Analysis of the Query Logs of a Web Site Search Engine

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

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

Olivia R. Liu Sheng
School of Accounting and Information Systems, The University of Utah, Salt Lake City, UT 84112. E-mail: olivia.sheng@business.utah.edu

A large number of studies have investigated the transaction log of general-purpose search engines such as Excite and AltaVista, but few studies have reported on the analysis of search logs for search engines that are limited to particular Web sites, namely, Web site search engines. In this article, we report our research on analyzing the search logs of the search engine of the Utah state government Web site. Our results show that some statistics, such as the number of search terms per query, of Web users are the same for general-purpose search engines and Web site search engines, but others, such as the search topics and the terms used, are considerably different. Possible reasons for the differences include the focused domain of Web site search engines and users' different information needs. The findings are useful for Web site developers to improve the performance of their services provided on the Web and for researchers to conduct further research in this area. The analysis also can be applied in e-government research by investigating how information should be delivered to users in government Web sites.

Introduction

The amount of information on the World Wide Web is growing rapidly, and search engines have become increasingly important in helping users in information retrieval and other activities on the Web. General-purpose search engines, such as Google (http://www.google.com/), Excite (http://www.excite.com/), and AltaVista (http://www.altavista.com/), have been widely used. Many users begin their Web activities by submitting a query to a search engine. These search engines use Internet spiders/crawlers that automatically collect Web pages and create an index that can be searched by users (Chau & Chen, 2003b). As these general-purpose search engines do not restrict themselves to particular domains or specialties, they often try to collect as many Web pages as possible. However, as the number of indexable pages on the Web has exceeded 3 billion, it has become more difficult for these search engines to keep an up-to-date and comprehensive search index; low precision and low recall rates often result.

On the other hand, many Web sites have their own search engines. A Web site search engine is one that allows users to search for pages only within a particular Web domain or Web host. It is often found in the main page of a Web site and only indexes pages in that particular Web site. For example, going to the homepage of Microsoft Corporation (http://www.microsoft.com/), one would see a text box on the page that allows users to perform a search restricted to the microsoft.com Web site. Because a Web site search engine only needs to work with a limited set of pages, it can provide customized features to users and can be updated much more frequently (e.g., daily or even hourly) than general-purpose search engines. In addition, a Web site search engine is often more comprehensive because it can index pages that are not accessible by search engine crawlers/spiders, and no general-purpose search engines can cover every single page in every single Web site on the Internet. Although Web site search engines are very useful for users looking for information on a particular Web site, these search engines often have different user interfaces, interpret queries in different ways, support different types of advanced search functionalities, and employ different search algorithms. From end users' point of view, dealing with an array of different interfaces and understanding each one's idiosyncrasies add much confusion and present an additional layer of information and cognitive overload. Most users find it difficult to adapt to different search engines and cannot

Received January 23, 2004; revised July 26, 2004; accepted August 24, 2004

 ● Published online 31 August 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20210

fully utilize their capabilities. Therefore, it is very important to understand better the information needs and the search behavior of users in order to build better Web site search engines.

Ideally, we would sit behind the users' back, watch how they perform searches using these search engines, and record their actions and search results. Such observation is, however, not feasible for large-scale evaluation of Web site search engines because of the large number of users who are scattered around the world. Although it is possible to capture users' actions on their computers (e.g., their clicking on a mouse button or scrolling a window on the screen) by using client-side monitoring techniques, doing so often requires the installation of specific software or "plug-ins" on the users' computers (Montgomery & Faloutsos, 2001; Fenstermacher & Ginsburg, 2003). As these techniques require extra time and effort from the users and introduce privacy concerns, most users are not willing to install such software.

