Mining the World Wide Web: An Information Search Approach

by George Chang, Marcus J. Healey (Editor), James A. M. McHugh, Jason T. L. Wang

Kluwer Academic Publishers, 2001
192 pages, list price $115
ISBN: 0-7923-2349-9

Review by:
Aris Ouksel, University of Illinois at Chicago
Department of Information and Decision Sciences, 2411 University Hall
aris@uic.edu
http://tigger.uic.edu/~aris/

Chang et al. present information retrieval from distributed data sources with a focus on the World Wide Web (WWW). The specific topics explored include: keyword and query based search engines; mediators and wrappers; multimedia search engines; data, text, and web mining; and web crawling agents. A few tools and some related technologies are introduced for each of these topics, but the descriptions are very brief and only scan the surface.

The book is primarily intended as a supplement for upper-level undergraduate or graduate courses in data mining, databases, and information retrieval and then as a reference manual for practitioners. However it falls short on the first objective. The coverage lacks both depth and comprehensiveness on some important conceptual issues. At best it surveys very superficially a handful of techniques. Its usefulness as a supplemental reading book would have required the inclusion of either a set of problems and/or research questions at the end of each section to help students in investigating those issues cursorily presented in the main textbook. At the same time, the main concepts utilized in each technique should have received better coverage.

The book meets the second objective. It can indeed very well serve as a quick reference manual for practitioners, who by the nature of their work have limited time to investigate in some depth new and future developments in their specific or related fields. This book gives them an opportunity to see what may be coming down the pipeline in the area of distributed sources of information on the web.

The book is divided into three parts: information retrieval on the web, data mining on the web, and a case study in environmental engineering. The first two chapters, which cover keyword and query based search engines, are excellent. The material is succinct and written in a pro-active mode, which does stimulate a reader's interest. Then a chapter on mediators and wrappers follows. In my view, its coverage has a lot of room for improvement, as the presentation does not do justice to the extensive research that took place in this specific topic in the last few years. The chapter on multimedia search engines could have actually been put in an appendix without significantly altering the flow in the book. Beyond a terse description of these engines, the rest of the chapter was hurriedly written and the coverage is poor. While multi-media databases are still emerging,

there is wide literature available in recent conference proceedings and in special-issue journals. These sources have not been included in the material.

Section two begins with a chapter on data and text mining. The coverage of the two issues however is far from being comprehensive. The authors attempted to summarize a field, which can stand on its own, in twelve pages. For excellent coverage on data mining and machine learning, please see Tom Mitchell's work *Machine Learning*, it still remains as one of the flagship texts in the field. I personally enjoyed the coverage of the last two chapters on web mining and web crawling agents. The details were presented in an easy to grasp manner and were comprehensive. Though the relevance of chapter three is a puzzle. The authors describe a case study of a system created at their labs, called "EnviroDaemon". While it is a web-based knowledge gathering system, its relevance and connection to the concepts presented is minimal. The readers would be better served if the author's developed the case in an expository manner, illustrating the main concepts and associating them to the various chapters. The list of references is however comprehensive and relevant.

The writing style and the presentation of material definitely are to be commended. The book is an easy read for individuals with an intermediate level knowledge in databases and data structures. Novices would definitely require some background to appreciate some of the technical and computational details. Writing style is kept concise throughout the text, the tone being relatively informal.

In conclusion, the merits of the book may lie in its value to practitioners who need to be kept up-to-date about recent developments in web based knowledge discovery. It is also an easy read for people with database and data structure backgrounds. The core topics are those on keyword and query based searches, web mining, and web crawling agents. One can envision another version of this book, with some of the changes following the remarks, which would be extremely useful to upper-undergraduate and graduate students.