Enhancing Web Search by Mining Search and Browse Logs

Daxin Jiang
Microsoft Research Asia
No. 5 Danling Street
Haidian, Beijing, China
djiang@microsoft.com

Jian Pei
Simon Fraser University
8888 University Drive
Burnaby, BC, Canada
jpei@cs.sfu.ca

Hang Li
Microsoft Research Asia
No. 5 Danling Street
Haidian, 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 web search results. 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 will focus on mining search and browse log data for search engines. We will start with an introduction of search and browse log data and an overview of frequently-used data summarization in log mining. We will then elaborate how log mining applications enhance the five major components of a search engine, namely, query understanding, document understanding, query-document matching, user understanding, and monitoring and feedbacks. For each aspect, we will survey the major tasks, fundamental principles, and state-of-the-art methods. Finally, we will discuss the challenges and future trends of log data mining.

The goal of this tutorial is to provide a systematic survey on large-scale search/browse log mining to the IR community. It may 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.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Algorithms, Experimentation

Keywords

Search and browse logs, log mining, search applications

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

Presenter Bios

Daxin Jiang, Ph.D., Researcher, Microsoft Research Asia. Daxin Jiang's research focuses on information retrieval and log data mining. 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. Daxin Jiang has been working on development of Microsoft search engines, including Live Search and Bing.

Jian Pei, Ph.D., Associate Professor, Associate Director, School of Computing Science, Simon Fraser University. Jian Pei's research focuses on data mining and analytic queries on various data repositories. With prolific publications in refereed journals and conferences, he is the recipient of several prestigious awards. He is the Associate Editor-in-Chief of IEEE Transactions on Knowledge and Data Engineering (TKDE), and an Associate Editor of 4 premier journals on data mining and analytics, including ACM Transactions on Knowledge Discovery from Data (TKDD). He has served regularly in the organization committees and the program committees of numerous international conferences and workshops, such as a PC co-chair of the ICDM 2011 conference. He is a senior member of both ACM and IEEE.

Hang Li, Ph.D., Senior Researcher and Research Manager, Microsoft Research Asia. Hang Li's 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 PC co-chair of WSDM 2011, senior PC member of WSDM 2010, senior PC member of KDD 2010, area chair of ACL 2010, and PC member of WWW 2010. Hang has been working on development of several products. These include NEC TopicScope, Microsoft SQL Server 2005, Microsoft Office 2007, Microsoft Live Search 2008, and Microsoft Bing 2009.