Alternatively, server-side search engine log data can provide much information about users' search behavior and information needs. Server-side data are easy to collect without any extra effort from the users, and they are much more comprehensive and scalable than client-side data as they can cover 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, Spink, Bateman, & Saracevic, 1998; Jansen, Spink, & Saracevic, 2000; Ross & Wolfram, 2000; Spink, Jansen, Wolfram, & Saracevic, 2002; Spink, Wolfram, Jansen, & Saracevic, 2001). Analysis of the AltaVista query logs also has been reported (Silverstein, Henzinger, Marais, & Moricz, 1999). Such analyses can provide much information about the information needs and searching behavior of search engine users. There is, however, little research reported on the analysis of Web site search engine logs. It has been shown that the information needs of users of Web site search engines can be quite different from those of users of general-purpose search engines (Wang, Berry, & Yang, 2003). It would be interesting to compare the various metrics across different types of search engines.

In this article, we report our research on the analysis of the search query logs collected from a Web site search engine. We study the information needs and search behavior of the users for the search engine and compare them with those of general-purpose search engine users. The article is structured as follows: In the second section we review related research in Web mining and search engine log analysis. We pose our research questions in the third section. The fourth section discusses the data and the methods we used in this research. In the fifth section we present and discuss the findings of our analysis. We conclude the article in the sixth section with a summary of our study and some future directions.

Related Studies

Analysis of Web log data or search engine log data can be categorized under the research area of Web mining. The term Web mining was first used by Etzioni (1996) to denote the use of data mining techniques to discover Web documents and services, extract information from Web resources, and uncover general patterns on the Web automatically. Over the years, Web mining research has been extended to cover the use of data mining as well as other similar techniques to discover resources, patterns, and knowledge from the Web and 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 (Kosala & Blockeel, 2000; Chau, & Chen, 2003a). *Web content mining* is the discovery of useful information from Web contents, including text, images, audio, and video. Web content mining research includes resource discovery from the Web (e.g., Chakrabarti, Chau et al., 2003; van den Berg, & Dom, 1999; Chau & Chen, 2003a), 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 in order to improve the performance of a Web site (e.g., Cohen, Krishnamurthy, & Rexford, 1998; Fang & Sheng, 2004).

The mining of search engine logs, usually focused on the study of how users use the search engines on the Web to satisfy their information needs, belongs to the category of Web usage mining. On the other hand, it also is a strongly rooted in information retrieval research. Many studies have reported analysis of user information behavior, search queries, and search sessions with various information retrieval and digital libraries systems (e.g., Fenichel, 1981; Bates, Wilde, & Siegfried, 1993). 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 studies 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 1 million queries submitted to the Excite search engine (Spink et al.,

2001, 2002; Wolfram, Spink, Jansen, & Saracevic, 2001). Researchers were also able to obtain interesting findings on the information needs and search behavior of users, such as the trends in Web searching (Spink et al., 2002), sexual information searching on the Web (Spink, Ozmutlu, & Lorence, 2004), and question format Web queries (Spink & Ozmultu, 2002). Another large-scale Web query analysis was performed by Silverstein and associates (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, and number of result pages viewed by each user. These metrics allow researchers to compare their findings across different types of search engines at different times.

The second category of research focused on analyzing the search logs of a specific Web site or system. However, only a few studies have been reported in this category. One example is the study of Croft, Cook, and Wilder (1995), which investigated the search queries submitted to the THOMAS system, an online searchable database consisting of U.S. legislative information. They analyzed 94,911 queries recorded in their system and identified the top 25 queries. They also found that 88% of all queries contain three or fewer words, a number much lower than that of traditional information retrieval systems. Jones, Cunningham, and McNam (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 queries were composed of three or fewer words. Their study also found that most people use the default settings of search engines without any modifications. Wang, Berry, and Yang (2003) analyzed the search queries submitted to the search engine of the University of Tennessee, Knoxville, over a period of 4 years. They performed a longitudinal analysis on the search queries and found that seasonal patterns exist in Web site searching. They also identified some differences between the search queries submitted to a general-purpose search engine and a Web site search engine, in terms of search topics, search term distribution, and mean length of queries.

