Web Searching in Chinese: A Study of a Search Engine in Hong Kong

Michael Chau
School of Business, University of Hong Kong, Pokfulam, Hong Kong. E-mail: mchau@business.hku.hk

Xiao Fang
College of Business Administration, University of Toledo, Toledo, OH 43606. E-mail: xiao.fang@utoledo.edu

Christopher C. Yang
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong. E-mail: yang@se.cuhk.edu.hk

The number of non-English resources has been increasing rapidly on the Web. Although many studies have been conducted on the query logs in search engines that are primarily English-based (e.g., Excite and AltaVista), only a few of them have studied the information-seeking behavior on the Web in non-English languages. In this article, we report the analysis of the search-query logs of a search engine that focused on Chinese. Three months of search-query logs of Timway, a search engine based in Hong Kong, were collected and analyzed. Metrics on sessions, queries, search topics, and character usage are reported. N-gram analysis also has been applied to perform character-based analysis. Our analysis suggests that some characteristics identified in the search log, such as search topics and the mean number of queries per sessions, are similar to those in English search engines; however, other characteristics, such as the use of operators in query formulation, are significantly different. The analysis also shows that only a very small number of unique Chinese characters are used in search queries. We believe the findings from this study have provided some insights into further research in non-English Web searching.

Introduction

The World Wide Web has become a major information resource. More people are routinely looking for useful information on the Web. When users need to search for information on the Web, they often use Web search engines such as Google (http://www.google.com) and MSN (http://search.msn.com). Many users begin their Web activities by submitting a query to a search engine. While most Web contents are in English, there are increasingly more Web pages that are authored in other languages. Besides, many Web users are not native English speakers; some even do not know English at all. It has been estimated that only around 36.5% of Internet users are native English speakers (Global Reach, 2004). Consequently, it is common for users to perform Web searches for non-English resources. While most large commercial search engines such as Google can handle queries in multiple languages, there also are many search engines that have been designed for particular languages. Examples include the search engine Fireball (http://www.fireball.de) for German Web pages, Goo (http://www.goo.ne.jp) for Japanese, and Ayna (http://www.ayna.com) for Arabic. Although some of these language-specific search engines also support multilingual searching, they are often tailor-made for the respective language by utilizing some language-specific indexing techniques and focusing on relevant contents to achieve better performance.

As the number of non-English searches on the Web increases, it is important to study users' Web-searching behavior for non-English information using language-specific search engines. Such research is important to improving the design of search engines and understanding users' information-searching behavior. One of the most popular approaches is to analyze the log files of search engines. This kind of server-side log data is easy to collect without any extra effort from the users and can record visits of every single user who has visited the Web page or search engine of interest. Several commercial and research projects have reported on the analyses of such data for various Web applications. For example, studies on the Excite query logs have been widely published (Jansen, Cunningham, & McNam, 1998; Jansen, Spink, & Saracevic, 2000; Ross & Wolfram, 2000; Spink, Jansen, Wolfram, & Saracevic, 2002; Spink, Wolfram, Jansen, &

Received June 6, 2005; revised August 5, 2006; accepted August 6, 2006

 ● Published online 2 April 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20592

Saracevic, 2001). Analysis on the AltaVista query logs also has been reported (Silverstein, Henzinger, Marais, & Moricz, 1999). Similar studies on peer-to-peer Gnutella network also have been conducted (Kwok & Yang, 2004; Yang & Kwok, 2005); however, most of these studies focused on English search queries only. The information needs and search behaviors of users of non-English search engines can be very different from those of English search engines because of the different nature (e.g., grammar) of these languages and also the different cultures of non-English users. It would be interesting to compare the various metrics across search engines designed for different languages.

In this article, we report our study on the search-query logs collected from a Hong Kong search engine called Timway. A study of Chinese search logs would be an interesting area because of the different language structure between Chinese and English (especially because Chinese is an ideographical, character-based language as opposed to the alphabetical, word-based nature of English). This could result in many different searching behaviors such as the average number of terms or characters used in a query. The rest of the article is structured as follows. First, we review related research in Web mining and search engine log analysis. Next, we pose our research questions. We then discuss the data and the methods used in this research. The following section presents and discusses the findings of our analysis. Finally, we summarize our study and suggest some future directions.

Related Work

Search engines have been widely used to locate useful information on the World Wide Web. When a user submits a query to a search engine, the query is usually stored by the search engine into its log file, along with other data such as the user's IP address or a timestamp. These search logs often contain a large number of queries submitted by a large number of users. The study of search logs is often categorized under the research area of Web mining, which was defined by Etzioni (1996) as the use of data-mining techniques to automatically discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web. In a broader definition, Web mining research covers the use of data mining as well as other similar techniques to discover resources, patterns, and knowledge from the Web and other Web-related data such as Web usage data or Web server logs. Web mining research can be classified into three categories: Web content mining, Web structure mining, and Web usage mining (Chen & Chau, 2004; Kosala & Blockeel, 2000). Web content mining refers to the discovery of useful information from Web contents, including text, images, audio, video, and so on. Web content mining research includes resource discovery from the Web (e.g., Chakrabarti, van den Berg, & Dom, 1999; Chau, Zeng, Chen, Huang, & Hendriawan, 2003), document categorization and clustering (e.g., Chen, Fan, Chau, & Zeng, 2001; Kohonen et al., 2000; Zamir & Etzioni, 1999), and information extraction from Web pages (e.g., Hurst, 2001). Web structure mining studies the model underlying the link structures of the Web. It usually involves the analysis of in-links and out-links information of a Web page, and has been used for search engine result ranking and other Web applications (Brin & Page, 1998; Kleinberg, 1998). Web usage mining focuses on using data-mining techniques to analyze search logs or other activity logs to find interesting patterns. One application of Web usage mining is to learn user profiles (e.g., Armstrong, Freitag, Joachims, & Mitchell, 1995). Data such as Web traffic patterns or usage statistics also can be extracted from Web usage logs to improve the performance of a Web site (e.g., Cohen, Krishnamurthy, & Rexford, 1998; Fang & Sheng, 2004).

