Preface

Web mining is moving the World Wide Web toward a more useful environment in which users can quickly and easily find the information they need. It includes the discovery and analysis of data, documents, and multimedia from the World Wide Web. Web mining uses document content, hyperlink structure, and usage statistics to assist users in meeting their information needs.

The Web itself and search engines contain relationship information about documents. Web mining is the discovery of these relationships and is accomplished within three sometimes overlapping areas. Content mining is first. Search engines define content by keywords. Finding contents' keywords and finding the relationship between a Web page's content and a user's query content is content mining. Hyperlinks provide information about other documents on the Web thought to be important to another document. These links add depth to the document, providing the multi-dimensionality that characterizes the Web. Mining this link structure is the second area of Web mining. Finally, there is a relationship to other documents on the Web that are identified by previous searches. These relationships are recorded in logs of searches and accesses. Mining these logs is the third area of Web mining.

Understanding the user is also an important part of Web mining. Analysis of the user's previous sessions, preferred display of information, and expressed preferences may influence the Web pages returned in response to a query.

Web mining is interdisciplinary in nature, spanning across such fields as information retrieval, natural language processing, information extraction, machine learning, database, data mining, data warehousing, user interface design, and visualization. Techniques for mining the Web have practical application in m-commerce, e-commerce, e-government, e-learning, distance learning, organizational learning, virtual organizations, knowledge management, and digital libraries.

BOOK OBJECTIVE

This book aims to provide a record of current research and practical applications in Web searching. This includes techniques that will improve the utilization of the Web by the design of Websites, as well as the design and application of search agents. This book presents this research and related applications in a manner that encourages addi-

tional work toward improving the reduction of information overflow that is so common today in Web search results.

AUDIENCE

Researchers and students in the fields of information and knowledge creation, storage, dissemination, and retrieval in various disciplines will find this book a starting point for new research. Developers and managers of Web technologies involved with content development, storage, and retrieval will be able to use this book to advance the state of the art in Web utilization.

ORGANIZATION OF THE BOOK

In any Web search effort the user wants information. To find information the search engine must be able to understand the intent of the user and the intent of the Web page author. Chapter I, "Metadata Management: A Requirement for Web Warehousing and Knowledge Management" by Gilbert Laware, brings into focus why Web mining is important and what is important to Web mining.

Understanding the user's intent leads to personalization of search. Personalization can influence user search results, and through collaboration, the results of others. Personalization spans the divisions of Web mining. Chapter II, "Mining for Web Personalization" by Penelope Markellou, Maria Rigou, and Spiros Sirmakessis, provides an introduction to this important area. Other chapters discussing personalization issues are Chapters III, V, XI, XIII, XIV, XVII, XVIII, and XIX.

In keeping with the three primary areas of Web mining the remaining chapters are organized in sections on content, structure, and usage.

Section II presents content mining. Content mining extracts and compares concepts from the content of Web pages and queries. Information retrieval techniques are applied to unstructured, semi-structured, and structured Web pages to find and rank pages in accordance with the user's information need.

In Chapter III, "Using Context Information to Build a Topic-Specific Crawling System," Fan Wu and Ching-Chi Hsu discuss several important criteria to measure the relevancy of Web pages to a given topic. This includes rank ordering the pages based on a relevancy context graph.

This is followed by "Ontology Learning from a Domain Web Corpus" from Roberto Navigli (Chapter IV), which describes a methodology for learning domain ontologies from a specialized Web corpus. This methodology extracts a terminology, provides a semantic interpretation of relevant terms, and populates the domain ontology automatically.

Chapter V, **"**MARS: Multiplicative Adaptive Refinement Web Search" by Xiannong Meng and Zhixiang Chen, applies a new multiplicative adaptive algorithm for user preference retrieval to Web searches. This algorithm uses a multiplicative query expansion strategy to adaptively improve and reformulate the query vector to learn the user's information preference.

Chapter VI, Neil C. Rowe's "Exploiting Captions for Web Data Mining," presents an indirect approach to indexing multimedia objects on Web pages by using captions.

This novel approach is useful because text is much easier for search engines to understand than multimedia, and captions often express the document's key points.

"Towards a Danger Theory Inspired Artificial Immune System for Web Mining" by Andrew Secker, Alex A. Freitas, and Jon Timmis is Chapter VII. As part of a larger project to construct a dynamic Web content mining system, the Artificial Immune System for E-mail Classification (AISEC) is described in detail.