Research Questions

Most previous studies focused on the users of general-purpose Web search engines, and their findings may not be applicable to Web site search engines that have their own characteristics and groups of users. Consequently, a Web site search engine should be designed in a customized way according to its users' information needs, which we suggest can be identified from the search engine's query log data.

We address the following research questions in this study: (1) What are the characteristics of the search queries submitted to a Web site search engine? (2) How do these queries compare to those submitted to general-purpose search engines? (3) How can these results be used to improve the design of the Web site search engine?

Data and Methods

In this study, we collected the search queries submitted to the Utah state government Web site in the United States (http://www.utah.gov/), captured over a period of 168 days from March 1, 2003, to August 15, 2003. The Utah state government Web site was one of the most advanced government Web sites and was named the best state government Web portal in the United States by the Center for Digital Government (Center for Digital Government, 2003). The Web site search engine is accessible from a text box with the text "Search Utah.gov" near the top of the main page of the Web site (see Figure 1). A user can enter a search query in the box and click on the Go button to submit the query to the Web site search engine and obtain the search results. Alternatively, the user can go to the Web page (http://info.utah.gov/) to type in the search query and specify the search options (see Figure 2).

In total, there are 1,895,680 records in the Web transaction log. Each record in the log file represents a request sent from a user to the search engine. A request can be a search query (requesting either the first page of search results or subsequent pages beyond the top 25 results), a request for viewing the actual document in the search result, or a request for an image file for display. Each record contains 14 fields, including date, time, Internet Protocol (IP) address, type of request submitted, and the parameters of the request. An example of a record is given in Table 1. We are most interested in the fields Date, Time, Client IP Address, Request method, Uniform Resource Identifier (URI) stem, URI query, and User cookie. In the following we discuss the use of these fields.

The first field we used for our processing is the URI Stem, in which the value "/search.asp" specifies that the transaction is search-related. Other non-search-related transactions have the value of a file name in the URI Stem field, e.g., "/Images/RankLow.gif." Because the requests for such static documents or image files are not relevant to our study, these transactions were removed from our records. Some other irrelevant transactions were also removed from the logs. Thus, the remaining logs represent two major types of transactions: a user submitting search query or a user viewing an actual document on the Web site in the search result. The second type of query is especially interesting because most previous research on search engine log analysis (such as the studies on AltaVista and Excite) did not capture or study whether users click on any actual documents in the search result pages.

Among the search queries, 443,342 queries were generated by Web spiders—programs that automatically collect pages from the Web. These spiders sent queries to the search engine to generate search results for search engine indexing (Chau & Chen, 2003b). The queries generated by search spiders can be identified by the User-Agent field in the transaction log. For example, the queries submitted by the search

