DATA MINING THE WEB

Uncovering Patterns in Web Content, Structure, and Usage

ZDRAVKO MARKOV AND DANIEL T. LAROSE

Central Connecticut State University
New Britain, CT

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

DATA MINING
THE WEB

The Wiley Bicentennial–Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce

William J. Pesce
President and Chief Executive Officer

Peter Booth Wiley

Peter Booth Wiley
Chairman of the Board

DATA MINING THE WEB

Uncovering Patterns in Web Content, Structure, and Usage

ZDRAVKO MARKOV AND DANIEL T. LAROSE

Central Connecticut State University
New Britain, CT

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Markov, Zdravko, 1956–
Data-mining the Web : uncovering patterns in Web content, structure, and usage /
by Zdravko, Markov & Daniel T. Larose.
p. cm.
Includes index.
978-0-471-66655-4 (cloth)
1. Data mining. 2. Web databases. I. Larose, Daniel T. II. Title.
QA76.9.D343M38 2007
005.74 – dc22

2006025099

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

For my children
Teodora, Kalin, and Svetoslav
– Z.M.

For my children
Chantal, Ellyriane, Tristan, and Ravel
– D.T.L.

CONTENTS

PREFACE xi

PART I

WEB STRUCTURE MINING

1 *INFORMATION RETRIEVAL AND WEB SEARCH* **3**

Web Challenges 3
Web Search Engines 4
Topic Directories 5
Semantic Web 5
Crawling the Web 6
Web Basics 6
Web Crawlers 7
Indexing and Keyword Search 13
Document Representation 15
Implementation Considerations 19
Relevance Ranking 20
Advanced Text Search 28
Using the HTML Structure in Keyword Search 30
Evaluating Search Quality 32
Similarity Search 36
Cosine Similarity 36
Jaccard Similarity 38
Document Resemblance 41
References 43
Exercises 43

2 *HYPERLINK-BASED RANKING* **47**

Introduction 47
Social Networks Analysis 48
PageRank 50
Authorities and Hubs 53
Link-Based Similarity Search 55
Enhanced Techniques for Page Ranking 56
References 57
Exercises 57

PART II
WEB CONTENT MINING

3 *CLUSTERING* **61**
Introduction 61
Hierarchical Agglomerative Clustering 63
k-Means Clustering 69
Probabilty-Based Clustering 73
Finite Mixture Problem 74
Classification Problem 76
Clustering Problem 78
Collaborative Filtering (Recommender Systems) 84
References 86
Exercises 86

4 *EVALUATING CLUSTERING* **89**
Approaches to Evaluating Clustering 89
Similarity-Based Criterion Functions 90
Probabilistic Criterion Functions 95
MDL-Based Model and Feature Evaluation 100
Minimum Description Length Principle 101
MDL-Based Model Evaluation 102
Feature Selection 105
Classes-to-Clusters Evaluation 106
Precision, Recall, and *F*-Measure 108
Entropy 111
References 112
Exercises 112

5 *CLASSIFICATION* **115**
General Setting and Evaluation Techniques 115
Nearest-Neighbor Algorithm 118
Feature Selection 121
Naive Bayes Algorithm 125
Numerical Approaches 131
Relational Learning 133
References 137
Exercises 138

PART III
WEB USAGE MINING

6 *INTRODUCTION TO WEB USAGE MINING* **143**
Definition of Web Usage Mining 143
Cross-Industry Standard Process for Data Mining 144
Clickstream Analysis 147

Web Server Log Files 148
Remote Host Field 149
Date/Time Field 149
HTTP Request Field 149
Status Code Field 150
Transfer Volume (Bytes) Field 151
Common Log Format 151
Identification Field 151
Authuser Field 151
Extended Common Log Format 151
Referrer Field 152
User Agent Field 152
Example of a Web Log Record 152
Microsoft IIS Log Format 153
Auxiliary Information 154
References 154
Exercises 154

7 *PREPROCESSING FOR WEB USAGE MINING* 156

Need for Preprocessing the Data 156
Data Cleaning and Filtering 158
Page Extension Exploration and Filtering 161
De-Spidering the Web Log File 163
User Identification 164
Session Identification 167
Path Completion 170
Directories and the Basket Transformation 171
Further Data Preprocessing Steps 174
References 174
Exercises 174