The mining of search engine logs belongs to 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 satisfy their information needs. 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-libraries systems (e.g., Bates, Wilde, & Siegfried, 1993; Fenichel, 1981). Since the Internet evolution, 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 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 datasets collected in 1997, 1999, and 2001, each containing at least 1 million queries submitted to the Excite search engine (Spink et al., 2001; Spink et al., 2002; Wolfram, Spink, Jansen, & Saracevic, 2001). Many interesting findings have been identified from these search logs, such as trends in Web searching (Spink et al., 2002), sexual-information searching on the Web (Spink, Ozmultu, & Lorence, 2004), and question-format Web queries (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, and so on. These metrics allow researchers to compare their findings across different types of search engines at different times.

The second category of Web search analysis focused on the search logs of a specific Web site or system. Compared to the widely published search-log analysis studies on general-purpose search engines, only a few studies have been reported in this category. One example is the study by Croft, Cook, and Wilder (1995), which investigated the search queries submitted to the THOMAS system, an online searchable

DOI: 10.1002/asi

database consisting of U.S. legislative information. Croft et al. analyzed 94,911 queries recorded in the system and identified the top 25 queries. They also found that 88% of all queries contained three or fewer words, a number much lower than that of traditional information-retrieval systems. Jones et al. (1998) analyzed the transaction logs of the New Zealand Digital Library that contained a collection of computer science technical reports. They obtained similar results regarding the number of words in queries: Almost 82% of the queries were composed of three or fewer words. Their study also found that most users use the default settings of search engines without any modifications. Chau, Fang, and Sheng (2005) studied the search queries submitted to the search engine on a government Web site, and Wang et al. (2003) analyzed the search queries submitted to the search engine of a university. In both studies, it was demonstrated that seasonal patterns exist in Web site searching. They also found that the search queries submitted to a general-purpose search engine were quite different from that submitted to a Web site search engine, in terms of search topics, search-term distribution, the mean number of queries per session, and the mean length of queries.

All Web search analysis studies discussed earlier have focused on search-log data that contain primarily English queries. Only a couple of studies focused on non-English search engine data; one of them is the Fireball study (Hölscher, 1998; Hölscher & Strube, 2000). The dataset contains about 16 million queries and 27 million non-unique terms collected from Fireball. Some summary statistics such as the average length of queries, the use of Boolean operators, and the use of phrase searching were discussed in the article. Their study showed that the average length of queries is only 1.66 words, which is much shorter than those identified in English search engines such as Excite ($n = 2.52$) or AltaVista ($n = 2.02$). This may be due to the differences between the English language and the German language (which allows compound words to be created relatively easily). However, as described by Jansen and Pooch (2000), there was limited discussion on query terms in the Fireball study.

Another search engine log study on non-English data was performed by the Academia Sinica in Taiwan on Chinese search engines. Instead of studying users' information needs and searching behaviors from the search logs, they utilized the query logs to provide term suggestion to users. In their study, Huang and colleagues (2003; Huang et al., 2001) analyzed the query logs from two Chinese search engines in Taiwan, namely GAIS (http://gais.cs.ccu.edu.tw) and Dreamer (no longer available). They also proposed a method to extract search sessions and search queries from proxy server log data. Their query log contained the search requests submitted to several general-purpose Web search engines in Taiwan, including Yahoo-Taiwan (http://tw.yahoo. com), Sina-Taiwan (http://www.sina.com.tw), PChome (http://www.pchome.com.tw), and Yam (http://www.yam.com). A total of 2,369,282 queries and 218,362 unique terms was collected in a period of 126 days. They also found that 74% of their search sessions contained only one query, which was similar to the number in the AltaVista study (Silverstein et al., 1999). Similar to the Fireball study, the major drawback of the study of Huang and colleagues is that only limited statistics were provided; there was no in-depth analysis of query terms and search topics.

Pu, Chuang, and Yang (2002) also performed analysis on Chinese search-log data. In their study, they analyzed the query logs from three search engines in Taiwan, namely Dreamer, GAIS, and Openfind (http://www.openfind.com.tw). Pu et al. reported some basic characteristics of the queries collected. They found that the average length of the Chinese queries in their logs was 3.18 characters, which was longer than that in English. They also reported that advanced search functions were seldom used, and that less than 5% of queries covered almost three fourths of the total frequencies (Pu et al., 2002). However, because their study focused on comparing the performance of human categorization and machine categorization of queries into different subjects, Pu et al. provided only basic statistics on the query logs, but not detailed analysis such as session/query analysis and character-level analysis.

Commercial search engines also publish their top search queries periodically. For example, Lycos 50 (http://50.lycos.com/) lists the top people, places, and other things which are most searched on the Lycos search engine. Another example is the Google Zeitgeist (http://www.google.com/press/zeitgeist.html), which shows the top queries submitted to Google and some interesting patterns such as seasonal patterns and top celebrity queries. Other search engines (e.g., Yahoo, MSN, AskJeeves) also publish their top queries. The strength of these analyses is that they are based on large search engine logs submitted to these popular search engines. Some of these datasets are not limited to a single day but extend over a period of several years. However, there are two major problems that limit the implications of these analyses. First, only data on search queries are presented; they do not reveal other patterns such as search-term usage and metrics such as the mean length of queries or the mean number of queries per session. Second, most of these studies performed filtering in their analyses. For example, categorical terms (e.g., "news" and "music"), queries on Internet tools (e.g., "Windows" and "mp3") pornographic queries, and foul language are not included in the analysis of Lycos 50. As such, many interesting patterns are not revealed in these studies.

Research Questions

As discussed earlier, it is interesting and important to study the searching behavior and the information needs in non-English search engines; however, previous in-depth search-log analysis studies have focused on search engines such as Excite and AltaVista, and most queries analyzed were in English. These findings may not be applicable to non-English search engines, which have their own characteristics and groups of users.

In this study, we try to address the following research questions: (a) What are the characteristics of the search

DOI: 10.1002/asi

queries submitted to a non-English search engine? (b) How do these queries compare to those of English search engines such as Excite and AltaVista? (c) What are the implications of these results on the design of non-English search engines?

Data and Methods

In this study, we captured the search queries submitted to the Timway search engine (http://www.timway.com). Timway is a search engine that was established in 1997 and is primarily designed for searching Web sites in Hong Kong. Timway indexes Web pages in both English and Chinese, and accepts search queries in both languages. A screenshot of the home page of the search engine is shown in Figure 1.

