# Book Review

## Mining the Web: Discovering Knowledge from Hypertext Data

Soumen Chakrabarti.
San Francisco, CA: Morgan Kaufmann, 2003. 345 pp. $54.95. (ISBN: 1-55860-754-4.)

After the invention of the Web many things in our lives (such as how we do research, communicate, entertain ourselves, and do personal finance) have changed. Simply, the Web has changed our lives. Perhaps, it even contributed to the world-wide obesity problem, especially among the youth. Like other technological inventions, it made our lives easier, lazier, and more enjoyable.

About ten years ago, soon after the Web's birth, Web "search engines" were first by word of mouth. Soon, however, automated search engines became a world wide phenomenon, especially AltaVista at the beginning. I was pleasantly surprised by the amount and diversity of information made accessible by the Web search engines even in the mid 1990's. The growth of the available Web pages is beyond most, if not all, people's imagination. The search engines enabled people to find information, facts, and references among these Web pages.

The characteristics of the Web environment (too much of good and bad things; abundance, redundancy, and misrepresentation; too many pages satisfying typical user queries) make researchers look for retrieval techniques which are more sophisticated than the ones used in traditional information retrieval systems. For this purpose data mining is a sensible choice.

Hidden statistical dependencies among the attributes of an object of interest can reveal relationships or structures which are not intuitively obvious. The search for such statistical dependencies, or patterns, is referred to as machine learning or data mining. For example, quantitative text attributes (such as the rate of occurrences of function words in text blocks) can be used to discover non-obvious patterns in writings, and they can be used for author attribution and other purposes. In the case of the Web, data mining aims to exploit the patterns among terms, Web pages, hyperlinks, hypertext markup, sites, and topic directories for making accessibility of Web information more effective and efficient, or simply to connect users to information they seek. An example of machine learning for Web applications would be Web personalization, for instance, modeling a user's topics of interest and recommending Web pages to a user based on past behavior. This book on Web mining focuses on the application of data mining in this phenomenal, dynamic, huge, uncontrolled information environment.

The book contains nine chapters, and after the first chapter, *Introduction*, it is divided into three parts. The contents of the chapters are briefly defined in the following. To reflect the emphasis of topics in the book, after each chapter title, I provide the number of pages in it and the approximate percentage of its size in the main body of the book, which is 306 pages, excluding references and the index section. Some of these pages (10, 3%) are left blank or just contain section titles for a roomy pleasurable presentation. In the following, for convenience, the percentage size of each of the three parts is also given.

*Introduction* (13, 4%), contains one chapter. Chapter 1 provides an overview of the book.

Part I, *Infrastructure* (20%), contains two chapters, Chapter 2: *Crawling the Web* (27 pages, 9%), and Chapter 3: *Web Search and Information Retrieval* (32, 11%). Chapter 2 provides crawler implementation details and the related performance issues and possible crawler problems, such as spider traps and duplicate pages under different names. Chapter 3 covers the fundamental concepts of the information

retrieval problem within the context of the Web. As indicated by the author, Part I can be skimmed if the reader is more interested in data mining.

Part II, *Learning* (40%), contains three chapters, Chapter 4: *Similarity and Clustering* (45, 15%); Chapter 5: *Supervised Learning* (52, 17%), Chapter 6: *Semisupervised Learning* (23, 8%). Chapter 4 first introduces the motivation behind clustering, i.e., the cluster hypothesis, then covers various clustering methods and the incorporation of the Web-based attributes to the clustering process. The following chapter considers the maintenance of an existing taxonomy in the dynamic Web environment. Chapter 6 discusses the incorporation of hyperlink-derived information to the learning process.

Part III, *Applications* (34%), contains three chapters, Chapter 7: *Social Network Analysis* (52, 17%), Chapter 8: *Resource Discovery* (34, 11%), and Chapter 9: *The Future of Web Mining* (18, 6%). The first chapter of this part studies link-based techniques for enhancing text-based retrieval and ranking strategies. The middle chapter of this part focuses on locating desired resources, such as preferential crawling, in distributed hypertext. Finally, the last chapter describes techniques for analyzing documents at the level of tokens, their proximity, and their relationships with each other.

Each chapter of the book, except the last one, ends with bibliographic notes. They will be especially useful for researchers. The book contains 220 references with 180 different first authors. This means 1.2 citations per cited first author which indicates a considerable amount of diversity among the cited references. Some of the references are to Web pages and the readers are reminded that if the cited URLs become unavailable, the (Internet Archive) Web site located at www.archive.org may be helpful in locating these pages in the future, provided that the "Archive" is still available.

The book targets researchers, and graduate and senior undergraduate students. All targeted audiences will find it valuable. Knowledge of elementary undergraduate statistics, algorithms, and networking will be sufficient to follow the book's material. Some of the deliberate omissions of the book include natural language processing, and Web usage mining.

It is printed on acid free paper with ample margins that can be used for personal notes. The index of the book is rather comprehensive and contains about 1000 items. The quality of binding is excellent. However, it would be nice if the publisher used a darker font in future editions of the book, since the current font is rather pale.

All in all this is an excellent book. I enjoyed the book and highly recommend it as a textbook for Web data mining classes at graduate or senior undergraduate levels. In future editions, the author may extend the book by adding end of chapter problems to make it even more attractive as a textbook. Researchers working on the general information retrieval problem will also find it useful and inspirational. The semi-formal presentation of the author switches between "I" and "we" and it flows very nicely. Chakrabarti has a rich vocabulary and is a gifted writer. I bet he will write new, good books in the future, and he should. I look forward to them.

**Fazli Can**
*Computer Science and Systems Analysis Department*
*Miami University*
*Oxford, OH 45056*
*E-mail: canf@muohio.edu*