Journal of Intelligent & Fuzzy Systems 16 (2005) 75
IOS Press

Book review

Mining the Web: Discovering Knowledge from Hypertext Data, Soumen Chakrabarti and Morgan Kaufmann, San Francisco, 2003, US$ 57.95, ISBN 1 558 60 754 4

This unique and comprehensive book provide a clear and in-depth introduction to mining the web. It starts from the very beginning, with the brief descriptions of HTML and http in Section 2.1, and goes into detailed description of web crawlers (Chapter 2): what are they, what algorithms they use, what are the problems that they face (traps, duplicate pages, need to refresh, etc.), and ideas of how to design a crawler. Chapter 3 describes the main problems and algorithms for web search:

- how to appropriately compress the stored page indices so as to make search faster,
- how to estimate the degree to which the page is relevant to the query,
- how to eliminate near duplicates:
  - duplicate individual pages – by the closeness of their contents, and
  - duplicate websites – by closeness of the corresponding graphs,

  and
- how to go from simple queries to complex Boolean queries.

This complete Part I of the book.

Part II describes different algorithms for (supervised) learning and (unsupervised) clustering, including fuzzy clustering algorithms, and how these algorithms can be used (and have been used) in mining the web.

Part III starts with a description of Google's PageRank algorithm and other similar algorithms that rank webpages based, crudely speaking, on how authoritative they are, i.e., on how many other pages link to these pages. After describing the basic algorithms, the author spends quite some time explaining how these basic algorithms need to be modified to avoid "nepotism" when several sites cite each other without being really authoritative.

To make algorithms like crawlers or PageRank faster, we must know the structure of the web. The results of statistical analysis of the World Wide Web as a graph are described in Section 7.6. These are the results that led to a well-known description of web as a "small world" graph in which the dependence of the number of pages with k links on k is described by a power law,

Chapter 8 provides additional ideas and algorithms related to the resource discovery on the web: efficient algorithms for graph-based search, algorithms that discover communities (closely related subgraphs of related nodes), etc.

Finally, Chapter 9 describes the directions of current research in web mining, from more efficient natural language processing that would enable the computer to understand queries typed in in natural language, to making search engines adjustable to the individual who uses them.

This book is a must for everyone who is interested in knowing how crawlers and search engines work and how we can use different statistics-based techniques for data mining.

It is a very good research book, and it can also be used as a textbook for advanced students; actually, not much of programming skill is required to understand this book, but the knowledge of matrix algebra and basic statistics is needed.

Vladik Kreinovich
Book Review Editor
Journal of Intelligent & Fuzzy Systems