Book Review: Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data

Olfa Nasraoui
Knowledge Discovery & Web Mining Lab
Computer Engineering and Computer Science department
Speed School of Engineering
University of Louisville
Louisville KY 40292

olfa.nasraoui@louisville.edu

ABSTRACT

This paper presents a review of the book “Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data” by Bing Liu. The review concludes that the breadth and depth of this book makes it a required staple for every Web mining researcher, student, or practitioner.

Keywords

Data mining, Web mining, text mining, Web information retrieval, crawling, Web search, data extraction, link analysis.

1. INTRODUCTION

So what does the author, Bing Liu know about Web data mining to write the book “Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data”[1] ? Fortunately the answer is “a lot!” This fact along with the title which had some cosine similarity with the names of my research lab and a graduate course that I have been teaching at the University of Louisville since 2004, and prior to that at the University of Memphis since 2000, are the reasons why I ordered a copy of this book. Bing Liu is a well seasoned researcher who has made significant contributions to association rule mining, in particular classification using association rule mining and association rule mining with multiple supports. He has also worked on Web data extraction, and more recently on opinion mining. In addition to the expertise of the author, two of the chapters, Chapter 8, Web Crawling, and Chapter 12, Web Usage Mining, were contributed by two leading experts in these respective areas, Filippo Menczer for the former and Bamshad Mobasher for the latter.

This book is appropriate for students at the graduate or senior undergraduate level, for practitioners in industry, and even as a good comprehensive reference for researchers in academia.

The Table of Contents held a surprise for someone who had always found it hard to limit the number of textbooks to one book in a web mining course that does not have data mining as prerequisite, and thus typically prescribes a good data mining book to introduce data mining techniques, in addition to a second book related to web mining. This book, on the other hand, has two parts, one devoted to data mining, and the other devoted to Web mining. While it was not a problem to find a very good data mining book (I have a few of them on my bookshelf), it was harder to find a book that addressed data mining and Web mining. It was also hard to find a good and comprehensive Web mining book, since most of them tend to focus on one or only two of the three main Web mining areas of Web structure, content, and usage mining (typically leaving Web usage mining in the dark, with just a small section, citing that it is an *emerging* area). This book, on the other hand, is a serious book on Web mining that also devotes a decent portion to data mining. I would describe the way the topics are presented as deep and rigorous enough in most chapters, which is in contrast to a large number of books on data mining and web mining. That said, because the book is full of simple examples that illustrate the methods being discussed, it is useful even for beginners, making it also appropriate for an introductory level course.

2. REVIEW

The book is well structured with the chapters divided into two parts. The first part consists of Chapters 2-5, covering the *data mining* tasks of association rule and sequential pattern mining, supervised learning, unsupervised learning, and partially supervised learning. The second part consists of Chapters 6-12, covering Information retrieval and search, link analysis, Web crawling, structured data extraction (wrapper generation), information integration, opinion mining, and Web usage mining.

Typically, a topic is presented starting with a motivation, followed by the pertinent notations, equations and definitions, then an algorithm and a concrete example illustrating the ideas. Furthermore, every chapter ends with a section titled *Bibliographic Notes* that places the presented methods in a historical context, and then points the curious reader toward more literature on the topic.

The book starts with an introduction (Chapter 1) that overviews the history of the WWW, and then discusses the challenges that distinguish Web mining from data mining. It does so by emphasizing those aspects of the Web that are not typical in most other data sets. For example, the Web is a place of interaction between people and automated services, it is huge, noisy, heterogeneous, and full of unstructured and semi-structured data. Also, because of reputation concerns, all data is not considered equal on the Web. After this motivation, definitions are given for data mining and for Web mining. The Web mining definition followed suit to the data mining definition, with a redirection of the source of data to the Web, thus the traditional trilogy of the Web's hyperlink structure, page content, and usage data. The author stresses that Web mining is *not* to be viewed as an *application* of traditional data mining! This view was supported by citing the distinguishing characteristics of the Web data, such as its *heterogeneity* and *lack of structure* which have led to the invention of *new* specialized mining *tasks* and *algorithms* in Web mining. As a researcher in this area, I concur with this view, which is not shared by some other data mining books that typically stow Web mining as an application. This, despite the fact that the Web mining area is so vast that many problems have no counterpart in traditional data mining, and these problems, and interest in them, have expanded over the years. I would cite as examples of specific web mining algorithms, the PageRank and the HITS algorithms for link analysis. The Web mining process is then distinguished further from that of data mining, citing the issue of *data collection* which can be substantial in the former compared to the latter. This claim, I found, provided a good justification for including an entire chapter (Chapter 8) on *Web crawling* in the book, as well as other chapters devoted to *data extraction* and to *data integration*. That said, in addition to being directed toward data gathering, I would add that Web crawling often requires powerful data mining methods (example: classification) in order to guide the crawler, particularly in the case of focused crawling. I would add here that compared to conventional data mining, even the data pre-processing and the evaluation and post-processing can be daunting in Web mining.

