

Guest Editors' Introduction: Special Section on Mining and Searching the Web

Bing Liu and Soumen Chakrabarti

WITH the phenomenal growth of the Web, there is an ever-increasing volume of information being published on numerous Web sites. This vast amount of accessible information has raised many new opportunities and challenges for knowledge discovery and data engineering researchers. For programs that seek to analyze Web content, the heterogeneity in authorship and the consequent lack of structure are formidable hurdles. Discovering and extracting novel and useful knowledge from Web sources call for innovative approaches that draw from a wide range of fields spanning data mining, machine learning, statistics, databases, information retrieval, artificial intelligence, and natural language processing.

In Web search, although general-purpose search engines are very useful, finding specific or targeted information can still be a frustrating experience. Highly effective, domain-specific, and personalized search techniques are not yet mainstream. In e-commerce, a whole range of online techniques are also needed to support such applications. For example, in online shopping, there are no human shop assistants to help customers. Instead, automated techniques are needed to learn from the behaviors of users in order to provide effective recommendations and assistance. Mining, extracting, and integrating Web information are challenging problems as well because there is still no mature technique to integrate information from structured (stored database), ad hoc structured (shopping sites), and unstructured (product reviews) sources. Clearly, format standards for semistructured data will not solve all of these problems.

This special issue of *IEEE Transactions on Knowledge and Data Engineering* brings together some of the latest research results in the field. It presents seven papers which deal with a wide range of problems. All of the accepted papers propose some novel and/or principled techniques to solve these problems. Of the seven papers, three focus on domain specific and personalized Web search, one proposes a principled technique for collaborative filtering, one studies Web page cleaning for identifying informative structures and content blocks in Web pages, one studies classification of Web pages based on positive and unlabeled training examples, and one studies the clustering of XML data for efficient storage and querying of such data.

- B. Liu is with the Department of Computer Science, University of Illinois at Chicago, 851 South Morgan Street, Chicago, IL 60607. E-mail: liub@cs.uic.edu.
- S. Chakrabarti is with the Computer Science and Engineering Department, Indian Institute of Technology, Bombay, Powai, Mumbai 400076 India. E-mail: soumen@cse.iitb.ac.in.

For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number 01-092003.

The first paper by Michelangelo Diligenti, Marco Gori, and Marco Maggini studies Web page scoring for Web search and resource discovery. Current methods for the purpose are mainly based on the analysis of hyperlinks. The structure of the hyperlinks is the result of collaborative activities of the community of Web authors. Web authors usually like to link resources they consider authoritative, and authority emerges from the dynamics of popularity of the resources on the Web. This paper proposes a general probabilistic framework based on random walk of links for Web page scoring that incorporates and extends many existing models. Their results show that the proposed framework is effective and is particularly suited for focused or vertical search.

The second paper by Satoshi Oyama, Takashi Kokubo, and Toru Ishida describes an interesting technique for domain specific Web search. The basic idea is to find a set of domain specific keywords (which the authors call *keyword spices*) that can be used as the context of the search queries in the domain. A nice algorithm based on text classification is given for identifying a reasonably complete set of such keyword spices. To perform text classification, it collects training pages from the Web through a search using an initial set of keywords of the domain. The main advantage of the proposed method is that it does not need to collect and index domain specific pages as most domain specific search engines do. The work is also related to research in query expansion and modification, but deals with a slightly different problem and offers different approaches.

The third paper by Fang Liu, Clement Yu, and Wei-Yi Meng also studies Web search, more specifically, personalized Web search. Since general-purpose search engines do not consider user's interests, their search results may not be interesting to a specific user. Personalized search aims at carrying out search for each user incorporating his/her interests. In this paper, the authors propose to employ a user profile and a general profile to constrain the search. The user profile is learned from the user's search history, which contains the user interested categories and weighted terms in the categories. The general profile is built using the categories from the Open Directory Project. The key advance of the technique is that it maps each user query to some categories. At the search time, the system first uses the profiles to infer the categories of the search terms in question. Then, the search terms are augmented with each category as the context to perform search. The search results are then merged to produce a single result ranking. A comprehensive experimental evaluation is described in the paper.

