



# NEW BOOKS IN REVIEW

EDITOR: *Naveen Donthu*

ASSOCIATE EDITORS: *Meryl P. Gardner*  
*Sandeep Krishnamurthy*  
*Stephanie Noble*

MODELING THE INTERNET AND THE WEB: PROBABILISTIC METHODS AND ALGORITHMS, Pierre Baldi, Paolo Frasconi, and Padhraic Smyth, West Sussex, UK: John Wiley & Sons, 2002, 285 pages, \$78.30.

MINING THE WEB: DISCOVERING KNOWLEDGE FROM HYPERTEXT DATA, Soumen Chakrabarti, San Francisco: Morgan Kaufman Publishers, 2003, 344 pages, \$57.95.

HANDBOOK OF GRAPHS AND NETWORKS: FROM THE GENOME TO THE INTERNET, Stefan Bornholdt and Heinz Georg Schuster, Weinheim, Germany: Wiley-VCH, 2002, 417 pages, \$145.00.

## MODELING AND MINING THE WEB

The World Wide Web is a vast repository of interconnected content. Today, the challenge is not scarcity of information, but rather the overabundance of it; at the time of writing this, Google's index had just doubled to more than eight billion Web pages, a daunting number, but a mere subset of the overall Web. Customer relationship management systems routinely collect terabytes of data, and new models are required to recognize patterns in this data. Data mining is an area that has emerged to tackle the challenges of this new informational environment. In this review, I discuss three books that cover state-of-the-art techniques of data mining, especially in the context of the Web.

Data mining is still a nascent field that draws heavily from multiple disciplines (for an early and influential article in the field, see Zaiane, Xin, and Han 1998). The three books I discuss herein provide a general understanding of the domain. All three books are written with the advanced reader in mind. Moreover, because all three books have a strong engineering rather than a social sciences orientation, the emphasis can sometimes be excessively on understanding the "innards" of data collection programs (especially search engine crawlers) and building better algorithms. Some readers may be unfamiliar with the "pseudocode"—a common practice in engineering and computer science books—in Chakrabarti's and the Baldi, Frasconi, and Smyth's books. Only portions of these books address understanding human behavior in the online context, and only small subsets address the commercial context. Readers who are familiar with the marketing literature will recognize many well-known concepts (e.g., k-means clustering) in an

unfamiliar context (e.g., categorizing data collected in search engine crawls). At the same time, some of the newer concepts that have emerged from the field of data mining are educational. For example, complicated correlation-based topological graphs help punctuate the point in Baldi, Frasconi, and Smyth's book.

Of the three books I review herein, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, by Pierre Baldi, Paolo Frasconi, and Padhraic Smyth, is the most accessible and does not presuppose an unduly high level of expertise in the field. I recommend this book to the reader who is interested in becoming familiar with this emerging domain. This book is also a good graduate-level textbook for data mining because it has good pedagogical features (e.g., exercises at the end of each chapter).

This book commences with a short mathematical tutorial on some of the basics that are necessary to comprehend the contents of the book. Then, the book provides a lucid description of the basic technologies that underlie the Web. The reader is exposed to the basics of HTML, TCP/IP, and search engine design. In contrast to Chakrabarti's book, this book does not discuss too many engineering details of search engine design. Rather, it is more statistical and thus closer to the mind-set of the marketing reader.

Next, the authors review the considerable literature on the Web as a graph. They discuss the now-famous bowtie study. The result of this study is that the Web is uneven in its connectedness. Some portions of the Web only have "outlinks," others only have "inlinks," and only some have both.

The chapter on text analysis may take many readers by surprise. Many marketing scholars are familiar with the textual analysis that accompanies verbal protocol or focus group data. However, this chapter is about textual databases that have billions of records. The findings in this chapter are drawn from the areas of machine learning and information retrieval and are useful for large data sets of textual documents. This chapter discusses probabilistic techniques to classify and then retrieve documents from a large data set (e.g., search engine index). The first part of the chapter explains the concepts clearly, and then it provides several interesting applications (e.g., classification of Web pages, news stories, spam filtering).

The chapter on link analysis draws from graph theory. Readers may be interested in an analysis of Google's algorithm (PageRank). The discussion of search engine persuasion, the phenomenon in which a group of accomplices link

to one another to inflate and conflate page authority, is especially interesting.

