Search and Browse Log Mining for Web Information Retrieval: Challenges, Methods, and Applications

Daxin Jiang
(Microsoft Research Asia, Beijing, China,
djiang@microsoft.com)
Jian Pei
(Simon Fraser University, Burnaby, BC
Canada, jpei@cs.sfu.ca)
Hang Li
(Microsoft Research Asia, Beijing, China,
hangli@microsoft.com)

Abstract: Huge amounts of search log data have been accumulated in various search engines. Currently, a commercial search engine receives billions of queries and collects tera-bytes of log data on any single day. Other than search log data, browse logs can be collected by client-side browser plug-ins, which record the browse information if users' permissions are granted. Such massive amounts of search/browse log data, on the one hand, provide great opportunities to mine the wisdom of crowds and improve search results as well as online advertisement. On the other hand, designing effective and efficient methods to clean, model, and process large scale log data also presents great challenges.

In this tutorial, we focus on mining search and browse log data for Web information retrieval. We consider a Web information retrieval system consisting of four components, namely, query understanding, document understanding, query-document matching, and user understanding. Accordingly, we organize the tutorial materials along these four aspects. For each aspect, we will survey the major tasks, challenges, fundamental principles, and state-of-the-art methods.

The goal of this tutorial is to provide a systematic survey on large-scale search/browse log mining to the IR community. It will help IR researchers to get familiar with the core challenges and promising directions in log mining. At the same time, this tutorial may also serve the developers of Web information retrieval systems as a comprehensive and in-depth reference to the advanced log mining techniques.

ACM Categories & Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: Search and browse logs, log data mining

Bios: ***Daxin Jiang*** is a Researcher at Microsoft Research Asia. His research focuses on data mining and information retrieval. He received Ph.D. in computer science from the State University of New York at Buffalo. He has published extensively in prestigious conferences and journals, and served as a PC member of many conferences. He received the Best Application Paper Award of SIGKDD'08 and the Runner-up for Best Application Paper Award of SIGKDD'04.

Jian Pei is an Associate Professor and the Associate Director, Research, of the School of Computing Science, Simon Fraser University. His research focuses on data mining and analytic queries in databases. With prolific publications in refereed journals and conferences, he is the recipient of several prestigious awards. He is an associate editor of ACM Transactions on Knowledge Discovery from Data (TKDD) and IEEE Transactions on Knowledge and Data Engineering (TKDE). He has served regularly in the organization committees and the program committees of numerous international conferences and workshops. He is a senior member of both ACM and IEEE.

Hang Li is a Senior Researcher and Research Manager at Microsoft Research Asia. His research areas include natural language processing, information retrieval, statistical machine learning, and data mining. He graduated from Kyoto University and holds a PhD in computer science from the University of Tokyo. Hang has about 80 publications in international conferences and journals. He is associate editor of ACM Transaction on Asian Language Information Processing and area editor of Journal for Computer and Science Technology, etc. His recent academic activities include senior PC member of SIGIR 2010, WSDM 2010, and KDD 2010, area chair of ACL 2010, and PC member of WWW 2010.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.