Mining Product Reviews in Web Forums ..... | 78 |
|--|----|

*S. Hariharan, Pavendar Bharathidasan College of Engineering and Technology, India*

*T. Ramkumar, A.V.C. College of Engineering, India*

## **Chapter 7**

|   |    |
|---|----|
| Schema Independent XML Compressor ..... | 95 |
|---|----|

*Baydaa Al-Hamadani, University of Huddersfield, UK*

*Zhongyu (Joan) Lu, University of Huddersfield, UK*

*Raad F. Alwan, Philadelphia University, Jordan*

## **Chapter 8**

|  |     |
|--|-----|
| Determination of Algorithms Making Balance Between Accuracy and Comprehensibility in<br>Churn Prediction Setting ..... | 116 |
|--|-----|

*Hossein Abbasimehr, K. N. Toosi University of Tech, Iran*

*Mohammad Jafar Tarokh, K. N. Toosi University of Tech, Iran*

*Mostafa Setak, K. N. Toosi University of Tech, Iran*

## **Chapter 9**

|   |     |
|---|-----|
| Virtual Community of Practice Ontocop: Towards a New Model of Information Science<br>Ontology (ISO) ..... | 132 |
|---|-----|

*Ahlam Sawsaa, University of Huddersfield, UK*

*Zhongyu (Joan) Lu, University of Huddersfield, UK*

## **Chapter 10**

|   |     |
|---|-----|
| Improved Parameterless K-Means: Auto-Generation Centroids and Distance Data<br>Point Clusters ..... | 156 |
|---|-----|

*Wan Maseri Binti Wan Mohd, University Malaysia Pahang, Malaysia*

*A.H. Beg, University Malaysia Pahang, Malaysia*

*Tutut Herawan, University Malaysia Pahang, Malaysia*

*A. Noraziah, University Malaysia Pahang, Malaysia*

*K. F. Rabbi, University Malaysia Pahang, Malaysia*

## **Chapter 11**

|  |     |
|--|-----|
| Ranking Tagged Resources Using Social Semantic Relevance ..... | 169 |
|--|-----|

*Anjali Thukral, University of Delhi, India*

*Hema Banati, University of Delhi, India*

*Punam Bedi, University of Delhi, India*

## **Chapter 12**

|   |     |
|---|-----|
| Model of E-Reading Process for E-School Book in Libya ..... | 188 |
|---|-----|

*Azza Abubaker, University of Huddersfield, UK*

*Zhongyu (Joan) Lu, University of Huddersfield, UK*

### **Chapter 13**

|   |     |
|---|-----|
| The Effect of Stemming on Arabic Text Classification: An Empirical Study..... | 207 |
|---|-----|

*Abdullah Wahbeh, Dakota State University, USA*

*Mohammed Al-Kabi, Yarmouk University, Jordan*

*Qasem Al-Radaideh, Yarmouk University, Jordan*

*Emad Al-Shawakfa, Yarmouk University, Jordan*

*Izzat Alsmadi, Yarmouk University, Jordan*

### **Chapter 14**

|  |     |
|--|-----|
| Biometrics: Electronic Resources Management in Jaykar Library, University of Pune, India ..... | 226 |
|--|-----|

*Prakash Dongardive, University of Mekelle, Ethiopia*

*Neela Deshpande, University of Pune, India*

### **Chapter 15**

|  |     |
|--|-----|
| MapReduce Based Information Retrieval Algorithms for Efficient Ranking of Webpages ..... | 250 |
|--|-----|

*Srinivasa K.G., M.S. Ramaiah Institute of Technology, India*

*Anil Kumar Muppalla, M.S. Ramaiah Institute of Technology, India*

*Bharghava Varun A., M.S. Ramaiah Institute of Technology, India*

*Amulya M., M.S. Ramaiah Institute of Technology, India*

### **Chapter 16**

|   |  |
|---|--|
| SAR: An Algorithm for Selecting a Partition Attribute in Categorical-Valued Information |  |
|---|--|

|                                    |     |
|------------------------------------|-----|
| System Using Soft Set Theory ..... | 266 |
|------------------------------------|-----|

*Rabiei Mamat, Universiti Malaysia Terengganu, Malaysia*

*Tutut Herawan, Universiti Malaysia Pahang & Data Engineering Research Centre, Malaysia*

