Journal of Statistical Software

April 2008, Volume 25, Book Review 1. http://www.jstatsoft.org/

Reviewer: Patrick Mair
Wirtschaftsuniversität Wien

Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage

Zdravko Markov and Daniel T. Larose
John Wiley & Sons, Hoboken, NJ, 2007.
ISBN 978-0-471-66655-4. xvi + 218 pp. USD 56.66.
http://www.dataminingconsultant.com/

This third volume of the Wiley series on data mining textbooks covers the topic of Web mining. Web mining, being a subdiscipline of data mining, covers the analysis of data stemming from Web applications. This introductory book is divided into three parts: Web structure mining (Part I of the book), Web content mining (Part II), and Web usage mining (Part III). In total, nine chapters are assigned to these parts. At the end of each chapter the authors provide numerous (real-life) exercises which allow the reader to understand more deeply the methodological explanations. Throughout the text, examples are computed using either the open-source software **Weka** (Part I and II) or the SPSS data mining software **Clementine** (Part III).

In the first chapter, the authors start with the description of the challenges emanating from Web data. By means of the Web crawler **WebSPHINX** the reader is introduced to basic structures of the Internet and it is shown how to extract relevant information contained in HTML files. Within this text mining context, basic terminologies such as term-document matrices, term frequency, term frequency-inverse document frequency, various similarity measures, etc., are explained. These measures are important for the further understanding of the book.

The subject of Chapter 2 is hyperlink-based ranking. The two main algorithms described are Google's PageRank and HITS (hyperlink induced topic search). Particularly the first one could have been described in more detail, since many readers are surely interested in what's really happening behind the Google application, without having to dig through Page and Brin's original papers (Brin and Page 1998; Page, Brin, Motwani, and Winograd 1998).

The Web content mining part begins with Chapter 3 dealing with cluster algorithms. The authors start with hierarchical clustering, proceed with k-means, and finally describe EM-based probabilistic clustering. Based on the latter two, sketches are given on how user profiling and Web personalization can be achieved (collaborative filtering). All these approaches are completely embedded into Web mining scenarios, and for more detailed formal descriptions the reader is referred to Larose's introductory book (Larose 2005) of the above mentioned Wiley series. How these clustering results can be evaluated in terms of similarity based and

probabilistic criterion functions, the principle of minimum description length, precision, recall, and entropy, is given in Chapter 4.

Whereas the previous two chapters are typically part of unsupervised learning, Chapter 5 is about classifiers which are considered as supervised learning algorithms. Web-relevant heuristics such as nearest-neighbor, naive Bayes, and relational learning, including additionally a special emphasis on feature selection, are explained.

Web usage mining is introduced in Chapter 6. The authors point out the difficulties of the data struture contained in a server log file and describe how to organize corresponding clickstream data. One of the most challenging and time-consuming tasks in clickstream analysis in practice is the pre-processing step. The authors could have provided some more details on the ETL-process. Nevertheless, they provide clear explanations regarding data cleaning and filtering, de-spidering the log file, and the crucial issue of user/session identification, a key point for a successful Web shop.

Chapter 8 covers exploratory data analysis. Basic descriptive results such as number of visit actions, session/individual page duration, and average page dwell times are, in practice, highly valuable for the provider. Advanced analyses in terms of clustering, association rules, and classifiers are presented in the final Chapter 9. The focus is on algorithms offered in **Clementine** such as Birch clustering also known as 2-step clustering, a-priori association rules, and decision tree algorithms like CART and the concluding C4.5.

Overall, this book keeps the high didactical standard of Larose's first two books (Larose 2005, 2006). It is accompanied by a comprehensive collection of online materials such as power point slides for each chapter and numerous data sets (although at the time of this review, January 2008, this collection is not yet accessible). However, some materials can be found on the authors' Web sites.

From a methodological point of view, the authors' descriptions of the algorithms are clear and invariably accompanied by interesting applications on Web data. However, in my opinion, the authors could also have presented some more complex, up-to-date approaches such as modeling page transistions based on Markov chains.

As a final point it has to be noted that this book is an excellent resource for conducting Web mining lectures or single units within a Data mining class. The data sets can be used for small as well as quite comprehensive business intelligence projects. The book's content is easy to access; even students with very basic statistical skills can get the flavor of the intriguing aspects of Web mining.

References

Brin S, Page L (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems Archive*, **30**(1–7), 107–117.

Larose DT (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley & Sons, Inc., New York.

Larose DT (2006). *Data Mining: Methods and Models.* John Wiley & Sons, Inc., New York.

Page L, Brin S, Motwani R, Winograd T (1998). "The PageRank Citation Ranking: Bringing

Order to the Web." *Technical report*, Stanford Digital Library Technologies Project. URL http://citeseer.ist.psu.edu/page98pagerank.html.

Reviewer:

Patrick Mair
Wirtschaftsuniversität Wien
Department of Statistics and Mathematics
Augasse 2–6
A-1090 Vienna, Austria
E-mail: Patrick.Mair@wu-wien.ac.at
URL: http://statmath.wu-wien.ac.at/~mair/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 25, Book Review 1 *Published:* 2008-04-30
April 2008