The query-log data were collected over a period of about 3 months, from December 1, 2003 to March 2, 2004. A total of 1,255,633 records comprised the search-log data. Each record in the log file represents a search query sent from a user to the search engine. A record consists of four fields: the search query, the number of hits, the user's IP address, and a timestamp. The search query is the text entered by the users in the search-query box on the Timway search engine Web site. The number of hits reveals how many relevant search results were found in Timway's database and returned to the users. The third field shows the Internet address from which the search query was received. Finally, the timestamp records the date and time when the query was received by the search engine. An example of our query log is shown in Table 1.

There are several characteristics of our data. First, as Timway accepts search queries in both Chinese and English, our log data contain queries in both languages. While the data used in other search engine studies also include some non-English data (e.g., Silverstein et al. (1999)), the proportion of those data is often insignificant; however, in our study, the number of queries in both languages is comparable. In our analysis, we define three types of queries:

TABLE 1. Example of a record in the search log data.

Field	Value
Query	中學
No. of hits	583
Client IP Address	158.182.xxx.xxx (hidden for privacy reasons)
Timestamp	Dec 1 00:00:26

- Pure English query: a query that contains only ASCII characters
- Pure Chinese query: a query that contains only double-byte characters
- Mixed query: a query that contains both ASCII and double-byte characters

Note that these definitions also would treat any Chinese queries that contain symbols or Arabic numbers typed in ASCII as mixed queries. In the present article, we will use these definitions for our analysis. The distribution of the languages used in the queries is shown in Table 2.

TABLE 2. Statistics on the languages used in the search queries.

Language	Queries	%
English	641,169	51.06
Chinese	536,814	42.75
Mixed (containing both Chinese and English/symbols)	77,650	6.18
Total	1,255,633	100.00

FIG. 1. Timway search engine.

DOI: 10.1002/asi

TABLE 3. Statistics on the encodings used in the Chinese search queries.

Encoding	Queries	%
Chinese (Big 5)	529,094	98.56
Chinese (GB-2312)	7,388	1.38
Chinese (GBK)	332	0.06
All Chinese Queries	536,814	100.00

Another characteristic of our dataset is that the Chinese queries have been received in different character encodings. Because Timway has been designed for Hong Kong users, the default encoding used on its Web site is the Big 5 encoding, the most popular Chinese language encoding scheme used in Hong Kong; however, some of the queries have been submitted using other encodings, including GB-2312 and GBK. As a preprocessing step, we used an open-source Java program to detect the encoding of each query (Peterson, 2004). Statistics of the different encodings of the pure Chinese queries as detected by the program are shown in Table 3.

The algorithm of the detection program was based on the statistics and distribution of common Chinese characters. To test the accuracy of the detection program on the dataset, we did a test on a set of 137 randomly selected Chinese queries from our data. A Chinese native speaker was asked to manually judge whether the encoding detected by the program was correct by looking at each query in both the original encoding scheme and the detected encoding scheme, and then deciding which one was more likely to be a search query. Of the 137 queries, the encoding scheme was incorrectly detected in only one query (0.73%) by the program, thus the program achieved an accuracy of 99.27%. Because there are overlaps in the byte ranges used in the different encodings, there is no algorithm that can guarantee 100% accuracy in the detection. As such, we believe that the high accuracy of the detection program would be sufficient for our analysis. As most of the Chinese queries in our data were originally in Big 5 encoding, we converted all queries in GB-2313 and GBK encodings into the Big 5 encoding (using the same program) to facilitate further analysis.

Analysis

In this section, we present the results of the analysis on data collected from the Timway search engine. There are 1,255,633 queries in total. Among these queries, there are 536,814 Chinese queries, 641,169 English queries, and 77,650 mixed queries.

Sessions, Queries, and Topics

The metrics developed by Spink et al. (2001) were used in this study. *Queries* are sets of one or more terms. *Unique queries* are all differing queries entered by one user in a session. *Repeat queries* are all multiple occurrences of the same query by one user or submitted by the system automatically to periodically update the list of results. Multiple occurrences of the same query also can be generated by the system when the user requests subsequent result pages. *Empty queries* are defined as queries without any terms.

TABLE 4. Statistics on queries.

Total queries	1,255,633 (100.00%)
Unique queries	1,033,182 (82.28%)
Repeat queries	222,451 (17.72%)
Empty queries	0 (not captured in the Timway log file)
Mean queries per session	2.03
Median queries per session	1
Mean unique queries per session	1.67
Median unique queries per session	1

Table 4 presents the occurrences and distribution of different types of queries. Of the 1,255,633 queries, 1,033,182 (82.28%) are unique queries, and 222,451 (17.72%) are repeat queries. Since empty queries were not captured in the Timway search log, we were not able to report the total here. The average number of queries per session is 2.03 ($Mdn = 1$), which is much lower than that ($n = 4.86$) reported in the Excite study (Spink et al., 2001) but very close to that ($n = 2.02$) reported in the AltaVista study (Silverstein et al., 1999). Since our definition of session is similar to that in the AltaVista study (whereas the definition used in the Excite study tends to assume longer sessions), we believe that the mean number of queries per session in our study is comparable to that in other English search engines. Upon further investigating the number of unique queries per session, the average number of unique queries per session is 1.67 ($Mdn = 1$), which is 18.74% lower than the average number of queries per session. Users do not tend to use repeat queries within the same session.

We further investigated the distribution of number of queries per session. There is a total of 671,612 sessions in the log file. We followed the definition of session suggested in Silverstein et al. (1999), which defines a session as a series of queries submitted by a single user within a small range of time. A session represents a set of queries relevant to a user's single information need. In this study, since other client information such as cookie or user id was not available, we used IP address to identify unique users and sessions. Note that IP address is less reliable in identifying unique users because the same user may be assigned two different IP addresses in different sessions, and two different users may share the same IP address. To identify sessions for each IP address (each user), we used a cutoff of 30 min. If a user submitted a query within 30 min from the previous query, these queries would be included in the same session. Otherwise, the second query would be considered the start of a new session.

Table 5 presents the distribution of the number of queries per session (range = 1–20), and Figure 2 shows the distribution of number of queries per session. Over 50% of sessions have only one query. Sessions with fewer than or equal to seven queries constitute over 90% of the sessions.

