Chapter XIX
Processing and Analysis of Search Query Logs in Chinese

Michael Chau
The University of Hong Kong, Hong Kong

Yan Lu
The University of Hong Kong, Hong Kong

Xiao Fang
The University of Toledo, USA

Christopher C. Yang
Drexel University, USA

Abstract

More non-English contents are now available on the World Wide Web and the number of non-English users on the Web is increasing. While it is important to understand the Web searching behavior of these non-English users, many previous studies on Web query logs have focused on analyzing English search logs and their results may not be directly applied to other languages. In this Chapter we discuss some methods and techniques that can be used to analyze search queries in Chinese. We also show an example of applying our methods on a Chinese Web search engine. Some interesting findings are reported.

Introduction

Search engines have been widely used for finding useful information on the World Wide Web. Many users start their Web activities using popular search engines such as Google (http://www.google.com). Many search engines were originally designed for English Web pages. While English content is popular on the Web, the number of Web users speaking other languages is increasing rapidly (Global Reach, 2004). To satisfy the information needs of non-English speaking users,
most large commercial search engines support multilingual searching. However, the underlying technologies used in these search engines may not be the best technique for searching non-English documents.

For instance, in mainland China, Google’s share of the search engine market is smaller than that of the market leader Baidu (http://www.baidu.com), a search engine tailor-made for Chinese Web pages (CNNIC, 2006). Because of such reasons, there are many language-specific search engines designed for particular languages. For example, the search engine Fireball (http://www.fireball.de) was designed for German Web pages, Goo (http://www.goo.ne.jp) for Japanese, and Ayna (http://www.ayna.com) for Arabic. The information needs and search behavior of non-English users are different from those of native English users because of different languages and different cultures (Chau et al., 2007). More importantly, some languages, such as Asian languages, have different characters, grammars, and structures that are significantly different from those of English. Consequently, the methods and techniques for processing search logs in these languages can be quite different from those for processing English search logs. In this chapter, we discuss methods and issues involved in processing search logs in Chinese. As one of the most widely used non-English languages, Chinese has its unique characteristics. On the other hand, it shares similar characteristics with some other Asian languages such as Japanese and Korean. We believe that we can extend methods in this chapter across these languages.

The chapter is structured as follows. In the next section, we give some background knowledge about the characteristics of the Chinese language. Then we discuss the methods and techniques used to analyze Chinese search queries. The section that follows presents the application of our methods on a Chinese Web search engine called Timway. The last section provides a summary of this chapter.

**BACKGROUND**

The analysis of search engine logs can be classified into the area of Web usage mining. Study on search engine logs usually has focused on how users use the search engines on the Web to find the information they need. On the other hand, it also has a strong root in information retrieval research. Before the Web became popular, many studies had reported analysis of user information behavior, search queries, and search sessions with various information retrieval and digital library systems (e.g., Fenichel, 1981; Bates et al., 1993). In recent years, we have seen many studies devoted to search engines and information systems on the Web. The first category of Web search engine log research focused on analyzing the search logs submitted to general-purpose search engines. In 1998, Jansen, Spink, and several others started a series of research projects on the search logs that were made available by Excite. Their first study analyzed a set of 51,473 queries submitted to the Excite search engine in 1997 (Jansen et al., 1998; Jansen et al., 2000). Subsequently, they expanded their research and analyzed three sets of data collected in 1997, 1999 and 2001, each containing at least one million queries submitted to the Excite search engine (Spink et al., 2001; 2002; Wolfram et al., 2001). Many interesting findings have been identified from these search logs, such as the trends in Web searching (Spink et al., 2002), sexual information searching on the Web (Spink & Ozmultu, 2002), and Web queries in question format (Spink & Ozmultu, 2002). Another large-scale Web query analysis was performed by Silverstein et al. (1999) on a set of 993 million requests submitted to the AltaVista search engine over a period of 43 days in 1998. Most of these studies used a set of similar metrics or statistics in their studies, including number of sessions, number of queries, number of queries in a session, number of terms in a query, percentage of queries using Boolean queries, number of result pages viewed by each user, etc. These metrics allow researchers