Enhancing Web Search through Query Log Mining

Ji-Rong Wen
Microsoft Research Asia, China

INTRODUCTION

Web query log is a type of file keeping track of the activities of the users who are utilizing a search engine. Compared to traditional information retrieval setting in which documents are the only information source available, query logs are an additional information source in the Web search setting. Based on query logs, a set of Web mining techniques, such as log-based query clustering, log-based query expansion, collaborative filtering and personalized search, could be employed to improve the performance of Web search.

BACKGROUND

Web usage mining is an application of data mining techniques to discovering interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Since the majority of usage data is stored in Web logs, usage mining is usually also referred to as log mining. Web logs can be divided into three categories based on the location of data collecting: server log, client log, and proxy log. Server log provides an aggregate picture of the usage of a service by all users, while client log provides a complete picture of usage of all services by a particular client, with the proxy log being somewhere in the middle (Srivastava, Cooley, Deshpande, & Tan, 2000).

Query log mining could be viewed as a special kind of Web usage mining. While there is a lot of work about mining Website navigation logs for site monitoring, site adaptation, performance improvement, personalization and business intelligence, there is relatively little work of mining search engines' query logs for improving Web search performance. In early years, researchers have proved that relevance feedback can significantly improve retrieval performance if users provide sufficient and correct relevance judgments for queries (Xu & Croft, 2000). However, in real search scenarios, users are usually reluctant to explicitly give their relevance feedback. A large amount of users' past query sessions have been accumulated in the query logs of search engines. Each query session records a user query and the corresponding pages the user has selected to browse. Therefore, a query log can be viewed as a valuable source containing a large amount of users' implicit relevance judgments. Obviously, these relevance judgments can be used to more accurately detect users' query intentions and improve the ranking of search results.

One important assumption behind query log mining is that the clicked pages are "relevant" to the query. Although the clicking information is not as accurate as explicit relevance judgment in traditional relevance feedback, the user's choice does suggest a certain degree of relevance. In the long run with a large amount of log data, query logs can be treated as a reliable resource containing abundant implicit relevance judgments from a statistical point of view.

MAIN THRUST

Web Query Log Preprocessing

Typically, each record in a Web query log includes the IP address of the client computer, timestamp, the URL of the requested item, the type of Web browser, protocol, etc. The Web log of a search engine records various kinds of user activities, such as submitting queries, clicking URLs in the result list, getting HTML pages and skipping to another result list. Although all these activities reflect, more or less, a user's intention, the query terms and the Web pages the user visited are the most important data for mining tasks. Therefore, a query session, the basic unit of mining tasks, is defined as a query submitted to a search engine together with the Web pages the user visits in response to the query.

Since the HTTP protocol requires a separate connection for every client-server interaction, the activities of multiple users usually interleave with each other. There are no clear boundaries among user query sessions in the logs, which makes it a difficult task to extract individual query sessions from Web query logs (Cooley, Mobasher, & Srivastava, 1999). There are mainly two steps to extract query sessions from query logs: user identification and session identification. User identification is the process of isolating from the logs the activities associated with an

individual user. Activities of the same user could be grouped by their IP addresses, agent types, site topologies, cookies, user IDs, etc. The goal of session identification is to divide the queries and page accesses of each user into individual sessions. Finding the beginning of a query session is trivial: a query session begins when a user submits a query to a search engine. However, it is difficult to determine when a search session ends. The simplest method of achieving this is through a timeout, where if the time between page requests exceeds a certain limit, it is assumed that the user is starting a new session.

Log-Based Query Clustering

Query clustering is a technique aiming at grouping users' semantically (not syntactically) related queries in Web query logs. Query clustering could be applied to FAQ detecting, index-term selection and query reformulation, which are effective ways to improve Web search. First of all, FAQ detecting means to detect Frequently Asked Questions (FAQs), which can be achieved by clustering similar queries in the query logs. A cluster being made up of many queries can be considered as a FAQ. Some search engines (e.g. Askjeeves) prepare and check the correct answers for FAQs by human editors, and a significant majority of users' queries can be answered precisely in this way. Second, inconsistency between term usages in queries and those in documents is a well-known problem in information retrieval, and the traditional way of directly extracting index terms from documents will not be effective when the user submits queries containing terms different from those in the documents. Query clustering is a promising technique to provide a solution to the word mismatching problem. If similar queries can be recognized and clustered together, the resulting query clusters will be very good sources for selecting additional index terms for documents. For example, if queries such as "atomic bomb", "Manhattan Project", "Hiroshima bomb" and "nuclear weapon" are put into a query cluster, this cluster, not the individual terms, can be used as a whole to index documents related to atomic bomb. In this way, any queries contained in the cluster can be linked to these documents. Third, most words in the natural language have inherent ambiguity, which makes it quite difficult for user to formulate queries with appropriate words. Obviously, query clustering could be used to suggest a list of alternative terms for users to reformulate queries and thus better represent their information needs.

The key problem underlying query clustering is to determine an adequate similarity function so that truly similar queries can be grouped together. There are mainly two categories of methods to calculate the similarity between queries: one is based on query content, and the other on query session. Since queries with the same or similar search intentions may be represented with different words and the average length of Web queries is very short, content-based query clustering usually does not perform well.

Using query sessions mined from query logs to cluster queries is proved to be a more promising method (Wen, Nie, & Zhang, 2002). Through query sessions, "query clustering" is extended to "query session clustering". The basic assumption here is that the activities following a query are relevant to the query and represent, to some extent, the semantic features of the query. The query text and the activities in a query session as a whole can represent the search intention of the user more precisely. Moreover, the ambiguity of some query terms is eliminated in query sessions. For instance, if a user visited a few tourism Websites after submitting a query "Java", it is reasonable to deduce that the user was searching for information about "Java Island", not "Java programming language" or "Java coffee". Moreover, query clustering and document clustering can be combined and reinforced with each other (Beeferman & Berger, 2000).

Log-Based Query Expansion

Query expansion involves supplementing the original query with additional words and phrases, which is an effective way to overcome the term-mismatching problem and to improve search performance. Log-based query expansion is a new query expansion method based on query log mining. Taking query sessions in query logs as a bridge between user queries and Web pages, probabilistic correlations between terms in queries and those in pages can then be established. With these term-term correlations, relevant expansion terms can be selected from the documents for a query. For example, a recent work by Cui, Wen, Nie, and Ma (2003) shows that, from query logs, some very good terms, such as "personal computer", "Apple Computer", "CEO", "Macintosh" and "graphical user interface", can be detected to be tightly correlated to the query "Steve Jobs", and using these terms to expand the original query can lead to more relevant pages.

Experiments by Cui, Wen, Nie, and Ma (2003) show that mining user logs is extremely useful for improving retrieval effectiveness, especially for very short queries on the Web. The log-based query expansion overcomes several difficulties of traditional query expansion methods because a large number of user judgments can be extracted from user logs, while eliminating the step of collecting feedbacks from users for ad-hoc queries. Log-based query expansion methods have three other important properties. First, the term correlations are pre-