In addition to query and session studies, we investigated the topics in the top queries of the Timway search engine. Among the top 100 queries, there are 54 English queries and 46 Chinese queries, but no mixed queries. Table 6 lists the

DOI: 10.1002/asi

TABLE 5. Number of queries per session.

Queries per session	Occurrences (%)	Queries per session	Occurrences (%)
1	355,956 (53.00)	11	1015 (0.15)
2	130,726 (19.46)	12	651 (0.10)
3	58,890 (8.77)	13	475 (0.07)
4	29,788 (4.43)	14	347 (0.05)
5	16,179 (2.41)	15	254 (0.04)
6	9,251 (1.38)	16	197 (0.03)
7	5,401 (0.80)	17	151 (0.02)
8	3,439 (0.51)	18	109 (0.02)
9	2,179 (0.32)	19	62 (0.01)
10	1,483 (0.22)	20	63 (0.01)

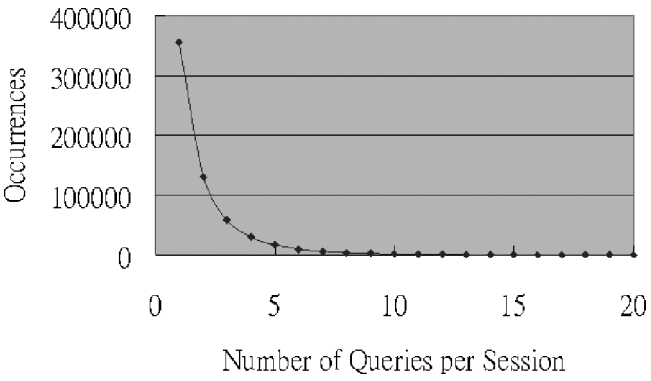

FIG. 2. Number of queries per session.

TABLE 6. Top 25 English queries and top 25 Chinese queries.

	Top 25 English queries	Occurrences (%)	Top 25 Chinese queries	Occurrences (%)
1	sex	25,721 (2.05)	一樓一	6,841 (0.54)
2	141	7,589 (0.60)	成人	4,621 (0.37)
3	sex141	5,619 (0.45)	色情	3,995 (0.32)
4	161	4,915 (0.39)	走光	2,869 (0.23)
5	man161	4,786 (0.38)	衣	2,558 (0.20)
6	bt	3,937 (0.31)	六合彩	2,555 (0.20)
7	mp3	3,840 (0.31)	貼圖	2,147 (0.17)
8	adult	3,740 (0.30)	性	1,473 (0.12)
9	map	3,010 (0.24)	酒店	1,348 (0.11)
10	one floor one	2,644 (0.21)	歌詞	1,289 (0.10)
11	hotel	2,610 (0.21)	一樓一鳳	1,264 (0.10)
12	wallpaper	2,548 (0.20)	學生妹	1,233 (0.10)
13	midi	2,272 (0.18)	絲襪	1,180 (0.09)
14	travel	2,148 (0.17)	情色	1,102 (0.09)
15	sauna	2,047 (0.16)	食譜	991 (0.08)
16	av	2,007 (0.16)	旅行社	990 (0.08)
17	169	2,004 (0.16)	窺	982 (0.08)
18	gay	1,749 (0.14)	拍賣	952 (0.08)
19	sextvb	1,709 (0.14)	地圖	951 (0.08)
20	BT	1,525 (0.12)	中原	896 (0.07)
21	macau	1,506 (0.12)	勞工處	880 (0.07)
22	best161	1,478 (0.12)	手機	877 (0.07)
23	car	1,400 (0.11)	桑拿	862 (0.07)
24	bank	1,287 (0.10)	論壇	850 (0.07)
25	yahoo	1,252 (0.10)	電影	846 (0.07)

top 25 English queries and the top 25 Chinese queries. Over 50% of the top queries either in English or Chinese were related to sexuality. Examples in Chinese include "一樓一" (prostitutes), "成人" (adult), and "色情" (pornography). This phenomenon is the same as the data reported in the search-log studies for other general-purpose search engines such as Excite (Jansen et al., 1998; Spink et al., 2001) and AltaVista (Silverstein et al., 1999). Among the other queries, some are related to traveling such as "map," "hotel," "travel," "macau," "car," "酒店" (hotel), "旅行社" (travel agency), and "地圖" (map). Some are related to electronic commerce such as "yahoo" and "拍賣" (auction). Some are related to computer systems such as "bt" (the BitTorrent software), "BT," "mp3," "wallpaper," and "midi." The others are miscellaneous, such as "bank," "六合彩" (Mark Six, a lottery in Hong Kong), "貼圖" (posted images), "歌詞" (lyrics), "食譜" (recipes), "中原" (Centaland, a real estate agency in Hong Kong), "勞工處" (labor department), and "手機" (mobile phones).

There is no mixed query in the top 100 queries. In Table 7, we present the top 10 mixed queries. We found that some of these mixed queries are names of movies and animations such as "頭文字D" and "外星BB撞地球." Some queries are mixed because the English part of the queries, such as "mp3" and "bt," does not have a popular Chinese translation. In the other queries, the English part or the Chinese part of the mixed queries has translation in another language, but the term is mixed in English and Chinese due to the culture in Hong Kong. For example, "卡拉ok" is originally a Japanese term and can be translated as "karaoke" in English; however, it is commonly translated as "卡拉ok" in Hong Kong.

Character Usage

In most previous studies in English search-log analysis, the use of terms in queries and the distribution of terms have been investigated (Spink et al., 2001; Wang et al., 2003). Such analysis can often reveal the search topics of the users and the use of function words and the co-occurrences of terms in search queries; however, in Chinese, the concept of term is quite different from that in English because Chinese is a character-based language. There is no space between terms in Chinese. For example, the query "旅行社" (travel agency) actually consists of two terms in Chinese "旅行" (travel) and "社" (agency), but it is not easy to tokenize the two terms automatically without advanced algorithms. On the other hand, character-based processing has been widely used in the analysis of Chinese text. Many Chinese information retrieval systems and Chinese processing applications are character-based

TABLE 7. Top 10 mixed queries.