8 *EXPLORATORY DATA ANALYSIS FOR WEB USAGE MINING* 177

Introduction 177
Number of Visit Actions 177
Session Duration 178
Relationship between Visit Actions and Session Duration 181
Average Time per Page 183
Duration for Individual Pages 185
References 188
Exercises 188

9 *MODELING FOR WEB USAGE MINING: CLUSTERING, ASSOCIATION, AND CLASSIFICATION* 191

Introduction 191
Modeling Methodology 192
Definition of Clustering 193
The BIRCH Clustering Algorithm 194
Affinity Analysis and the A Priori Algorithm 197

Discretizing the Numerical Variables: Binning 199
Applying the A Priori Algorithm to the CCSU Web Log Data 201
Classification and Regression Trees 204
The C4.5 Algorithm 208
References 210
Exercises 211

INDEX **213**

PREFACE

DEFINING DATA MINING THE WEB

By *data mining the Web*, we refer to the application of data mining methodologies, techniques, and models to the variety of data forms, structures, and usage patterns that comprise the World Wide Web. As the subtitle indicates, we are interested in uncovering patterns and trends in the content, structure, and use of the Web. A good definition of data mining is that in *Principles of Data Mining* by David Hand, Heikki Mannila, and Padhraic Smyth (MIT Press, Cambridge, MA, 2001): "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner." *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage* demonstrates how to apply data mining methods and models to Web-based data forms.

THE DATA MINING BOOK SERIES

This book represents the third volume in a data mining book series. The first volume in this series, *Discovering Knowledge in Data: An Introduction to Data Mining*, by Daniel Larose, appeared in 2005, and introduced the reader to this rapidly growing field of data mining. The second volume in the series, *Data Mining Methods and Models*, by Daniel Larose, appeared in 2006, and explores the process of data mining from the point of view of model building—the development of complex and powerful predictive models that can deliver actionable results for a wide range of business and research problems. Although *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage* serves well as a stand-alone resource for learning how to apply data mining techniques to Web-based data, reference is sometimes made to more complete coverage of certain topics in the earlier volumes.

HOW THE BOOK IS STRUCTURED

The book is presented in three parts.

Part I: Web Structure Mining

In Part I we discuss basic ideas and techniques for extracting text information from the Web, including collecting and indexing web documents and searching and ranking

web pages by their textual content and hyperlink structure. Part I contains two chapters, Chapter 1, *Information Retrieval and Web Search*; and Chapter 2, *Hyperlink-Based Ranking.*

Part II: Web Content Mining

Machine learning and data mining approaches organize the Web by content and thus respond directly to the major challenge of *turning web data into web knowledge.* In Part II we focus on two approaches to organizing the Web, clustering and classification. Part II consists of three chapters: Chapter 3, *Clustering*; Chapter 4, *Evaluating Clustering*; and Chapter 5, *Classification.*

Part III: Web Usage Mining

Web usage mining refers to the application of data mining methods for uncovering usage patterns from Web data. Web usage mining differs from web structure mining and web content mining in that web usage mining reflects the behavior of humans as they interact with the Internet. Part III consists of four chapters: Chapters 6, *Introduction to Web Usage Mining*; Chapter 7, *Preprocessing for Web Usage Mining*; Chapter 8, *Exploratory Data Analysis for Web Usage Mining*; and Chapter 9, *Modeling for Web Usage Mining: Clustering, Association, and Classification.*

WHY THE BOOK IS NEEDED

The book provides the reader with:

- The models and techniques to uncover hidden nuggets of information in Web-based data
- Insight into how web mining algorithms really work
- The experience of actually performing web mining on real-world data sets

"WHITE-BOX" APPROACH: UNDERSTANDING THE UNDERLYING ALGORITHMIC AND MODEL STRUCTURES

The best way to avoid costly errors stemming from a blind black-box approach to data mining, is to apply, instead, a white-box methodology, which emphasizes an understanding of the algorithmic and statistical model structures underlying the software. The book, applies this white-box approach by:

- Walking the reader through various algorithms
- Providing examples of the operation of web mining algorithms on actual large data sets

- Testing the reader's level of understanding of the concepts and algorithms
- Providing an opportunity for the reader to do some real web mining on large Web-based data sets

Algorithm Walk-Throughs

The book walks the reader through the operations and nuances of various algorithms, using small sample data sets, so that the reader gets a true appreciation of what is really going on inside an algorithm. For example, in Chapter 1, we demonstrate the nuts and bolts of relevance ranking, similarity searching, and other topics, using a particular small web data set. The reader can perform the same analysis in parallel, and therefore understanding is enhanced.