Chapter 2 (*Association Rules and Sequential Patterns*) covers two important data mining tasks that are particularly important in Web usage mining. It is easy to notice how this chapter has benefitted from the girth of the author's expertise in this subject, as it starts with the *Apriori* algorithm, but then moves on to more advanced variants that are rarely covered in other data mining books, in particular *mining with multiple minimum supports*, which can be considered crucial in most of the data sets dealt with in Web usage mining. This is due to the vast difference between the supports of the different items, stemming from the power law distribution of most Web related data. For instance, we are all too familiar with the presence of a long tail of infrequent items in most e-commerce transactions that are less constrained by the physical limitations of the warehouse of their offline counterparts. The same phenomenon occurs when one considers the support of words in text corpora. After this, the chapter covers mining *class association rules*, which are useful for transactional data and certain kinds of categorical data that are common in e-commerce applications. Finally, the chapter concludes with mining sequential rules based on GSP and based on PrefixSpan, both with and without multiple minimum supports. Missing in this chapter is the FP-tree approach, which can significantly compress Web transaction data for subsequent AR mining.

Chapter 3 discusses *Supervised Learning*, i.e. classification, and most importantly talks about it within the context of classifying text documents. The emphasis on text is further exemplified in the choice of evaluation metrics (precision, recall, F1) that were discussed, which are suitable for imbalanced data, as is often the case in text classification, and especially in information filtering. The chapter could have benefited from also presenting ROC curves for more than two classes. That said, the text is rigorous, including for instance the derivations of the equations for Support Vector Machines (SVM), for the separable and non-separable cases.

Chapter 4 (*Unsupervised Learning*) presents clustering algorithms, in particular the K-means algorithm and hierarchical clustering, which in my opinion are insufficient for the Web mining field. I would have preferred a more thorough chapter on clustering, including the EM algorithm for mixture models and the Spherical K-Means which in my opinion, are more scalable and more suitable for Web documents and Web sessions. The author also discussed *data standardization* at the end of this chapter. This is an important part of pre-processing, and should have been presented at the start of the Supervised Learning chapter which came earlier. For some of the presented supervised learning methods (e.g. decision trees), standardization may not be crucial, but for many supervised learning methods such as neural networks, it is very critical.

Chapter 5 (*Partially Supervised Learning*) talks about the case when some of the data is labeled, while the rest is unlabeled (LU learning), then moves to the two-class case of positive labels versus no labels (PU learning). This chapter is very valuable in Web mining, as one is often overwhelmed by the massive size of web pages to label on the Web, thus ending up with labeling only a small sample. Because the sample is very small however, the model accuracy may not be satisfactory. Fortunately, LU learning or semi-supervised classification is based on taking advantage of a larger set of *unlabeled* samples in order to improve the accuracy of the learned classification model. The latter case is particularly suitable for text information filtering where one has only a few examples of documents from a certain topic, and would like to find more similar (topic-wise) documents from a large collection that contains all kinds of topics. The only thing missing in this chapter is the mention that LU learning is also useful for clustering (semi-supervised clustering), where a limited sample of *labeled* web pages can guide the clustering of a larger set of unlabeled samples.

Chapter 6 (*Information Retrieval and Web Search*) presents Web search, a very popular problem, as the single most important application of the much older field of Information Retrieval (IR). Yet, the author stresses that Web search is *not* only a simple *application* of IR because there are many unique characteristics in Web data, for example the hyperlink information, the deceptive Web content such as Web spam, and the massive size that rules out all but the most scalable search engines. This is analogous to the distinctions made earlier between Web mining and data mining. Web IR was presented rather rigorously. It started with the theory, i.e. the IR models (Boolean, vector space, probabilistic), then moved on to *relevance feedback* with the *Rocchio* model and machine learning methods, including a relevant connection to the previous chapter (LU and PU learning). The chapter then presents some IR evaluation measures and delves into text pre-processing which includes very search-specific issues such as duplicate detection. Part of this section (tokenization, stop work elimination, stemming) should have been included at the start of Chapter 3 (Supervised Learning) because it is a pre-requisite to text classification. The chapter continues with a detailed discussion of *inverted indexing*, including various compression methods (e.g. Elias gamma, Elias delta, and Golomb coding) that are well illustrated with examples.

Section 6.7 is devoted to latent semantic indexing, while Section 6.8 attacks the "*search*" part of Web IR, followed by a good discussion of *meta-search* and various ranking combination schemes in Sec. 6.9. The chapter ends with a section on Web spamming.

Chapter 7 (*Link Analysis*) starts with a nice overview of social network analysis, thus defining important metrics in this area, such as *closeness* and *betweenness centrality*, then it moves to citation analysis, before delving into the PageRank and HITS algorithms in detail (again with several examples), and then ending with community discovery.