1	頭文字D	6	h漫
2	mp3機	7	外星BB撞地球
3	卡拉ok	8	台灣kiss
4	H漫	9	BT下載
5	AV女優	10	bt下載

DOI: 10.1002/asi

(e.g., Chau, Qin, Zhou, Tseng, & Chen, 2005). Character-based processing methods are often based on a statistical model (e.g., Chien, 1997). Among these methods, character-based bigrams or n-grams analysis is simple and easy to conduct for Chinese text. In bigram analysis, all occurrences of any two consecutive characters are extracted and the frequencies are recorded (Li, Ding, & Tan, 2002). N-gram analysis is similar except that all occurrences of *n* consecutive characters are processed.

Next, we discuss the following characteristics of our data: the most commonly used characters in the queries, the distribution of character usage, and the most frequent n-grams in our data.

Characters used in queries. The mean number of characters used in the pure Chinese queries is 3.380, which is slightly larger than that ($n = 3.18$) reported in Pu et al. (2002) on their Taiwan search engine logs. Our number also is much greater than the mean number of terms used in English queries as reported in previous studies. For example, the mean number of terms per query is 2.16 in Spink et al. (2001) and 2.35 in Silverstein et al. (1999). However, as discussed earlier, terms in English and characters in Chinese cannot be directly compared as they are different linguistic units. For instance, many two-character queries such as "地圖" or "電影" are actually the counterparts of one English term ("maps" and "movies," respectively). Further research will be needed to study the characteristics of query length in Chinese search queries. Nonetheless, we report the numbers here so that comparison will be possible in future research.

We also analyzed the data to study the distribution of characters in Chinese queries. We found that of the 536,814 Chinese queries in our data, there are only 7,303 unique Chinese characters. On the other hand, 140,279 unique English terms were identified in approximately 1 million queries analyzed in the Excite study (Spink et al., 2001).

We identified the top 50 characters that appeared in the search log, as shown in Table 8. One interesting observation is that the top 50 characters constitute 25.25% of the total occurrences of characters in the search log. In other words, about one of every four characters submitted to the search engine would fall within the top 50 list. This finding concurs with previous studies which showed a high degree of usage of the most frequent terms (Jansen et al., 1998; Spink et al., 2001). This number is actually much higher than that reported in the Excite study, in which the top 75 terms constituted only about 9% of the terms occurring in unique queries (Spink et al., 2001).

Besides the fact that Chinese characters and English terms have different linguistic units, one possible reason for these differences is that the number of Chinese characters is relatively small compared with the number of terms in English. This has important implications for the design of Chinese search engines, which we will discuss later in the article.

Another interesting finding is in the study of the occurrences of these top 50 characters in other Internet applications.

TABLE 8. Top 50 characters in (a) the Timway search log data and (b) a Usenet newsgroup corpus.

Timway search log			Usenet newsgroup (Tsai, 1996)		
Character	Occurrences	%	Character	Occurrences	%
人	24,090	1.19	的	6,538,132	3.80
中	19,906	0.98	是	3,200,626	1.86
港	19,841	0.98	不	2,831,612	1.65
香	19,077	0.94	我	2,584,497	1.50
電	18,569	0.92	一	2,542,556	1.48
情	17,492	0.87	有	2,289,333	1.33
成	16,761	0.83	大	1,891,383	1.10
色	15,564	0.77	在	1,715,554	1.00
學	14,616	0.72	人	1,598,855	0.93
小	13,258	0.66	了	1,507,218	0.88
會	12,800	0.63	中	1,322,363	0.77
圖	12,151	0.60	到	1,310,850	0.76
文	11,955	0.59	資	1,115,608	0.65
美	11,265	0.56	要	1,034,142	0.60
大	11,252	0.56	以	994,958	0.58
國	11,239	0.56	可	992,842	0.58
手	11,064	0.55	這	986,130	0.57
機	10,972	0.54	個	933,857	0.54
女	10,630	0.53	你	915,385	0.53
樓	10,461	0.52	會	894,569	0.52
日	9,837	0.49	好	860,232	0.50
網	9,672	0.48	為	847,332	0.49
天	9,233	0.46	上	828,394	0.48
生	9,192	0.45	來	812,950	0.47
子	8,939	0.44	學	806,783	0.47
性	8,859	0.44	就	803,921	0.47
公	8,637	0.43	交	728,005	0.42
影	8,491	0.42	也	712,260	0.41
地	8,340	0.41	用	695,290	0.40
新	8,038	0.40	能	668,264	0.39
星	7,963	0.39	如	659,275	0.38
下	7,385	0.37	時	658,186	0.38
遊	7,145	0.35	文	651,140	0.38
行	7,085	0.35	說	638,724	0.37
明	6,834	0.34	沒	638,184	0.37
金	6,765	0.33	他	635,766	0.37
的	6,677	0.33	看	632,561	0.37
片	6,602	0.33	那	610,340	0.36
車	6,595	0.33	問	601,742	0.35
心	6,483	0.32	生	601,668	0.35
光	6,231	0.31	提	599,147	0.35
理	6,214	0.31	下	589,356	0.34
工	6,151	0.30	過	586,922	0.34
馬	6,066	0.30	請	576,417	0.34
水	5,936	0.29	們	571,155	0.33
物	5,807	0.29	天	569,684	0.33
合	5,714	0.28	所	558,469	0.32
歌	5,629	0.28	多	542,911	0.32
畫	5,573	0.28	麼	535,402	0.31
司	5,514	0.27	小	530,133	0.31

We compared the top 50 characters identified in our data to the top 50 identified in a corpus consisting of Usenet newsgroup articles (Tsai, 1996). There are only 11 overlapping characters in the list (22.0%). This may present an important problem as what people are searching for may be quite different from what is available on the Web. Further research will be needed to study the reason for the difference

DOI: 10.1002/asi

and its effect on issues such as search engine design and effectiveness.

N-gram analysis. In search-log data, because each query is short and generally not a complete sentence, it is difficult to identify context information and linguistic features that are required in many analysis methods. Therefore, we decided to use the n-grams method to analyze our data with different values of *n*. The top n-grams would allow us to identify the most frequent "terms" (each consisting of two or more characters) and thus the most popular topics.

In total, there are 162,678 unique bigrams in our data. The top 50 bigrams are shown in Table 9. Of these bigrams, only 4 are invalid Chinese terms ("六合," "合彩," "生妹," and "限公"). The other 46 bigrams represent some of the most searched topics in the search engine. The top five bigrams are "香港" (Hong Kong), "成人" (adult), "色情" (pornography), "公司" (company), and "下載" (download). Although the term "香港" (Hong Kong) is not in the top 25 queries shown in Table 6, it is the most frequent bigram in the logs. It shows that this term is not frequently used alone as a two-character query, but is often used together with other characters to form a query.