Many marketing scholars will perhaps go directly to Chapter 7 of this book, “Modeling and Understanding Human Behavior on the Web.” In this chapter, the book moves from an engineering approach to somewhat of a social science approach. This chapter begins with a description of the issues that are idiosyncratic to data collection on the Web (e.g., proxy servers, caching) and then describes the typical data that are available from Web pages. The chapter then enters into a detailed analysis of Markov models for page prediction (i.e., how users transition from page to page within a given Web site). Beginning with a simple model, the chapter moves on to more complicated Markov models and discusses how these models describe and predict observed behavior. The chapter discusses one specific surfing model by Huberman and colleagues (1998). At the end of this chapter, the authors present statistical models that predict consumer querying behavior on search engines.

The next chapter, “Commerce on the Web: Models and Applications,” may also be of interest to readers. This chapter discusses the familiar Bass model (in the context of Hotmail adoption) with reference to a marketing article by Montgomery (2001). The chapter examines building better recommender systems by using collaborative filtering techniques, a topic not entirely unfamiliar to marketing scholars. The section in this chapter on Web path analysis for purchase prediction is a must-read. The analysis models a consumer’s path on a commercial Web site and, using a transition matrix, attempts to predict when the person will actually purchase the product.

Overall, Baldi, Frasconi, and Smyth’s book is a winner. The authors attempt to elucidate a complicated area to readers rather than intimidate them. They make every attempt to explain the basic concepts before moving on to more complicated ones. Exercises and applications enhance readability. Graduate students who wish to work in the area of e-commerce should at least peruse this book.

More than the other two books, *Mining the Web: Discovering Knowledge from Hypertext Data*, by Soumen Chakrabarti, focuses extensively on building a better search engine crawler. Therefore, it may be beneficial to review the basics of search engines. Search engines collect information periodically by using software programs called crawlers. These programs visit pages and follow the links on to other pages. New pages are collected and added to a database called an index. When a user queries a search engine (e.g., Google), he or she queries this database, which is the last snapshot of the Web. Search engines differ in the algorithms they use to order the relevance of results. Google’s PageRank is perhaps the most well known; it categorizes pages on the basis of the number of inlinks. Thus, it treats an inlink as a rare commodity and a vote of confidence in any given page.

Chakrabarti’s book begins with a discussion of search engine crawlers in a chapter titled “Crawling the Web.” The discussion in this chapter is technical and detailed. Readers learn about features such as the robots.txt file, a file that can be written in a certain way to stop crawlers from visiting a page. Chapter 3 is about organizing the data in the index in a useful way, and Chapter 4 is about relevance; that is, how a person maximizes the similarity between a user query and

documents in a search engine’s index. Readers will see many familiar concepts in this chapter (e.g., clustering, multidimensional scaling).

The most interesting part of the book is perhaps Chapter 7, “Social Network Analysis.” In this chapter, the author presents the most famous search engine algorithms (e.g., PageRank, HITS, SALSA). The presentation is rigorous and terse and may intimidate a beginner. We learn that the most important element of search engine design is the identification of the authoritative source and how the different algorithms achieve this. The author presents interesting problems that search engines face (e.g., nepotistic links, pages on the same host that link to one another; clique attacks, a group of people that link to one another in the hope of boosting their relevance; mixed hubs, pages that may score high on multiple words).

The problem with the Web is that different parts of it change with different levels of frequency. Thus, a news page may change every few hours, whereas a faculty member’s homepage may change only two or three times a year. Using one crawler to crawl both pages is inefficient. There are already examples of pages that are dynamically constructed by crawling a small population of pages at a high frequency (e.g., Google News crawls 4500 news sources to build a news page dynamically). Chapter 8 presents techniques on doing this.

In the last chapter, the author discusses the future of data mining and, more specifically, what future search engines might look like. A way that search engines may evolve is question answering; a person could simply ask a search engine a question, such as “What is the capital of Alaska?” and the search engine will respond. Personalized search results will also prove useful. The bibliography at the end of the book will be useful to all readers. Although most of the sources are drawn from the engineering literature, they should be a great starting point for researchers.