Chapter 8 (*Web Crawling*) was contributed by an expert on this topic, Filippo Menczer, who starts with *universal* crawlers or spiders, and finishes with *focused* (examples of positive and negative pages are available) crawlers and *topical* (only some positive examples are used as seeds) crawlers. The chapter even addresses certain implementation issues such as parsing, spider traps, and concurrency. Then Menczer presents a brilliant discussion on *adaptation* in crawlers, delighting the reader with details about *InfoSpiders*, that adapt through reinforcement learning. The chapter concludes with a section crawler ethics and etiquette, thus addressing such gimmicks as *cloaking* and such anti-crawler tricks as the graphic-based inverse Turing tests known as CAPTCHAs. The last discussion on new developments is enlightening and up to date, as it mentions the future of peer to peer search, for example.

Chapter 9 is titled *Structured Data extraction: Wrapper Generation*. Not to disappoint anyone, this is not the hard coded template based extraction of the late 90's, but rather the data mining oriented data extraction. The latter uses training samples to learn automated extractors that are essential to some data gathering tasks where one wants to collect, say, the Data Rich Pages that list product features and prices on an e-commerce website. The chapter is detailed and contains numerous examples with nice diagrams that make it easy to quickly grasp the concepts.

Chapter 10 (*Information Integration*) may at first appear to be in the wrong book (on databases). However, starting with the first page, the author makes it clear that this is a *sequel* to the previous chapter on structured data extraction, for the case when we want to extract data from *multiple* websites. This subject turns out to be essential to extracting the right data from the *deep Web* (Web databases), and should thus be considered a crucial step in *deep Web mining*.

Chapter 11 is devoted to *Opinion Mining*, and thus focuses attention on *user generated* content, as is common on blogs and discussion forums. The author presents methods for *sentiment mining*, then *feature-based opinion mining and summarization*, then *comparative sentence and relation mining*, and finally *opinion search* and *opinion spam* and its detection. Just to emphasize the thoroughness of this chapter, I would say that it contains 12 examples!

This finally leads us to Chapter 12 (*Web Usage Mining*), itself contributed by a leading expert in the field, Bamshad Mobasher. The chapter starts with a detailed explanation of data collection and pre-processing, followed by a discussion of data modeling for web usage mining. It then delves into the methods of Web usage pattern discovery, including clustering, association rule and sequential pattern mining, and classification and prediction (with collaborative filtering, an essential social filtering method used to provide recommendations on e-commerce websites). This chapter provides an excellent foundation in Web usage mining (WUM), but could benefit from adding more contemporary Web mining tasks such as query log mining and addressing challenges such as scalability and adapting to the evolution of user activity. A section on privacy would also make this chapter more rounded. That said, this chapter is a welcome and practical treatment of WUM that has not been thoroughly presented in previous books.

3. CONCLUSIONS

I will most likely keep using this book for teaching Web mining, continue to recommend it to other researchers in the area, and continue to make it a *required reading* for every student working in my lab. Fortunately for those who are considering using this book for teaching, the author has made available a complete list of lecture slides for each chapter (except for Chapter 12) on http://www.cs.uic.edu/~liub/WebMiningBook.html. The book could benefit from adding problems, at the end of each chapter, that would not only spark the curiosity of the casual reader, but also support using the book for teaching. The few shortcomings that were mentioned in this review are actually easy to mend, because the book started on a solid foundation.

I must mention that there are several other good Web mining books that I consider a requirement on every Web miner's bookshelf, however they tend to have a different composition than Bing Liu's book. For example one of the best in my opinion, Chakrabarty's book [2] is a solid book on Web content and structure mining, but it has only one small section on Web usage mining tucked within a chapter on emerging problems at the end. Another very good book, by Baldi et al. [3], is rather focused on *modeling* the Internet, and most of the techniques therein emphasize a *probabilistic* approach, leaving in obscurity methods like association rule mining. Also, most other books are in need of an update to cover the more recent methods in this fast moving area and the interesting problems that have emerged in the last few years (e.g. opinion mining).

4. REFERENCES

[1] B. Liu, *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*, Springer Series on Data-Centric Systems and Applications, 2007.

[2] S. Chakrabarty, *Mining the Web: Analysis of Hypertext and Semi Structured Data*, Morgan Kauffman, 2002.

[3] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web - Probabilistic methods and algorithms*, Wiley, 2003.

Olfa Nasraoui is an Associate Professor in computer engineering and computer science at the University of Louisville, where she is also the endowed Chair of e-commerce. Her research activities include data mining, Web mining, mining evolving data streams, personalization, and computational intelligence. She has served on the organizing and program committees of several conferences and workshops, including co-chairing the ACM KDD workshops on Web mining and Web usage analysis (WebKDD) from 2004 to 2008. She is the recipient of a National Science Foundation CAREER Award, and is a member of IEEE, IEEE Women in Engineering, and ACM. (http:/www.webmining.spd.louisville.edu/)