Applications of Algorithms and Models to Large Data Sets

The book provides examples of the application of the various algorithms and models on actual large data sets. For example, in Chapter 7 data cleaning, de-spidering, session identification, and other tasks are carried out on two real-world large web log databases, from the Web sites for NASA and Central Connecticut State University. All data sets used throughout the book are available for free download from the book series Web site, `www.dataminingconsultant.com`.

Chapter Exercises: Checking to Make Sure That You Understand It

The book includes over 100 chapter exercises, which allow readers to assess their depth of understanding of the material, as well as to have a little fun playing with numbers and data. These include exercises designed to (1) clarify some of the more challenging concepts in data mining, and (2) challenge the reader to apply the particular data mining algorithm to a small data set and, step by step, to arrive at a computationally sound solution. For example, in Chapter 4 readers are asked to run a series of experiments comparing the efficacy of a variety of clustering algorithms applied to the "Top 100 Websites" data set.

Hands-on Analysis: Learn Data Mining by Doing Data Mining

Nearly every chapter provides the reader with *hands-on analysis problems*, representing an opportunity for the reader to apply his or her newly acquired data mining expertise to solving real problems using large data sets. Many people learn by doing. The book provides a framework by which the reader can learn data mining by doing data mining. For example, in Chapter 8 readers are challenged to provide detailed reports and summaries for real-world web log data. The 34 tasks include finding the average time per page view, constructing a table of the most popular directories, and so on.

DATA MINING AS A PROCESS

The book continues the coverage of data mining as a process. The particular standard process used is the CRISP-DM framework: the cross-industry standard process for data mining. CRISP-DM demands that data mining be seen as an entire process, from communication of the business problem through data collection and management, data preprocessing, model building, model evaluation, and finally, model deployment. Therefore, this book is not only for analysts and managers, but also for data management professionals, database analysts, decision makers, and others who would like to leverage their repositories of Web-based data.

THE SOFTWARE

The software used in this book includes the following:

- WEKA open-source data mining software
- Clementine data mining software suite.

The Weka (Waikato Environment for Knowledge Analysis) machine learning workbench is open-source software issued under the GNU General Public License, which includes a collection of tools for completing many data mining tasks. The book uses Weka throughout Parts I and II. For more information regarding Weka, see `http://www.cs.waikato.ac.nz/~ml/`. Clementine (`http://www.spss.com/clementine/`) is one of the most widely used data mining software suites and is distributed by SPSS. Clementine is used throughout Part III.

THE COMPANION WEB SITE: www.dataminingconsultant.com

The reader will find supporting materials for both this book and the other data mining books in this series at the companion Web site, `www.dataminingconsultant.com`. There one may download the many data sets used in the book, so that the reader may develop a hands-on feeling for the analytic methods and models encountered throughout the book. Errata are also available, as is a comprehensive set of data mining resources, including links to data sets, data mining groups, and research papers.

The real power of the companion Web site is available to faculty adopters of the textbook, who will have access to the following resources:

- Solutions to all the exercises, including hands-on analyses
- Powerpoint presentations of each chapter, ready for deployment in the classroom

- Sample data mining course projects, written by the authors for use in their own courses, and ready to be adapted for your course
- Real-world data sets, to be used with the course projects.
- Multiple-choice chapter quizzes
- Chapter-by-chapter web resources

DATA MINING THE WEB AS A TEXTBOOK

The book naturally fits the role of a textbook for an introductory course in web mining. Instructors may appreciate:

- The "white-box" approach, emphasizing an understanding of the underlying algorithmic structures
  - Algorithm walk-throughs
  - Application of the algorithms to large data sets
  - Chapter exercises
  - Hands-on analysis
- The logical presentation, flowing naturally from the CRISP-DM standard process and the set of web mining tasks
- The companion Web site, providing the array of resources for adopters detailed above

The book is appropriate for advanced undergraduate or graduate-level courses. An introductory statistics course would be nice, but is not required. No prior computer programming or database expertise is required.