*Handbook of Graphs and Networks: From the Genome to the Internet*, edited by the two physicists Stefan Bornholdt and Heinz Georg Schuster, contains chapters by various scholars in many areas related to graphs and networks. The underlying theme of this book is that networks are important and that they underlie a wide range of phenomena. In a way, the primary message of this book is that a network is just that—a network. In other words, this book encourages readers to think at the unit of analysis of a network rather than of an individual actor. People are exposed to physical networks (highway traffic), biological networks (genome), ecological networks (food web), and virtual networks (the Internet). This fascinating juxtaposition of networks from different domains encourages thought about what is new with the Internet and about the intersection of concepts across network-related domains. Do the same principles apply in traffic networks and on the Internet? Are epidemic models applicable to message dissemination on the Internet? These are the questions that this book provokes. Indeed, one chapter compares yeast molecular networks to the Internet, finding that yeast protein interaction networks are dissimilar because they are more densely interconnected.

Readers will inevitably be drawn to Alan Kirman’s chapter on economic networks. Kirman provides a survey of economic research that emphasizes the microlevel interac-

tions among economic agents. Instead of talking about a market in the abstract, Kirman encourages thought about how trading relationships evolve and about the nature of exchange among economic actors in a network. The chapter begins with a discussion of exogenous and predefined network structures and moves on to the more interesting area of evolving random networks.

Kirman describes the case of a simplex with three sellers and multiple buyers. Buyers are represented as points within an equilateral triangle. The centroid of this triangle represents equal loyalty to all three sellers. Kirman then discusses how this may evolve over time under different conditions. He presents provocative simulation results in which many buyers quickly become loyal to one seller and gravitate to a vertex of the triangle. However, some buyers remain close to the center. These results are interesting, and future studies should envisage methods to track such evolutionary behavior in a lab setting.

Readers will also be fascinated by Dorogovtsev and Mendes's chapter, which identifies the phenomenon of network acceleration and its impact on network structure. Many networks exhibit accelerated growth (i.e., "the mean number of connections per vertex increases with time" [p. 318]). The authors provide a model that allows for nonlinear growth of networks, and they show that in these cases, acceleration affects the structure of the network, which leads to highly asymmetric outcomes. Therefore, their model fits well with the nature of the World Wide Web, which is extremely top-heavy.

Weisbuch and Solomon's chapter titled "Social Percolators and Self-Organized Criticality" discusses the idea that in a network in which people have incomplete information, they will learn by exchanging information with others with whom they are connected (i.e., neighbors in a network). This model can be helpful in modeling person-to-person communications in the marketing literature. The contribution may well be to model the tatonnement process. However, this chapter suffers from a problem that is common to

many chapters in this book; namely, the authors use simulated data. Researchers should view this as an opportunity to extend their results in marketing domains.

The final chapter by Jain and Krishna examines autocatalytic networks and how networks self-organize and evolve over time. The authors identify three phases of growth: the random phase, the growth phase, and the organized phase. They also study the impact of perturbations on networks.

In summary, all three books have something valuable to offer to the reader who is interested in building competence in data mining. I encourage readers to view this body of work through the lens of the authors rather than trying to evaluate the books using criteria developed in the marketing modeling literature. The authors of these books are trying to make sense of the immense data available in online environments and to build better technology. Marketing readers can benefit from this approach and adapt the findings of the books to marketing problems. Exposure to the innards of cutting-edge technologies should help the marketing discipline by focusing research questions and areas of inquiry. Data mining as a research methodology and as an area of managerial practice is bound to grow. Marketing scholars will certainly have a major input in the way this discipline evolves, and these books are a good starting point.

SANDEEP KRISHNAMURTHY  
*University of Washington Bothell*

#### REFERENCES

- Huberman, B.A., P.L.T. Pirolli, J.E. Pitkow, and R.M. Lukose (1998), "Strong Regularities in World Wide Web Surfing," *Science*, 280 (5360), 95-97.
- Montgomery, A.L. (2001), "Applying Quantitative Marketing Techniques to the Internet," *Interfaces*, 30 (2), 90-108.
- Zaiane, Osmar R., M. Xin, and J. Han (1998), "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," in *Proceedings of the Advances in Digital Libraries Conference*. Washington, DC: IEEE Computer Society, 19-29.