MINING THE WORLD WIDE WEB
An Information Search Approach

THE KLUWER INTERNATIONAL SERIES ON INFORMATION RETRIEVAL

Series Editor

W. Bruce Croft
University of Massachusetts, Amherst

Also in the Series:

MULTIMEDIA INFORMATION RETRIEVAL: ***Content-Based Information Retrieval from Large Text and Audio Databases,*** *by Peter Schäuble;* ISBN: 0-7923-9899-8

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** *by Gerald Kowalski;* ISBN: 0-7923-9926-9

CROSS-LANGUAGE INFORMATION RETRIEVAL, *edited by Gregory Grefenstette;* ISBN: 0-7923-8122-X

TEXT RETRIEVAL AND FILTERING: ***Analytic Models of Performance,*** *by Robert M. Losee;* ISBN: 0-7923-8177-7

INFORMATION RETRIEVAL: UNCERTAINTY AND LOGICS: ***Advanced Models for the Representation and Retrieval of Information,*** *by Fabio Crestani, Mounia Lalmas, and Cornelis Joost van Rijsbergen;* ISBN: 0-7923-8302-8

DOCUMENT COMPUTING: ***Technologies for Managing Electronic Document Collections,*** *by Ross Wilkinson, Timothy Arnold-Moore, Michael Fuller, Ron Sacks-Davis, James Thom, and Justin Zobel;* ISBN: 0-7923-8357-5

AUTOMATIC INDEXING AND ABSTRACTING OF DOCUMENT TEXTS, *by Marie-Francine Moens;* ISBN 0-7923-7793-1

ADVANCES IN INFORMATIONAL RETRIEVAL: ***Recent Research from the Center for Intelligent Information Retrieval,*** *by W. Bruce Croft;* ISBN 0-7923-7812-1

INFORMATION RETRIEVAL SYSTEMS: ***Theory and Implementation,*** **Second Edition,** *by Gerald J. Kowalski and Mark T. Maybury;* ISBN: 0-7923-7924-1

PERSPECTIVES ON CONTENT-BASED MULTIMEDIA SYSTEMS, *by Jian Kang Wu; Mohan S. Kankanhalli;Joo-Hwee Lim;Dezhong Hong;* ISBN: 0-7923-7944-6

MINING THE WORLD WIDE WEB: ***An Information Search Approach,*** *by George Chang, Marcus J. Healey, James A. M. McHugh, Jason T. L. Wang;* ISBN: 0-7923-7349-9

INTEGRATED REGION-BASED IMAGE RETRIEVAL, *by James Z. Wang;* ISBN: 0-7923-7350-2

MINING THE WORLD WIDE WEB
An Information Search Approach

by

George Chang
Kean University, Union, NJ

Marcus J. Healey
Mobilocity, New York, NY

James A. M. McHugh
New Jersey Institute of Technology, Newark, NJ

Jason T. L. Wang
New Jersey Institute of Technology, Newark, NJ

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

Mining the World Wide Web : an information search approach / George Chang ... [et al.].
p. cm. – (The Kluwer international series on information retrieval ; 10)
Includes bibliographical references and index.
ISBN 978-1-4613-5654-7 ISBN 978-1-4615-1639-2 (eBook)
DOI 10.1007/978-1-4615-1639-2
1. Data mining. 2. Web databases. 3. World Wide Web. I. Change, George. II. Series.

QA76.9.D343 M56 2001
006.3—dc21

Copyright © 2001 Springer Science+Business Media New York, Second Printing 2002.
Originally published by Kluwer Academic Publishers in 2001
Softcover reprint of the hardcover 1st edition 2001
All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, mechanical, photo-copying, recording, or otherwise, without the prior written permission of the publisher, Springer Science+Business Media, LLC

Printed on acid-free paper.

The Publisher offers discounts on this book for course use and bulk purchases. For further information, send email to <scott.delman@wkap.com>.

This printing is a digital duplication of the original edition.

This book is dedicated
to our teachers,
especially our parents,
for their
encouragement,
inspiration and love.