ACKNOWLEDGMENTS

The material for web content and structure mining is based on the web mining course that I developed and taught for the graduate CIT program at Central Connecticut State University. The student projects and some exercises from this course were then used in the artificial intelligence course that I taught for the CS program at the same school. Some material from my data mining and machine learning courses taught for the data mining program at CCSU is also included. I am grateful to my students from all these courses for their inspirational enthusiasm and valuable feedback. The book was written while I was on sabbatical leave, spent in my home country, Bulgaria, sharing my time between family and writing. I wish to thank my children, Teodora and Kalin, and my wife, Irena, for their patience and understanding during that time.

Zdravko Markov, Ph.D.
Department of Computer Science
Central Connecticut State University
`www.cs.ccsu.edu/~markov/`

I would like to thank all the folks at Wiley, especially editor Paul Petralia, for their guidance and support. Je suis également reconnaissant à ma rédactrice et amie Val Moliere, qui a insisté pour que cette série de livres devienne réalité. I also wish to thank Dr. Chun Jin, Dr. Daniel S. Miller, Dr. Roger Bilisoly, Dr. Darius Dziuda, and Dr. Krishna Saha, my colleagues in the Master of Science in data mining program at Central Connecticut State University, Dr. Timothy Craine, Chair of the Department of Mathematical Sciences at CCSU, Dr. Dipak K. Dey, Chair of the Department of Statistics at the University of Connecticut, and Dr. John Judge, Chair of the Department of Mathematics at Westfield State College. Thanks to my daughter, Chantal, for her precious love and gentle insanity. Thanks to my twin children, Tristan and Ravel, for sharing the computer and for sharing their true perspective. Above all, I extend my deepest gratitude to my darling wife, Debra J. Larose, for her support, understanding, and love. "Say you'll share with me one love, one lifetime. . . ."

Daniel T. Larose, Ph.D.
Professor of Statistics
Director, Data Mining @CCSU
Department of Mathematical Sciences
Central Connecticut State University
www.math.ccsu.edu/larose

PART I

WEB STRUCTURE MINING

In the first part of the book we discuss basic ideas and techniques for extracting text information from the Web, including collecting and indexing web documents and searching and ranking web pages by their textual content and hyperlink structure. We first discuss the motivation to organize the web content and find better ways for web search to make the vast knowledge on the Web easily accessible. Then we describe briefly the basics of the Web and explore the approaches taken by web search engines to retrieve web pages by keyword search. To do this we look into the technology for text analysis and search developed earlier in the area of information retrieval and extended recently with ranking methods based on web hyperlink structure.

All that may be seen as a preprocessing step in the overall process of data mining the web content, which provides the input to machine learning methods for extracting knowledge from hypertext data, discussed in the second part of the book.

Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage
By Zdravko Markov and Daniel T. Larose

CHAPTER 1

INFORMATION RETRIEVAL AND WEB SEARCH

WEB CHALLENGES

CRAWLING THE WEB

INDEXING AND KEYWORD SEARCH

EVALUATING SEARCH QUALITY

SIMILARITY SEARCH

WEB CHALLENGES

As originally proposed by Tim Berners-Lee [1], the Web was intended to improve the management of general information about accelerators and experiments at CERN. His suggestion was to organize the information used at that institution in a graphlike structure where the nodes are documents describing objects, such as notes, articles, departments, or persons, and the links are relations among them, such as "depends on," "is part of," "refers to," or "uses." This seemed suitable for a large organization like CERN, and soon after it appeared that the framework proposed by Berners-Lee was very general and would work very well for any set of documents, providing flexibility and convenience in accessing large amounts of text. A very important development of this idea was that the documents need not be stored at the same computer or database but rather, could be distributed over a network of computers. Luckily, the infrastructure for this type of distribution, the Internet, had already been developed. In short, this is how the Web was born.

Looking at the Web many years later and comparing it to the original proposal of 1989, we see two basic differences:

1. The recent Web is huge and grows incredibly fast. About 10 years after the Berners-Lee proposal, the Web was estimated to have 150 million nodes (pages) and 1.7 billion edges (links). Now it includes more than 4 billion pages, with about 1 million added every day.

Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage
By Zdravko Markov and Daniel T. Larose

2. The formal semantics of the Web is very restricted—nodes are simply web pages and links are of a single type (e.g., "refer to"). The meaning of the nodes and links is not a part of the web system; rather, it is left to web page developers to describe in the page content what their web documents mean and what types of relations they have with the documents to which they are linked. As there is neither a central authority nor editors, the relevance, popularity, and authority of web pages are hard to evaluate. Links are also very diverse, and many have nothing to do with content or authority (e.g., navigation links).