We also found that many invalid bigrams were produced because they are part of a longer term. For example, "合彩" is not a valid term, but is part of a longer term "六合彩" (Mark Six lottery). Therefore, we applied trigram analysis to the search logs as well. The top 25 trigrams are shown in Table 10.

TABLE 9. Top 50 bigrams.

Bigram	Frequency	Bigram	Frequency
香港	17,134	中文	2,416
成人	13,768	鈴聲	2,416
色情	9,922	電話	2,372
公司	5,045	寫真	2,304
下載	4,858	卡通	2,235
貼圖	4,111	旅遊	2,213
日本	4,006	中心	2,209
走光	3,996	賽馬	2,200
中國	3,850	中原	2,116
電影	3,841	歌詞	2,114
情色	3,573	馬會	1,994
酒店	3,441	足球	1,965
遊戲	3,414	討論	1,934
六合[a]	3,374	旅行	1,913
小說	3,372	電子	1,885
漫畫	3,283	生妹[a]	1,879
免費	3,221	網上	1,860
合彩[a]	3,185	有限	1,829
手機	3,109	限公[a]	1,819
明星	3,095	網頁	1,806
電腦	3,082	數碼	1,773
二手	3,036	論壇	1,765
學生	2,876	食譜	1,750
內衣	2,778	日報	1,737
地圖	2,650	世界	1,730

[a]Invalid terms.

TABLE 10. Top 25 trigrams.

Trigram	Frequency
六合彩	3,177
有限公[a]	1,816
限公司[a]	1,815
學生妹	1,814
旅行社	1,631
討論區	1,418
中原地[a]	1,158
情色文[a]	1,141
賽馬會	1,118
圖書館	1,098
貼圖區	1,049
手提電[a]	1,047
百老匯	923
成人小[a]	899
勞工處	899
模擬器	896
領事館	871
成人漫[a]	867
提電話[a]	865
百分百	839
數碼相	795
寫真集	771
新世界	771
碼相機[a]	750
人漫畫[a]	747

[a]Invalid terms.

The top five trigrams are "六合彩" (Mark Six lottery), "有限公" (incomplete term), "限公司" (incomplete term), "學生妹" (young female students), and "旅行社" (travel agencies). Some other valid trigrams include "討論區" (discussion forums), "賽馬會" (jockey club), "圖書館" (libraries), and "勞工處" (labor department). However, there are more invalid terms in this list. Most of these trigrams consist of one valid bigram and an extra character [e.g., "有限公" and "限公司", which together form the quadragram "有限公司" (limited companies)]. To better analyze the search topics in our queries, we extract all n-grams (with $n \geq 3$) from the search log and manually identify the valid ones that occur 500 times or more in the log. The results are shown in Table 11.

Advanced Search Features

The use of advanced search features has been studied in most previous search engine studies (e.g., Chau, Fang, & Sheng, 2005; Jansen et al., 2000; Spink et al., 2001). It is important to study how advanced features such as Boolean operators are used when users formulate Web queries. In the user interface of the Timway search engine, the functions of the operators are not clearly specified. Therefore, it may be difficult for users to decide what operators to use (e.g., "AND" vs. "+"). Note that in Chinese search engines, English words or symbols are often used as operators; operators based on Chinese characters are rare.

We analyzed all Chinese queries and mixed queries in our data to identify the use of operators. In total, 614,464 queries

DOI: 10.1002/asi

TABLE 11. Top *n*-grams with n ≥ 3 and frequency ≥ 500.

n-gram	Frequency
六合彩	3,177
學生妹	1,814
有限公司	1,812
旅行社	1,631
討論區	1,418
賽馬會	1,118
圖書館	1,098
貼圖區	1,049
百老匯	923
勞工處	899
模擬器	896
領事館	871
手提電話	864
百分百	839
數碼相	795
寫真集	771
新世界	771
數碼相機	746
情色文學	736
成人漫畫	726
聊天室	716
夜總會	684
香港賽馬	655
手機鈴聲	654
中原地區	633
心理測驗	630
二手車	622
林心如	599
網頁素材	597
香港賽馬會	594
成人小說	589
成人電影	588
楊千嬅	585
電子書	576
天文台	574
中文大學	569
輸入法	567
工聯會	560
張柏芝	556
情人節	541
倚天屠龍記	541
香港小姐	514
聯交所	505

are either in pure Chinese or contain mixed characters. The analysis results are shown in Table 12. The data reported for the Excite study in Spink et al. (2001) also are listed for comparison. Among the operators, the "+" operator is the most widely used in our queries. This is similar to the finding in Excite where this operator was among the most popular; however, two other popular operators in Excite, namely "AND" and quotation, are not widely used in Chinese searching. Overall, we found that operators are much less used in Chinese searching than they are in the Excite search engine. In our data, only 0.3624% of the Chinese/mixed queries used one or more operators. In other words, the majority of queries do not utilize any such operators. This concurs with the findings of Pu et al. (2002), who reported that less than 1% of the queries in their Taiwan search-engine logs utilized

TABLE 12. Usage of operators in query formulation.

	This study		Excite study	
Feature	Queries	%	Queries	%
AND	241	0.0392	29,146	2.8410
OR	18	0.0029	1,149	0.1120
NOT	4	0.0007	307	0.0299
+ (plus)	1,085	0.1766	44,320	4.3201
− (minus)	258	0.0420	21,951	2.1397
" "	63	0.0103	52,354	5.1032
( )	488	0.0794	(not reported)	(not reported)
Any of the above	2,227	0.3624	(not reported)	(not reported)

advanced features. One possible reason for the low utilization of operators is the language difference. Because the operators are either English words (e.g., AND) or symbols that originated from the Western culture, they fit more naturally with English search terms; however, since Chinese is a character-based language, it may be less natural for users to incorporate these operators in their query-formulation process.