*Mustafa Mat Deris, Universiti Tun Hussein Onn Malaysia, Malaysia*

### **Chapter 17**

|   |     |
|---|-----|
| XRecursive: Connecting XML with Relational Databases..... | 281 |
|---|-----|

*Mohammed Adam Ibrahim Fakharaldien, Universiti Malaysia Pahang, Malaysia*

*Jasni Mohamed Zain, Universiti Malaysia Pahang, Malaysia*

*Norrozila Sulaiman, Universiti Malaysia Pahang, Malaysia*

*Tutut Herawan, Universiti Malaysia Pahang, Malaysia*

|                                 |     |
|---------------------------------|-----|
| Compilation of References ..... | 293 |
|---------------------------------|-----|

|                              |     |
|------------------------------|-----|
| About the Contributors ..... | 314 |
|------------------------------|-----|

|            |     |
|------------|-----|
| Index..... | 317 |
|------------|-----|

# Detailed Table of Contents

|                      |    |
|----------------------|----|
| <b>Preface</b> ..... | XV |
|----------------------|----|

## **Chapter 1**

|  |   |
|--|---|
| XKFilter: A Keyword Filter on XML Stream ..... | 1 |
|--|---|

*Weidong Yang, Fudan University, China*

*Fei Fang, Fudan University, China*

*Nan Li, Fudan University, China*

*Zhongyu (Joan) Lu, University of Huddersfield, UK*

Most existing XML stream processing systems adopt full structured query languages, such as XPath or XQuery, but they are difficult for ordinary users to learn and use. Keyword search is a user-friendly information discovery technique that has been extensively studied for text documents. This paper presents an XML stream filter system called XKFilter, which is the first system for supporting keyword search over XML stream. In XKFilter, the concepts of XLCA (eXclusive Lowest Common Ancestor) and XLCA Connecting Tree (XLCACT) are used to define the search semantic and results of keywords, and present an approach to filter XML stream according to keywords. The prototype XKFilter is implemented in the experiments.

## **Chapter 2**

|  |    |
|--|----|
| On the Design and Implementation of Interactive XML Applications ..... | 19 |
|--|----|

*Jeff Brown, University of North Carolina Wilmington, USA*

*Rebecca Brown, University of North Carolina Chapel Hill, USA*

*Chris Velado, University of North Carolina Wilmington, USA*

*Ron Vetter, University of North Carolina Wilmington, USA*

This paper describes issues and challenges in the design and implementation of interactive client-server applications where program logic is expressed in terms of an extensible markup language (XML) document. Although the technique was originally developed for creating interactive short message service (SMS) applications, it has expanded and is used for developing interactive web applications. XML-Interactive (or XML-I) defines the program states and corresponding actions. Because many interactive applications require sustained communication between the client and the underlying information service, XML-I has support for session management. This allows state information to be managed in a dynamic way. The paper describes several applications that are implemented using XML-I and discusses design issues. The software framework has been implemented in a Java environment.

### Chapter 3

|   |    |
|---|----|
| A Roadmap to Integrate Document Clustering in Information Retrieval ..... | 31 |
|---|----|

*R. Subhashini, Sathyabama University, India*

*V.Jawahar Senthil Kumar, Anna University, India*

The World Wide Web is a large distributed digital information space. The ability to search and retrieve information from the Web efficiently and effectively is an enabling technology for realizing its full potential. Information Retrieval (IR) plays an important role in search engines. Today's most advanced engines use the keyword-based ("bag of words") paradigm, which has inherent disadvantages. Organizing web search results into clusters facilitates the user's quick browsing of search results. Traditional clustering techniques are inadequate because they do not generate clusters with highly readable names. This paper proposes an approach for web search results in clustering based on a phrase based clustering algorithm. It is an alternative to a single ordered result of search engines. This approach presents a list of clusters to the user. Experimental results verify the method's feasibility and effectiveness.