Contents

List of Figures ix
List of Tables xi
Preface xiii
Acknowledgments xvii

Part I Information Retrieval on the Web

1. KEYWORD-BASED SEARCH ENGINES 3
   1. Search Engines 4
   2. Web Directories 15
   3. Meta-Search Engines 16
   4. Information Filtering 17
2. QUERY-BASED SEARCH SYSTEMS 19
   1. W3QS/W3QL 20
   2. WebSQL 26
   3. WAQL 29
3. MEDIATORS AND WRAPPERS 35
   1. LORE 38
   2. ARANEUS 41
   3. AKIRA 47
4. MULTIMEDIA SEARCH ENGINES 51
   1. Text or Keyword-Based Search 53
   2. Content-Based Search 59

Part II Data Mining on the Web

5. DATA MINING 67
   1. What is Data Mining? 67
   2. On-Line Analytical Processing 68
   3. Pattern Extraction Processes 69

6. TEXT MINING 81
   1. What is Text Mining? 81
   2. Association Discovery 82
   3. Trend Discovery 84
   4. Event Detection 87

7. WEB MINING 93
   1. What is Web Mining? 93
   2. Web Usage Mining 95
   3. Web Structure Mining 100

8. WEB CRAWLING AGENTS 105
   1. What is a Web Crawling Agent? 105
   2. Web Crawling Architecture 107
   3. Crawling Algorithms 110
   4. Topic-Oriented Web Crawling 114

Part III A Case Study in Environmental Engineering

9. ENVIRODAEMON 119
   1. Background 119
   2. EnviroDaemon (ED) 121
   3. ED with Hierarchical Search 130
   4. A Hierarchical Query Language 134
   5. Summary 136

References 137

Index 161

List of Figures

1.1	The architecture of a search engine.	6
1.2	String edit operations.	10
1.3	An inverted file built based on a B-tree.	11
1.4	Index structure of a Web search engine.	13
1.5	Recall-precision curve.	15
1.6	The architecture of a meta-search engine.	17
1.7	The building process of a topic-specific search engine.	18
2.1	Web querying system architecture.	21
2.2	Structure-based query.	22
2.3	Example of a pattern graph.	24
3.1	A client/server architecture.	36
3.2	A data warehouse architecture.	36
3.3	A mediator architecture.	37
3.4	An OEM database.	39
3.5	Lore architecture.	41
3.6	ARANEUS data transformation process.	42
3.7	AuthorPage scheme for ARANEUS.	43
3.8	A unique page-scheme.	44
3.9	Relational view on VLDB papers.	45
3.10	Page-schemes to organize papers by year.	46
3.11	HTML page generating schemes.	46
3.12	A `Fragment` class.	47
3.13	A segment of HTML.	48

3.14	Concept classes for `Person` and `MP3`.	49
3.15	AKIRA system architecture.	50
4.1	Keyword-image inversion index.	55
4.2	Object extraction from an image.	61
4.3	Shape templates for clustering seeds.	61
4.4	Color histogram.	62
5.1	A 3-D data cube and OLAP operations.	70
5.2	Concept hierarchy.	72
5.3	Visualization methods for generalized data.	73
5.4	A decision tree for the concept "online shopping".	76
6.1	Algorithm for association rules generation.	84
6.2	An uptrend shape query.	87
6.3	Temporal histograms of two fictitious news events.	89
7.1	Hyperlink relationships.	95
7.2	Hubs and authorities.	101
7.3	Root-set and base set.	103
8.1	A Web crawling architecture.	108
8.2	A simple Retrieving Module.	109
8.3	Illustration of front pages.	111
9.1	Front-end interface for EnviroDaemon.	123
9.2	Front-end interface for EnviroDaemon with HIST.	132
9.3	HIST system architecture.	133
9.4	Converting a hypertext document to a labeled tree using DTD.	134
9.5	Hierarchical query & HTML tree.	135

List of Tables

1.1	Searchable Web page fields.	7
1.2	Granularity of inverted file indexes.	12
1.3	Yahoo! directory categories.	15
2.1	Operators specifying the traversal method on a specific link type.	31
3.1	A list of HTML fragments.	48
4.1	Media types and file extensions.	52
4.2	Categories of Yahoo! Image Surfer.	59

Preface

The World Wide Web (Web), which emerged in the early 1990s, has made great strides in the late 1990s. Its explosive growth is expected to continue into the next millennium. The contributing factors to this explosive growth include the widespread use of microcomputers, advances in hardware (microprocessors, memory and storage) technologies, increased ease of use in computer software packages, and most importantly – tremendous opportunities the Web offers for all businesses. The consequence of the popularity of the Web as a global information system is that it has flooded us with a large amount of data and information. In this sea of data and information, searching for a piece of information is like finding a needle in a haystack. Finding useful information on the Web is often a tedious and frustrating experience. Therefore, new tools and techniques are needed to assist us in intelligently searching for and discovering useful information on the Web.

This book explores the concepts and techniques of *Web mining*, a promising and rapidly growing field of computer science research with great potential in e-business. Web mining is a multidisciplinary field, drawing on such areas as artificial intelligence, databases, data mining, data warehousing, data visualization, information retrieval, machine learning, markup languages, pattern recognition, statistics, and Web technology. Moreover, depending on the type of data and approach used, techniques from other disciplines may be applicable.