Discussion

Previous Web search studies based on search-log analysis have been conducted mostly on English queries, with only a few exceptions such as the Fireball study on German queries. This study is one of the first that has investigated the characteristics of Web searching in an Asian language. While previous studies have reported similar results for English queries, our current study has identified several important characteristics in Chinese searching. The key findings of this study and their implications on search-engine design and development are:

- Pornographic materials are the most sought after in general-purpose search engines such as Excite or Timway, whether the search media is in English or in Chinese. Online searching for free pornography is a worldwide behavior. Since the proportion of these queries is large, attention needs to be given to these requests in search-engine design (e.g., filtering search results when appropriate).
- The mean number and median number of queries per session in our search log are comparable to those in other studies. This suggests that the number of queries per session may be independent of the language used. In addition, most users submitted only one query per session; they did not modify their queries because they got the pages they wanted on the first attempt, switched to another search engine, or completely gave up. This is an issue that needs to be researched further.
- The mean number of characters used in Chinese queries is 3.380, which is larger than the mean number of terms in English queries as reported in Excite ($n = 2.16$) and AltaVista ($n = 2.35$); however, terms and characters are different and cannot be compared directly. Each Chinese character corresponds to a morpheme (i.e., the smallest meaningful unit in a language) (Norman, 1988). Although these characters can

DOI: 10.1002/asi

stand alone and have their own meanings, most terms in modern Chinese are disyllabic or trisyllabic words, consisting of two or three characters, respectively. Therefore, three Chinese characters may be equivalent to only one or two Chinese terms. In that sense, Chinese queries may be shorter than English queries; however, because there is no clear word boundary in Chinese (e.g., the "space" counterpart in English), it is not easy to automatically analyze the exact number of words in a Chinese query. One possible solution is to use techniques such as mutual information (Chien, 1997; Yang, Yen, Yung, & Chung, 1998) for automatic Chinese-word segmentation. More research is needed on this topic.

- In all the Chinese search queries, there are only 7,303 unique Chinese characters in total, which is much lower than the number of unique terms in English queries. One reason is that Chinese characters are generally bounded to a closed class. New characters are seldom created. It is often suggested that knowledge of 3,000 commonly used characters will be sufficient for basic literacy in Chinese; most of the commonly used words are formed by these characters. Another reason is that Chinese input on computers is based on predefined input methods (e.g., Pinyin or Changjie). Typos in these input methods can result only in a "wrong character" rather than a new character whereas in English typos often result in new words. In fact, the Big 5 encoding only supports about 13,000 characters, which has set the upper bound on the number of Chinese characters in our study, while there is virtually no limit in the possible permutations of the English alphabet. It also is surprising to find that the top 50 Chinese characters already represent one quarter of all the Chinese characters in the search log. This proportion is much higher than that in the Excite search log. While Chinese characters and English words are quite different linguistically, both are often used as indexing units in search engines. It implies that a pure Chinese search engine would have a much smaller number of index items, but each item (character) would have a much longer list of document IDs. Chinese search engines and bilingual search engines should be designed to make use of these characteristics.
- The use of Boolean operators in Chinese query formulation is stunningly infrequent, when compared with previous data found in English search logs. Although the different designs of search engines (e.g., the default search options and the explanation of these operators) may have contributed partially to the low usage of operators in our data, it appears that language differences also have played an important role. For example, the operators "AND," "OR," and "NOT" are English words. While it is natural to include them in English queries, it is not the case in Chinese. Similar reasons may apply to symbol operators such as "+," "−," and quotations. We suppose that these operators are less used in Chinese queries because it is more natural to include these symbols in the English alphabet and numerals than in Chinese ideographs.

In addition, while most previous Web-search studies in English have utilized term-based analysis, we proposed the use of character-based analysis and bigram/n-gram analysis to study the search topics in Chinese queries. We believe this has set an important example and basis for comparison for future research in non-English search-log analysis.

Conclusion and Future Directions

In this article, we reported an exploratory study of Web searching in non-English languages. We analyzed the search log of Timway, a search engine in Hong Kong that focuses on Chinese queries. To our knowledge, this article provides detailed analyses and discussions on Chinese Web-search queries that were not reported in previous studies on Chinese search queries (e.g., Pu et al., 2002). These include analysis on queries with mixed English and Chinese, different encodings of Chinese queries, detailed search-session analysis, character-level analysis, detailed analysis of search operators, and their implications on search engine design.

As more non-English resources are available and more non-English users are using the Internet, it is important to study the information-seeking behavior in non-English search engines. The current study has shed some light on the analysis of search engines in Chinese, and several interesting findings have been identified. Future studies on the search logs in other languages, especially those belonging to a family of languages that has not been investigated in search-log analysis (e.g., Arabic), would be highly desired. More linguistic analyses also could be proposed and used in the study of these search logs to see whether these characteristics can be explained by the linguistic features of each language.

Note that the current study was conducted on data from a search engine in Hong Kong. Due to the differences in culture and dialects, the findings may not be directly applicable to other Chinese users (e.g., Mainland or Taiwan users). Future research will be needed to study whether there are any differences in the search behavior of these groups of users.

Another important direction is the study of search engines that involve more than one language. For example, most general-purpose search engines nowadays accept multilingual queries. Timway also accepts both Chinese and English queries, although we mainly focused on Chinese queries in the current study. It would be interesting to compare and contrast the queries in different languages in the same search engine as well as to study the characteristics of mixed queries.

Acknowledgments

This research was supported in part by a Seed Funding for Basic Research to M. Chau granted by the University of Hong Kong. We would like to thank Timmy Yu from Timway Hong Kong Search Engine Limited for his help in providing the search-log data used in this study. We also thank Jackey Ng and Raygen Lam from the University of Hong Kong for their help in data processing.

References

Armstrong, R., Freitag, D., Joachims, T., & Mitchell, T. (1995). WebWatcher: A learning apprentice for the World Wide Web. In Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments (pp. 6–12).

DOI: 10.1002/asi

Bates, M.J., Wilde, D.N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars: The Getty Online Searching Project Report. Library Quarterly, 63(1), 1–39.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the 7th WWW Conference, Brisbane, Australia. Retrieved from http://infolab.stanford.edu/~backrub/google.html

Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. In Proceedings of the 8th International World Wide Web Conference, Toronto. Retrieved from http://www8.org/w8-papers/5a-search-query/crawling/index.html

Chau, M., Fang, X., & Sheng, O.R.L. (2005). Analysis of the query logs of a Web site search engine. Journal of the American Society for Information Science and Technology, 56(13), 1363–1376.