The Web is now the largest, most open, most democratic publishing system in the world. From a publishers' (web page developers') standpoint, this is a great feature of the Web—any type of information can be distributed worldwide with no restriction on its content, and most important, using the developer's own interpretation of the web page and link meaning. From a web user's point of view, however, this is the worst thing about the Web. To determine a document's type the user has to read it all. The links simply refer to other documents, which means again that reading the entire set of linked documents is the only sure way to determine the document types or areas. This type of document access is directly opposite to what we know from databases and libraries, where all data items or documents are organized in various ways: by type, topic, area, author, year, and so on. Using a library in a "weblike" manner would mean that one has first to read the entire collection of books (or at least their titles and abstracts) to find the one in the area or topic that he or she needs. Even worse, some web page publishers cheat regarding the content of their pages, using titles or links with attractive names to make the user visit pages that he or she would never look at otherwise.

At the same time, the Web is the largest repository of knowledge in the world, so everyone is tempted to use it, and every time that one starts exploring the Web, he or she knows that the piece of information sought is "out there." But the big question is how to find it. Answering this question has been the basic driving force in developing web search technologies, now widely available through web search engines such as Google, Yahoo!, and many others. Other approaches have also been taken: Web pages have been manually edited and organized into topic directories, or data mining techniques have been used to extract knowledge from the Web automatically.

To summarize, the challenge is to bring back the semantics of hypertext documents (something that was a part of the original web proposal of Berners-Lee) so that we can easily use the vast amount of information available. In other words, we need to *turn web data into web knowledge*. In general, there are several ways to achieve this: Some use the existing Web and apply sophisticated search techniques; others suggest that we change the way in which we create web pages. We discuss briefly below the three main approaches.

Web Search Engines

Web search engines explore the existing (semantics-free) structure of the Web and try to find documents that match user search criteria: that is, to bring semantics into the process of web search. The basic idea is to use a set of words (or terms) that the user

specifies and retrieve documents that include (or do not include) those words. This is the *keyword search* approach, well known from the area of information retrieval (IR). In web search, further IR techniques are used to avoid terms that are too general and too specific and to take into account term distribution throughout the entire body of documents as well as to explore document similarity. Natural language processing approaches are also used to analyze term context or lexical information, or to combine several terms into phrases. After retrieving a set of documents ranked by their degree of matching the keyword query, they are further ranked by importance (popularity, authority), usually based on the web link structure. All these approaches are discussed further later in the book.

Topic Directories

Web pages are organized into hierarchical structures that reflect their meaning. These are known as *topic directories*, or simply *directories*, and are available from almost all web search portals. The largest is being developed under the Open Directory Project (`dmoz.org`) and is used by Google in their Web Directory: "the Web organized by topic into categories," as they put it. The directory structure is often used in the process of web search to better match user criteria or to specialize a search within a specific set of pages from a given category. The directories are usually created manually with the help of thousands of web page creators and editors. There are also approaches to do this automatically by applying machine learning methods for classification and clustering. We look into these approaches in Part II.

Semantic Web

Semantic web is a recent initiative led by the web consortium (`w3c.org`). Its main objective is to bring formal knowledge representation techniques into the Web. Currently, web pages are designed basically for human readers. It is widely acknowledged that the Web is like a "fancy fax machine" used to send good-looking documents worldwide. The problem here is that the nice format of web pages is very difficult for computers to understand—something that we expect search engines to do. The main idea behind the semantic web is to add formal descriptive material to each web page that although invisible to people would make its content easily understandable by computers. Thus, the Web would be organized and turned into the largest knowledge base in the world, which with the help of advanced reasoning techniques developed in the area of artificial intelligence would be able not just to provide ranked documents that match a keyword search query, but would also be able to answer questions and give explanations. The web consortium site (`http://www.w3.org/2001/sw/`) provides detailed information about the latest developments in the area of the semantic web.

Although the semantic web is probably the future of the Web, our focus is on the former two approaches to bring semantics to the Web. The reason for this is that web search is the data mining approach to web semantics: extracting knowledge from web data. In contrast, the semantic web approach is about turning web pages into formal knowledge structures and extending the functionality of web browsers with knowledge manipulation and reasoning tools.