We present the Web mining material in this book from an *information search* perspective. In this sense, we focus on issues relating to the efficiency, feasibility, scalability and usability of searching techniques for Web mining. As a

result, this book is not intended as an introduction to databases, data mining or information retrieval, etc. However, we do provide enough background information to facilitate the reader's comprehension in each area related to our book.

The first part of the book focuses on information retrieval (IR) on the Web. IR systems deal with the automated storage, retrieval, organization, and representation of documents. It is widely used in libraries and government agencies where a large amount of document storage and retrieval is necessary. Research on IR includes categorization, classification, filtering, modeling, query language, system architecture, user interface, etc.

IR mainly deals with natural language text that is often either not structured or semistructured in nature. Consequently, the semantics of the text may be ambiguous. In contrast, data retrieval tools in a database management system (DBMS) deal with structured data that is semantically well defined. As long as a record satisfies a user's query defined by a regular expression or relational algebra, it will be retrieved by the DBMS.

Although DBMSs provide a satisfactory solution to data retrieval, they do not provide users with search by topic or by subject. IR systems are an attempt to answer the shortcomings of data retrieval systems. Their main goal is the retrieval of documents relevant to a specific topic or subject according to the user's information need. Thus, the notion of relevance is at the heart of an IR system. Since IR systems must interpret the user's query, retrieving as few irrelevant documents as possible is also their primary goal.

We dedicate the first part of the book to IR techniques because of their importance in Web mining. Searching is fundamental to mining useful information and knowledge in any media type. Having the ability to find relevant documents on the Web is an essential process in Web mining. The cleaner the data set, the better the information and knowledge that can be extracted from it. In this part, keyword-based and multimedia search engines, query-based search systems, and mediators and wrappers are discussed.

The second part of the book reviews data mining (DM) on the Web. DM, which emerged during the late 1980s, has made a great impact in both academia and industry in the 1990s. Commercial products have appeared as a result of many years of research prototype development.

In reality, DM is one of the steps in the evolution of information technology. The main reason that it has gained attention in industry and scientific research is because of the vast amount of raw data that has been accumulated over the

past several decades. Putting these huge quantities of data to good use is the objective of data mining.

DM, sometimes called *knowledge discovery in databases* (KDD), refers to the extraction of knowledge from large amounts of data. DM has many stages including data cleaning, data integration, data filtering, data transformation, pattern recognition, pattern analysis, and knowledge presentation. Traditionally, DM techniques have been applied to data warehouses, relational databases, and transactional databases, where data are well structured and semantically well defined. In principle, DM techniques can be applied to any kind of information repository.

The Web is a distributed repository linked by millions of hyperlinks embedded in hypertext documents called HTML. This large repository of information provides many opportunities for DM. Unlike relational databases with structured, well defined semantics, the Web is semistructured in nature. Therefore, in order to apply data mining to the Web, previous relational data mining techniques require modification or new techniques must be invented.

In the second part of this book, basic concepts on data mining are first introduced. Topics on text mining, Web data mining, and Web crawling are then discussed, respectively.

The last part of the book is a case study. The case study focuses on a domain specific search engine prototype called EnviroDaemon. This search engine uses tools freely available on the Web to gather environmental science related information on the Web.

Web mining, which emerged in the late 1990s, is a cross-disciplinary field. This book is an effort to bridge the gap between information search and data mining on the Web. The book provides enough background information on both fields by presenting the intuition and mechanics of the best tools in each field, as well as how they work together. We hope that this book will encourage people with different backgrounds to contribute to Web mining.

The book is targeted towards researchers and developers of Web information systems. It can serve as a supplemental text book to a one-semester undergraduate/graduate course on data mining, databases and information retrieval.

Acknowledgments

We would like to express our sincere gratitude to all those who have worked with us in all aspects related to this book. Those include Francine Abeles, Jeff Cheng, Eunice W. Healey, Stanley H. Lipson, Amit Revankar, Gunjan Samtani, and Jeanie Hsiang for the cover design. We also wish to thank Bruce Croft, Series Editor of the Information Retrieval series, Scott Delman, Senior Publishing Editor, and Melissa Fearon, Editorial Assistant of Kluwer Academic Publishers, for their guiding us through the materialization and publication of this book. G. Chang's work was partially supported by Kean University's Untenured Faculty Research Grants. J. McHugh's work was supported by New Jersey Information-Technology Opportunities for the Workforce, Education, and Research (NJ I-TOWER) grant, a project funded by the NJ Commission on Higher Education (contract #: 01-801020-02).