### Chapter 4

|  |    |
|--|----|
| Query Recommendation for Improving Search Engine Results ..... | 46 |
|--|----|

*Hamada M. Zahera, Menoufiya University, Egypt*

*Gamal F. El-Hady, Menoufiya University, Egypt*

*W. F. Abd El-Wahed, Menoufiya University, Egypt*

As web contents grow, the importance of search engines become more critical and at the same time user satisfaction decreases. Query recommendation is a new approach to improve search results in web. In this paper a method is proposed that, given a query submitted to a search engine, suggests a list of queries that are related to the user input query. The related queries are based on previously issued queries, and can be issued by the user to the search engine to tune or redirect the search process. The proposed method is based on clustering processes in which groups of semantically similar queries are detected. The clustering process uses the content of historical preferences of users registered in the query log of the search engine. This facility provides queries that are related to the ones submitted by users in order to direct them toward their required information. This method not only discovers the related queries but also ranks them according to a similarity measure. The method has been evaluated using real data sets from the search engine query log.

### Chapter 5

|  |    |
|--|----|
| XML Documents Normalization Using GN-DTD ..... | 54 |
|--|----|

*Zurinahmi Zainol, University of Hull, UK, & Universiti Sains Malaysia, Malaysia*

*Bing Wang, University of Hull, UK*

Designing a well-structured XML document is important for the sake of readability, maintainability and more importantly to avoid both data redundancies and update anomalies. This paper proposes to improve and simplify XML structural design using a normalization process. To achieve this, Graphical Notation for Document Type Definition (GN-DTD) is used to describe the structure of XML document at the schema level. Multiple levels of normal forms for GN-DTD are proposed and the corresponding normalization rules to transform from poorly designed into well-designed XML documents. A case study is presented to show the application of these normal forms and normalization algorithm.



## Chapter 6

|  |    |
|--|----|
| Mining Product Reviews in Web Forums ..... | 78 |
|--|----|

*S. Hariharan, Pavendar Bharathidasan College of Engineering and Technology, India*

*T. Ramkumar, A.V.C. College of Engineering, India*

Internet has brought a major drift in user community. Apart from its well-known usage, it also promotes social networking. Research on such social networking has advanced significantly in recent years which have been highly influenced by the online social websites. People perceive the web as a social medium that allows larger interaction among people, sharing of knowledge, or experiences. Internet or social web forums act as an agent to reproduce some general information that would benefit the users. A product review by the user is a more accurate representation of its real-world performance and web-forums are generally used to post such reviews. Though commercial review websites allow users to express their opinions in whatever way they feel, the number of reviews that a product receives could be very high. Hence, opinion mining techniques can be used to analyze the user-reviews, classify the content as positive or negative, and thereby find out how the product fares. This paper focuses its attention on providing a recommendation to the products available on the web by analyzing the context to score the sentences for each review by identifying the opinion and feature words using a novel algorithm.

## Chapter 7

|   |    |
|---|----|
| Schema Independent XML Compressor ..... | 95 |
|---|----|

*Baydaa Al-Hamadani, University of Huddersfield, UK*

*Zhongyu (Joan) Lu, University of Huddersfield, UK*

*Raad F. Alwan, Philadelphia University, Jordan*

XML has become the standard way for representing and transforming data over the World Wide Web. The problem with XML documents is that they have a very high ratio of redundancy, which makes these documents demanding a large storage capacity and large network band-width for transmission. This study designs a system for compressing and querying XML documents (XMLCQ) which compresses the XML document without the need to its schema or DTD to minimize the amount of technologies associated with these documents. XMLCQ first compressed the XML document by separating its data into containers according to the path of these data from the root to the leaf, then it compressed these containers using a back-end compression technique. The compressed file then could be retrieved with any kind of queries applied. Only the required information is decompressed and submitted to the user. Depending on several experiments, the query processor part of the system showed the ability to answer different kinds of queries ranging from simple exact match queries to complex ones. Furthermore, this paper introduced the idea of retrieving information from more than one compressed XML documents.

## Chapter 8

|   |     |
|---|-----|
| Determination of Algorithms Making Balance Between Accuracy and Comprehensibility in Churn Prediction Setting ..... | 116 |
|---|-----|

*Hossein Abbasimehr, K. N. Toosi University of Tech, Iran*

*Mohammad Jafar Tarokh, K. N. Toosi University of Tech, Iran*

*Mostafa Setak, K. N. Toosi University of Tech, Iran*

Predictive modeling is a useful tool for identifying customers who are at risk of churn. An appropriate churn prediction model should be both accurate and comprehensible. However, reviewing the past researches in this context shows that much attention is paid to accuracy of churn prediction models than comprehensibility of them. This paper compares three different rule induction techniques from three