The fourth paper by Hung-Yu Kao, Shian-Hua Liu, Jan-Ming Ho, and Ming-Syan Chen focuses on the cleaning of

Web pages from news Web sites. In other words, it mines informative link structures and informative content blocks of Web pages. It is well-known that commercial Web pages typically contain a large amount of redundant or irrelevant information, such as company logos, navigation panels, advertisements, copyright, and privacy notices. Although such information items are functionally useful for human viewers and important to the Web site owners, they often hamper automated information gathering, Web data mining, and Web search. In their paper, Kao et al. propose an entropy-based method to remove redundant or irrelevant links and contents from the pages of news Web sites in order to find informative structures and contents blocks of the pages. Their experimental results on a number of news sites demonstrate the effectiveness of the technique.

The fifth paper by Kai Yu, Anton Schwaighofer, Volker Tresp, Xiaowei Xu, and Hans-Peter Kriegel considers the problem of collaborative filtering and recommender systems, which are commonly used in online shopping sites for recommending products and services to users based on the ratings and behaviors of other like-minded users. Yu et al. propose a principled approach to collaborative filtering in a probabilistic framework. Most existing methods are based on heuristics. The proposed framework uses the memory-based model and has a number of nice techniques for dealing with the problems of memory-based collaborative filtering, e.g., computational efficiency and accommodating new data and new users.

The sixth paper by Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang studies the problem of the classification of Web pages based on only positive examples and an unlabeled set. This is different from traditional classification, which requires both labeled positive and negative training examples. The key feature of this problem is that there is no labeled negative document for learning, which makes traditional classification techniques inapplicable. Empirical study of the problem only started recently. Yu et al. propose an SVM-based technique with a preprocessing step to build classifiers based on only positive and unlabeled data. Their experimental results demonstrate the effectiveness of the technique.

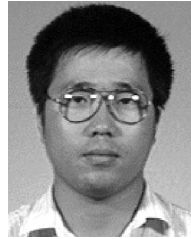
The final paper by Wang Lian, David W. Cheung, Nikos Mamoulis, and Siu-Ming Yiu studies clustering of XML documents. As XML is becoming the standard data representation for interoperability over the Internet, managing XML documents requires urgent study. Clustering XML documents helps to manage XML data in terms of both storage and querying. For example, similar XML documents can be stored together and, thus, can be queried more efficiently. In their paper, Lian et al. propose an effective and efficient clustering technique to group XML documents according to their structures. The key to the proposal is a new similarity metric that can be computed efficiently and also produces accurate clustering results.

In summary, the seven papers represent some of the latest and most promising research results in the new and exciting field of Web mining and Web search, which continues to make significant impact on real-world applications. We are

confident that this special issue will stimulate further research in this area.

Finally, we would like to thank all the reviewers who have helped to review more than 60 submissions. Without their generous assistance, this special section would not be possible.

Bing Liu
Soumen Chakrabarti
Guest Editors



Bing Liu received the PhD degree from the University of Edinburgh, U.K. He is an associate professor of computer science at the University of Illinois at Chicago (UIC). His research interests include data mining, Web and text mining, and bioinformatics. Before joining UIC, he was with the National University of Singapore. Since 1996, Dr. Liu has been active in data mining research and published extensively in leading conferences and journals (e.g., AAAI, IJCAI, SIGKDD, WWW, ICML, and IEEE transactions) in the areas of data mining, machine learning, and artificial intelligence. He is on the editorial boards of three international journals, including the *IEEE Transactions on Knowledge and Data Engineering*. He has also served (serves) on the technical program and/or organizing committees of many international conferences.



Soumen Chakrabarti received the PhD degree from the University of California, Berkeley. He is an associate professor of computer science and engineering at the Indian Institute of Technology (IIT), Bombay. Prior to joining IIT, he worked on hypertext information retrieval and data mining at the IBM Almaden Research Center. He has developed several systems for Web mining, including IBM's CLEVER search and Focused Crawler, published extensively, and acquired eight US patents on his inventions to date. Dr. Chakrabarti has served as a vice-chair or program committee member for many conferences, including WWW, SIGIR, ICDE, VLDB, KDD, and SODA.