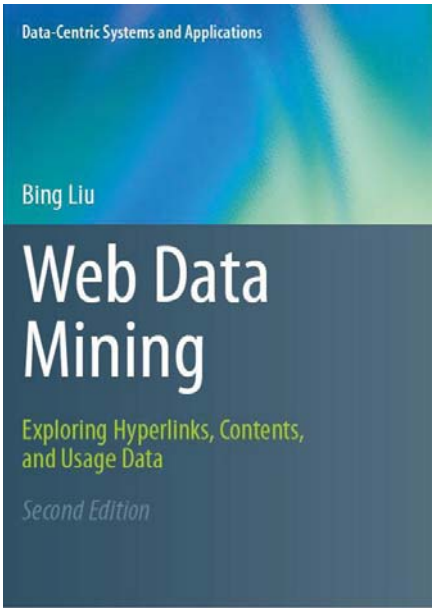
Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data

BY BING LIU– ISBN 978-3-642-19459-7



REVIEWED BY RUSSELL JOHNSON

Not so long ago only a small proportion of the information generated by organizations – that stored in structured data sources like databases and spreadsheets – was accessible for systematic computer-based search, classification and analysis. The techniques and algorithms of data mining were developed to extract useful patterns and knowledge from these structured sources. Today, the explosive growth of the World-Wide-Web, and a parallel development of private intranets, means that a much greater amount of information is potentially available for search and analysis, in a wider variety of formats and encompassing structured, semi-structured and unstructured data, from organized tables to multi-media clips. The established techniques of data mining have proved insufficient for this task.

Over the past decade, new techniques and algorithms have been developed which aim to discover useful information or knowledge from computer analysis of the hyperlink structures, page contents, and usage data of Web resources. *"Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data"* by University of Illinois Professor Bing Liu provides an in-depth treatment of this field.

In the introduction, Liu notes that to explore information mining on the Web, it is necessary to know data mining, which has been applied in many Web mining tasks. However, he points out that Web mining is not entirely an application of data mining. Due to the richness and diversity of information and other Web specific characteristics discussed above, Web mining has developed many of its own algorithms.

One of the standout features of Liu's book is that it encompasses both data mining and Web mining. The first half of his book outlines the major aspects of data mining which Liu lists as supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining, and provides examples of how these techniques are used in Web mining.

In the second half, the author focuses on specific web mining techniques. Based on the primary kinds of data used in the mining process, Liu categorizes these into three types: Web structure mining, Web content mining and Web usage mining.

Web structure mining abstracts useful knowledge from the hyperlink structure of the Web, which search engines do to discover important Web pages. It is also possible to discover communities of users who share common interests. Liu points out that traditional data mining cannot perform such tasks because relational tables do not have link structures. *Web content mining* extracts useful information from the contents of Web pages, such as automatically classifying and clustering pages according to their topics. These tasks are similar to those in traditional data mining. However, additional tasks can be performed such as mining customer reviews and forum postings to determine customer opinions which are not traditional data mining tasks.

Web usage mining involves discovering user access patterns from Web usage logs, which record each user's mouse clicks, and applying data mining algorithms to them.

Summary of Chapters

It is useful at this point to provide a summary of the chapter contents. As previously explained the book is divided into two main parts. Chapters 2-5 cover the major topics of data mining while the remaining chapters cover Web mining, including a chapter on Web search.

Chapter 1 provides a brief history of the web and of web mining. It also provides a useful summary of the contents of the book and how to read it.

Chapter 2 studies Association Rules and Sequential Patterns, which have been used in many Web mining tasks, especially in Web usage and content mining. Association rule mining finds sets of data items that occur together frequently. Sequential pattern mining finds sets of data items that occur together frequently in some sequences, and can be used to find regularities in the Web data.

Supervised learning (Chapter 3), or classification, is frequently used in both practical data mining and Web mining. It aims to learn a classification function (called a classifier) from data that are labeled with pre-defined classes or categories. The resulting classifier is then applied to classify future data instances into these classes.

In unsupervised learning (Chapter 4), the learning algorithm has to find the hidden structures or regularities in the data without any pre-defined classes, such as with clustering, which organizes data instances into groups or clusters according to their similarities (or differences).

Chapter 5 explains the Partially Supervised Learning model replacing the large number of manually-labeled examples required for the supervised learning model with a small set of labeled examples (data instances) and a large set of unlabeled examples for learning.

The section on web mining begins with Chapter 6 on Information Retrieval and Web Search. The vast scale of the web means that web search algorithms must not only be accurate but also efficient.

Chapter 7 (Social Network Analysis) outlines how hyperlinks are exploited for efficient Web search including looking at Google's hyperlink-based ranking algorithm, PageRank, which originated from social network analysis, and community finding algorithms which underpin social networking sites.

In Chapter 8 examines the techniques of web crawling – traversing the Web's hyperlink structure and locating pages linked by topic – which is usually the essential first step of Web mining or building a Web search engine

In Chapter 9 (Structured Data Extraction: Wrapper Generation) data mining techniques are explored for analyzing web pages based on structured data, possibly extracted from databases, to identify the underlying patterns and extract the data to provide value-added services such as comparative shopping.

Information Integration (Chapter 10) involves matching and integrating information from different web sources, especially structured data, to provide consistent and coherent database.

Chapter 11 (Opinion Mining and Sentiment Analysis) looks at techniques for analyzing and categorizing the huge amount of unstructured text on the Web to mine people's opinions and sentiments expressed in product reviews, forum discussions and blogs.

Finally, Chapter 12 on Web usage mining aims to study user clicks and their applications to e-commerce and business intelligence. The objective is to capture and model behavioral patterns and profiles of users who interact with a Web site to support *recommender* systems and other commercial applications.

Conclusion

Originally published in 2007, this revised and updated second edition is primarily intended as a textbook for an advanced computer science course. Each chapter has a reference section for future reading. In addition, a companion website at http://www.springer.com/3-540-37881-2 provides updates, lecture slides, implemented examples and other useful teaching resources.

Two chapters are mainly contributed by three other researchers Filippo Menczer, Bamshad Mobasher, and Olfa Nasraoui. Menczer contributed Chapter 8 on Web crawling, while Bamshad and Olfa wrote most of Chapter 12 on Web usage mining.

The content of *"Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data"* is comprehensive and in-depth. While the book's focus in web mining Liu recommends that students without a background in machine learning should not skip the sections on data mining. He also suggests that the early sections could provide the basis for an introductory data mining course, especially if integrated with the chapters on search engines and social networking which are of topical interest among students.

THE BOOK:

LIU, BING (2011) WEB DATA MINING, 2ND EDITION. 2011, XX, 622 P. ILLUS. SPRINGER.
ISBN: 978-3-642-19459-7

ABOUT THE REVIEWER:

RUSSELL JOHNSON
School of Engineering and Advanced Technology, Massey University, New Zealand. Contact him at: r.s.johnson@massey.ac.nz