Web Search/Browse Log Mining: Challenges, Methods, and Applications

WWW'10 Tutorial Overview

Daxin Jiang
Microsoft Research Asia
49 Zhichun Road
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
49 Zhichun Road
Haidian, Beijing, China
hangli@microsoft.com

ABSTRACT

Huge amounts of search and browse log data has been accumulated in various search engines. Such massive search/browse log data, on the one hand, provides great opportunities to mine the wisdom of crowds and improve Web search as well as online advertisement. On the other hand, designing effective and efficient algorithms and tools to clean, model, and process large scale log data presents great challenges.

In this tutorial, we give a systematic survey on the applications, challenges, fundamental principles and state-of-the-art methods of mining large scale search and browse log data. We start with an introduction of search and browse log data and an overview of various log mining applications. Then, we focus on four popular areas of log mining applications, namely query understanding, document understanding, query-document matching, and user understanding. For each area, we review the major tasks, analyze the challenges, and exemplify several representative solutions. Finally, we discuss several new directions in search/browse log mining.

The tutorial slides are available at the authors' homepages after the tutorial is presented.

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 and advertisement applications

1. INTRODUCTION

Huge amounts of search log data has been accumulated in various search engines. Nowadays, a commercial search engine receives billions of queries and collects tera bytes of log data on any single day. Moreover, many commercial search engine companies provide toolbars at the client side and collect browse information if users' permissions are granted. Such massive search/browse log data, on the one hand, provides great opportunities to mine the wisdom of crowds and improve Web search as well as online advertisement. On the other hand, designing effective and efficient algorithms and tools to clean, model, and process large scale log data presents great challenges.

In this tutorial, we focus on mining search and browse log data for Web search applications. We consider a Web search system consisting of four components, namely query understanding, document understanding, query-document matching, and user understanding. Accordingly, we organize the tutorial materials in these four areas. For each area, 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 WWW community. We only assume that the audience has the basic concepts of probability and statistics. We do not assume that the audience has deep background knowledge about statistics, sampling, probability, or any other mathematical principles. We use sufficient examples to explain the ideas and intuitions.

Our tutorial is designed to serve three categories of audience. First, researchers working on search/browse log mining or the highly related problems will find our tutorial a good summary and analysis of the state-of-the-art methods and a stimulating discussion on the core challenges and promising directions. Particularly, for researchers planning to start the investigation in this direction, the tutorial can serve as a short introduction course which leads them to the frontier quickly.

Second, general data mining audience will find the tutorial informative. They can get a global picture of the current research on search/browse log data mining. Moreover, researchers in other fields who need to tackle problems of similar nature can quickly understand the on-the-shelf techniques that they can borrow to solve their problems.

Third, industrial search engine practitioners will find our tutorial a comprehensive and in-depth reference to the advanced log mining techniques. This tutorial will serve as a bridge between the research frontier and the industrial practice. The ideas and solutions introduced in this tutorial may motivate the search engine developers to turn research fruits into product reality.

Copyright is held by the author/owner(s).
WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

2. OUTLINE OF THE TUTORIAL

1. Introduction
   (A) **Introduction to search/browse logs**: general formats of search/browse logs, differences between the two types of logs.
   (B) **Overview of log mining applications**: a taxonomy of log mining applications, four popular areas (query understanding, document understanding, query-document matching, and user understanding) and important tasks.
   (C) **Basic data structures of search/browse logs**: query histogram, click-through bipartite, click patterns, and session patterns; construction algorithms, data pre-processing techniques.
2. Query Understanding by Log Mining
   (A) **Basic statistics of queries**: distributions of query lengths, topics, and frequencies; evolvement of distributions over time and across regions; query statistics on different search devices (e.g., computers, iphones, and conventional mobile phones).
   (B) **Query categorization**: categorizing queries into navigational, informational, or transactional ones; classifying queries into a pre-defined set of topics.
   (C) **Query expansion, substitution, and suggestion**: common features of various approaches to these three tasks, and differences.
   (D) **Temporal and spatial features of queries**: different metrics used to create temporal and spatial profiles of queries, evolvement of query volumes and topics over time, event detection, semantic similarity. identification, and search precision prediction using temporal features of queries; geographic distribution of queries; localizable queries.
   (E) **Extracting aspects and entities from queries**: latest approaches; the differences from aspect and entity extraction from documents.
3. Document Understanding by Log Mining
   (A) **URL annotation**: term-level annotation, query-level annotation, URL annotation by tag data and query logs.
   (B) **Document summarization**: document summarization using user clicks.
   (C) **Search result clustering**: clustering search results, increasing diversity of answers.
   (D) **Evaluating page and site importance**: leveraging user browse patterns to identify pages/sites favored by users.
4. Query-Document Matching by Log Mining
   (A) **Mining preference pairs**: observations from eye-tracking experiments, generating training examples from search log data for learning to rank search results.
   (B) **Mining preference lists**: user clicks as implicit feedback, position bias in user clicks for Web search, different assumptions about the user behavior on browsing search results, various click models, effectiveness and efficiency.
   (C) **Predicting click-through rate of ads**: search advertising and contextual advertising, click-through rate estimation methods in three aspects: the type of log data used, the features to describe ads, and the models to predict clicks.
5. User Understanding by Log Mining
   (A) **Analyzing variations in user behavior**: variation of search and browse behavior between different users and different queries, influence of user interface factors to user behavior (e.g., snippets of search results).
   (B) **Personalized search**: three questions: how much search logs can help personalized search; whether all queries need to be personalized; how to personalize search results using log data.
   (C) **User segmentation**: segmentation by search history, segmentation by browse history, and segmentation by a combination of demographic profiles and search/browse history.
   (D) **Modeling user behavior**: frequent sequential patterns, hidden Markov model, and Bayesian network.
6. Summary: Challenges and Future directions:
   (A) **Challenges**: noise filtering, privacy preserving, parallelization of large-scale log data mining, etc.
   (B) **Future directions**: context-aware search, structured search, OLAP on log data, etc.

3. ACKNOWLEDGMENTS

Jian Pei's research has been supported in part by an NSERC Discovery grant and an NSERC Discovery Accelerator Supplements grant. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.