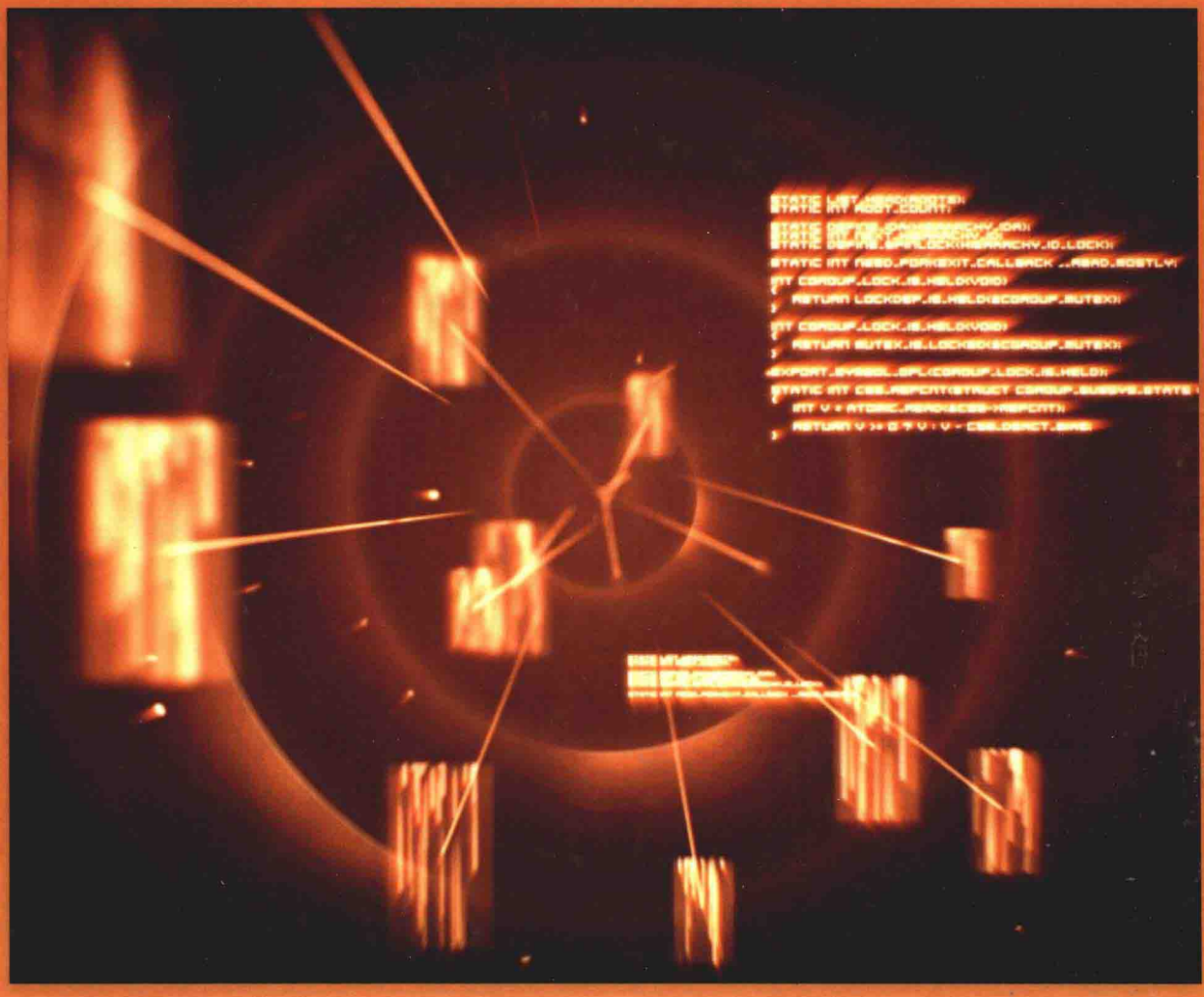
PREMIER REFERENCE SOURCE

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
Web site: http://www.igi-global.com

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.

Associate Editors

List of Reviewers

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 (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 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 Mereotopolgy 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

Detailed Table of Contents