Chau, M., Qin, J., Zhou, Y., Tseng, C., & Chen, H. (2005). SpidersRUs: Automated development of vertical search engines in different domains and languages. In Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, CO (pp. 110–111).

Chau, M., Zeng, D., Chen, H., Huang, M., & Hendriawan, D. (2003). Design and evaluation of a multi-agent collaborative Web mining system. Decision Support Systems, 35(1), 167–183.

Chen, H., & Chau, M. (2004). Web mining: Machine learning for Web applications. Annual Review of Information Science and Technology, 38, 289–329.

Chen, H., Fan, H., Chau, M., & Zeng, D. (2001). MetaSpider: Meta-searching and categorization on the Web. Journal of the American Society for Information Science and Technology, 52(13), 1134–1147.

Chien, L.-F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. In Proceedings of the 1997 ACM SIGIR, Philadelphia (pp. 50–58). New York: ACM Press.

Cohen, E., Krishnamurthy, B., & Rexford, J. (1998). Improving end-to-end performance of the Web using server volumes and proxy filters. In proceedings of the ACM SIGCOM conference (pp. 241–253). New York: ACM Press.

Croft, W.B., Cook, R., & Wilder, D. (1995). Providing government information on the Internet: Experiences with THOMAS. In Proceedings of the Digital Libraries Conference, Austin, TX (pp. 19–24).

Etzioni, O. (1996). The World Wide Web: Quagmire or gold mine? Communications of the ACM, 39(11), 65–68.

Fang, X., & Sheng, O.R.L. (2004). LinkSelector: A Web mining approach to hyperlink selection for Web portals. ACM Transactions on Internet Technology, 4(2), 209–237.

Fenichel, C.H. (1981). Online searching: Measures that discriminate among users with different types of experience. Journal of the American Society for Information Science, 32(1), 23–32.

Global Reach. (2004). Global Internet statistics. Retrieved March 10, 2007, from http://global-reach.biz/globalstats/index.php3

Hölscher, C. (1998). How Internet experts search for information on the Web. In Proceedings of the World Conference of the World Wide Web, Internet, and Intranet, Orlando, FL.

Hölscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. In Proceedings of the 9th International World Wide Web Conference (WWW9), Amsterdam. Retrieved from http://www9.org/w9cdrom/81/81.html

Huang, C.K., Chien, L.F., & Oyang, Y.J. (2003). Relevant term suggestion in interactive Web search based on contextual information in query session logs. Journal of the American Society for Information Science and Technology, 54(7), 638–649.

Huang, C.K., Oyang, Y.J., & Chien, L.F. (2001). A contextual term suggestion mechanism for interactive search. In N. Zhong, Y. Yao, J. Liu, & S. Ohsuga (Eds.), Web Intelligence: Research and Development, Proceedings of the First Web Intelligence Conference (WI'2001), Japan (pp. 272–281), Lecture Notes in Computer Science 2198, Springer.

Hurst, M. (2001). Layout and language: Challenges for table understanding on the Web. In Proceedings of the 1st International Workshop on Web Document Analysis, Seattle, WA (pp. 27–30).

Jansen, B.J., & Pooch, U. (2000). Web user studies: A review and framework for future work. Journal of the American Society for Information Science and Technology, 52(3), 235–246.

Jansen, B.J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. ACM SIGIR Forum, 32(1), 5–17.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. Information Processing and Management, 36, 207–227.

Jones, S., Cunningham, S.J., & McNam, R. (1998). Usage analysis of a digital library. In Proceedings of the 3rd ACM Conference on Digital Libraries, Pittsburgh, PA (pp. 293–294). New York: ACM Press.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, San Francisco (pp. 668–677).

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3), 574–585.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. ACM SIGKDD Explorations, 2(1), 1–15.

Kwok, S.H., & Yang, C.C. (2004). Searching the peer-to-peer networks: The community and their queries. Journal of the American Society for Information Science and Technology, 55(9), 783–793.

Li, Y., Ding, X., & Tan, C.L. (2002). Combining character-based bigrams with word-based bigrams in contextual postprocessing for Chinese script recognition. ACM Transactions on Asian Language Information Processing (TALIP), 1(4), 297–309.

Norman, J. (1988). Chinese. New York: Cambridge University Press.

Peterson, E. (2004). Chinese encoding converter. Retrieved October 7, 2004, from http://www.mandarintools.com/

Pu, H.T., Chuang, S.-L., & Yang, C. (2002). Subject categorization of query terms for exploring Web users' Search interests. Journal of the American Society for Information Science and Technology, 53(8), 617–630.

Ross, N.C.M., & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information Science, 51(10), 949–958.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. ACM SIGIR Forum, 33(1), 6–12.

Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-commerce: Web search changes. IEEE Computer, 35(3), 107–109.

Spink, A., & Ozmultu, H.C. (2002). Characteristics of question format Web queries: An exploratory study. Information Processing and Management, 38, 453–471.

Spink, A., Ozmutlu, H.C., & Lorence, D.P. (2004). Web searching for sexual information: An exploratory study. Information Processing and Management, 40, 113–123.

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. Journal of the American Society for Information Science and Technology, 52(3), 226–234.

Tsai, C.-H. (1996). Frequency and stroke counts of Chinese characters. Retrieved May 20, 2005, from http://technology.chtsai.org/charfreq/

Wang, P., Berry, M.W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. Journal of the American Society for Information Science and Technology, 54(8), 743–758.

Wolfram, D., Spink, A., Jansen, B.J., & Saracevic, T. (2001). Vox Populi: The public searching of the Web. Journal of the American Society for Information Science and Technology, 52(12), 1073–1074.

Yang, C.C., & Kwok, S.H. (2005). Changes of queries in Gnutella peer-to-peer networks. Journal of Information Science, 31(2), 124–135.

Yang, C.C., Yen, J., Yung, S.K., & Chung, K.L. (1998). Chinese indexing using mutual information. In Proceedings of the 1st Asia Digital Library Workshop, Hong Kong (pp. 57–64). Hong Kong: The University of Hong Kong Press.

Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to Web search results. In Proceedings of the 8th World Wide Web Conference, Toronto. Retrieved from http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html

DOI: 10.1002/asi