Chapter VIII, "XML Semantics" by Yasser Kotb, Katsuhiko Gondow and Takuya Katayama, introduces a novel technique to add semantics to XML documents by attaching semantic information to the XML element tag attributes. This approach is based on the same concept used by attribute grammars in attaching and checking static semantics of programming languages.

Existing classifiers built on flat files or databases are infeasible for classifying large data sets due to the necessity for multiple passes over the original data. "Classification on Top of Data Cube" by Lixin Fu, Chapter IX, gives a new approach for classification by designing three new classifiers on top of a data cube for both transactional and analytical purposes.

Structure mining, Section III, is concerned with the discovery of information through the analysis of Web page in and out links. This type of analysis can establish the authority of a Web page, help in page categorization, and assist in personalization.

In Chapter X, "Data Cleansing and Validation for Multiple Site Link Structure Analysis," Mike Thelwall provides a range of techniques for cleansing and validating link data for use in Web structure mining. He uses multiple site link structure analysis to mine patterns from themed collections of Websites.

Mohamed Salah Hamdi's "Extracting and Customizing Information Using Multi-Agents," Chapter XI, discusses the challenge of complex environments and the information overload problem. To cope with such environments and problems, he customizes the retrieval system using a multi-agent paradigm.

The Web is a graph that needs to be navigated. Chapter XII, "Web Graph Clustering for Displays and Navigation of Cyberspace" by Xiaodi Huang and Wei Lai, presents a new approach to clustering graphs, and applies it to Web graph display and navigation. The approach takes advantage of the linkage patterns of graphs, and utilizes an affinity function in conjunction with k-nearest neighbor.

Section IV on usage mining applies data mining and other techniques to discover patterns in Web logs. This is useful when defining collaboration between users and refining user personal preferences.

Web usage mining has been used effectively as an approach to automatic personalization and as a way to overcome deficiencies of traditional approaches such as collaborative filtering. Chapter XIII, "Integrating Semantic Knowledge with Web Usage Mining for Personalization" by Honghua Dai and Bamshad Mobasher, discusses the issues and requirements for successful integration of semantic knowledge from different sources, including the content and the structure of Websites. Presented is a general framework for fully integrating domain ontologies with Web usage mining and personalization processes using preprocessing and pattern discovery.

Search engine logs not only keep navigation information, but also the queries made by users. In particular, queries to a search engine follow a power-law distribution, which is far from uniform. Queries and user clicks can be used to improve the search engine user interface, index performance, and results ranking. Ricardo Baeza-Yates in Chapter XIV, "Web Usage Mining in Search Engines," presents these issues.

Efficient mining of frequent traversal path patterns, that is, large reference sequences of maximal forward references from very large Web logs, is a fundamental problem in Web mining. Chapter XV, "Efficient Web Mining for Traversal Path Patterns" by Zhixiang Chen, Richard H. Fowler, Ada Wai-Chee Fu, and Chunyue Wang, discusses two new algorithms to solve this problem.

"Analysis of Document Viewing Patterns of Web Search Engine Users," Chapter XVI by Bernard J. Jansen and Amanda Spink, discusses viewing patterns of Web results pages and Web pages by search engine users. This chapter presents the advantages of using traditional transaction log analysis in identifying these patterns.

Chapter XVII, "A Java Technology Based Distributed Software Architecture for Web Usage Mining" by Juan M. Hernansáez, Juan A. Botía, and Antonio F.G. Skarmeta reviews a technique of each of the Web usage mining approaches: clustering, association rules, and sequential patterns. These techniques are integrated into the learning architecture, METALA.

In Chapter XVIII, "Web Usage Mining: Algorithms and Results," Yew-Kwong Woon, Wee-Keong Ng, and Ee-Peng Lim focus on mining Web access logs, analyzing algorithms for preprocessing and extracting knowledge from such logs, and proposing techniques to mine the logs in a holistic manner.

We conclude the book with one chapter in Section V, a summary of Web mining and personalization and their application in another part of the Internet. Chapter XIX, "The Scent of a Newsgroup: Providing Personalized Access to Usenet Sites through Web Mining," by Giuseppe Manco, Riccardo Ortale, and Andrea Tagarelli discusses the application of Web mining techniques to the problem of providing personalized access to Usenet services. It focuses on the analysis of the three main areas of Web mining techniques: content mining, in which particular emphasis is posed to topic discovery and maintenance; structure mining, in which the structure of Usenet news is studied by exploiting structure mining techniques; and usage mining, in which techniques for tracking and profiling users' behavior are devised by analyzing access logs. The chapter ends with an overview of personalization techniques, and the description of a specific personalization method to the case of Usenet access.