Book Review

An introduction to Search Engines and Web Navigation. By Mark Levene. Addison Wesley Publisher, 392 pages, 2006. ISBN: 0321306775, US$97.5.

Effective searching and navigating on the Web (World Wide Web) have become important to various users in recent years, since the Web has already become the largest and most important information resource for extremely diverse applications and users. We can also easily observe the ever-increasing utilization of the Web for the presentation and exchange of information in daily life and the ability of an individual or organization to publish almost anything on the Web at very little cost. Without the technological advancements in search engines and navigation tools within Web browsers, using the huge Web resource would become almost impossible, since it is always growing in size and evolving in structure complexity.

This book attempts to introduce the readers to the subjects of Web searching and navigation and covers comprehensive issues in the rapidly growing field of Web technology. In its ten chapters, the book presents very readable materials on subjects from the background of search engines as early as the primitive hypertext prototype, the Bush's *memex* machine, dating back to 1945, to the state-of-the-art technologies for Web searching on mobile phones. In terms of its organization, the book is divided into three parts. The first part (Chapters 1 to 3) presents an account of existing search engines and highlights the underlying problems of Web interactions such as the surprising low Web coverage by existing Web search engines. The second part (Chapters 4 to 6) discusses the Web search engines and includes an evaluation of search quality and the details of search mechanisms. The third part (Chapters 7 to 9) looks at the more advanced issues, such as mobile and social network technologies in the context of search and navigation. In addition to these three parts, the last chapter offers a fascinating glimpse into the future prospects of search and navigation on the Web.

We now review each part of the book in detail as follows:

The first part of the book considers the following three topics: historical materials and the Web basics, the problems of searches, and the problems of Web navigation. Chapter 1 introduces the historical *memex* machine, which was the first hypertext design based on the important ideas of associate indexing and cross referencing of information. The chapter presents an interesting account of how the early ideas are related to today's Web system. The connections drawn are interesting, but this chapter could be further strengthened with the addition of better explanations of the figures, especially for those obtained from old information sources, which are in fact blurry. Chapter 2 is a particularly well-written chapter. It presents not only the basics of the Web, for example the intriguing bow-tie Web structure and some interesting statistics related to Web size and Web usage, which definitely help the reader to gain a deeper understanding of the current Web's state; but it also analyses the problems of information seeking and Web search. For example, we are given a clear explanation for why conventional information retrieval techniques, which are successful in library information searching, are not appropriate solutions for Web searching. The author also further clarifies the difference between search and navigation, and between local search and global search. Chapter 3 discusses the infamous navigation problem of "getting lost in hyperspace" and the application of machine learning techniques to improve search navigation tools. The issues in this chapter are appealing to IT scholars, since the author explains the application of some advanced computing technologies, such as the naïve Bayes classifier and the Markov chain model. However, it seems that the author may be slightly ambitious in including so many sophisticated methods and models in a single chapter. The content of this chapter may be relatively difficult to those having with weak computing backgrounds. Overall, the first part of the book gives sufficient background information to understand the general problems of search and navigation on the Web. It does a good job in motivating the readers to continue reading the book.

The second part essentially covers the architectural aspects of search engines. Chapter 4 presents major search engines and describes underlying components, such as the roles, functionalities and structures of a crawler and an indexer. The story of the competition among the three search engines of Google, Yahoo! and MSNSearch is interesting, from which we gain an understanding of the interaction between Web technology and the business environment. The downside of the chapter is that in the query engine part, many forward references to Chapters 5 and 6 are made, which might be confusing to readers. Chapter 5 gets down to the details of how a search engine actually works. For example, the questions of how to determine a page reference from the content and how to make use of hyperlinks to access important information on a Web page are answered in excellent detail. This is a clearly written chapter that provides very rich information on various metrics used in a search engine. The author also explains link-based metrics, such as PageRank, and the advantages and minefields associated with using link analysis to score Web pages in commercial applications. Chapter 6 is a particularly well-written chapter that organizes and presents the underlying ideas of a wide spectrum of search engines, from meta-search engines to special purpose search engines. The author undoubtedly recognises the importance and potential of search engine personalization. Thus, the discussion of the issues of Web personalized results, privacy and scalability, relevance feedback and personalized searches is well organized in this chapter. Other interesting issues, such as question answering and image search, are also included. Overall, this part of the book is able to provide the readers with fruitful knowledge of Web searching and the operations of a search engine.

The third part focuses on Web navigation and presents two advanced topics on the mobile Web and social networks. Chapter 7 contains useful details on browser navigation tools, such as a history list and the structural analysis of a Web site. However, the discussion of Web data mining as part of the chapter is inadequate. For example, the information on Web usage mining for personalization is too brief. Chapter 8 explains the different settings of the stationary Web and the mobile Web. The chapter discusses the implications of many technological challenges for search engines arising from the difference in these settings, which are currently hot research issues. It is interesting to see the comparison of various display constraints in mobile devices and the impacts of such constraints related to browsing in standard machines, Pocket PCs and mobile phones.

Chapter 9 is an interesting chapter. It focuses on high-level networks over the Web, such as social networks and P2P networks. The author also discusses how these networks give rise to new computing technologies such as collaborative filtering, managing Weblogs (Blogs) and so on. Overall, this part is a good attempt to guide our vision beyond the fundamental Web searching and navigation issues. One might appreciate and understand better the significance of Web development in the coming decades after reading this last section of the book.

This book has provided an excellent account of the main issues of the scope of searching and navigating over the Web. It is possible to learn from this book how Web size, page content, link structures and user preferences impact on the design and development of search engines and navigation tools. It is also possible to learn how the dynamic nature of the Web has posed serious challenges to search engines, which aim to cover as large a portion of the Web as possible. Finally, the reader gains understanding on why search engines should have scalable architectures, intelligent scheduling strategies, efficient update algorithms for ranking metrics, personalized Web searching technologies, and so on.

Although the book covers and analyses the topic well, it does not address in sufficient depth the challenging problem concerning the ability for a search engine to understand the structure and semantics of XML data, which is recognized to be a key feature for next-generation engines. The author presents only a brief discussion of this topic in Chapter 10. Regarding various personalization categories for tackling the diversified user interests, existing personalized search techniques mostly assume that the user’s interests are persistent. It is thereby feasible to learn a user’s current interests from a search history as discussed in Chapter 6. However, user interests often change. The problem of how to model a user’s changing interests over time is still a challenge as well as an opportunity for future developments. In addition, the problem of how to make more sophisticated and personalizable ranking functions is an interesting area deserving future study.

To sum up, the book is well written and most of the chapters should be comprehensible to those with or without strong IT backgrounds. The book also provides extensive materials related to Web search engines and navigation, including references and Web sources. The book’s organization is highly reader-friendly, since all the chapters are accompanied by a large number of illustrations and examples. Each chapter includes adequate pointers such as a listing of objectives at the very beginning and an excellent summary at the end. The questions and exercises make the book a suitable as a textbook for undergraduates or postgraduates.

Wilfred Ng
Department of Computer Science,
Hong Kong University of Science and Technology,
Clear Water Bay Road, Hong Kong
Email: wilfred@cs.ust.hk