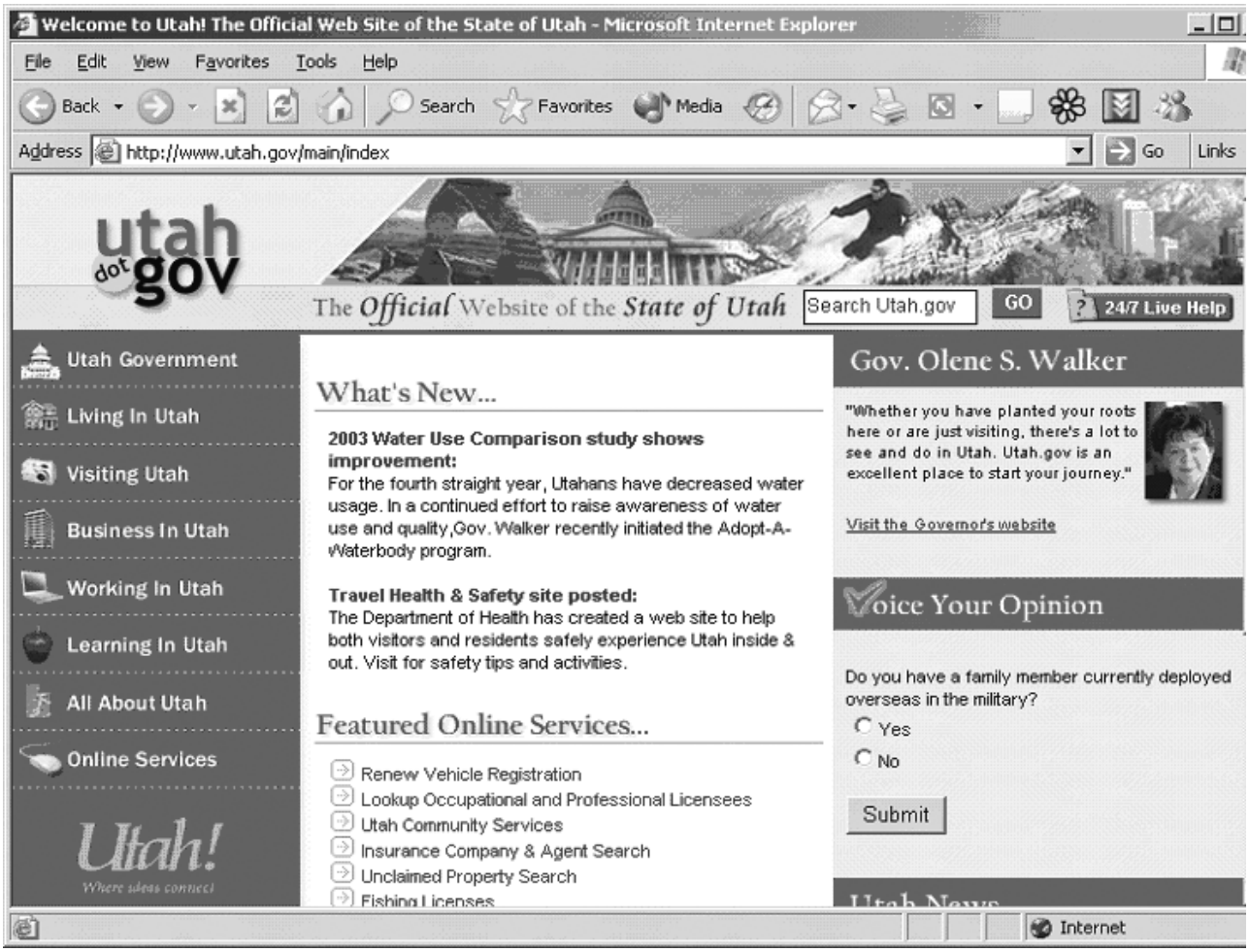

FIG. 1. The main page of the Utah state government Web site with access to the search engine.

spiders from FirstGov.gov, a search engine for U.S. government information, have the value "FirstGov.gov+Search+-+POC:firstgov.webmasters@gsa.gov" in the User-Agent field in the log, but the same field for a transaction submitted by a real user would have the name and features of the Web browser used by the user, such as "Mozilla/4.0+ (compatible;+MSIE+6.0;+Windows+NT+5.0)." Because the queries generated by the search spiders do not represent the real information needs of real users, we also removed these transactions from the logs. One should note that there may still be some automatically generated queries in the log data after the filtering because search engine spiders can "fake" themselves as real users when submitting their queries to the Utah search engine. However, the number of such queries should be relatively small and hence negligible.

The data were then loaded into a relational database for further processing and analysis. The next step was to identify the different sessions in the search query data. A session is a series of queries submitted by a single user within a small range of time (Silverstein et al., 1999). A session represents a set of queries relevant to a user's single information need. To identify session information, we first had to identify the unique users in the transaction logs. We did so on the basis of the cookie information and the IP address of each user. As each browser was assigned a unique cookie by a Web site, we could identify each unique browser used by the users. Although it is possible that users may share the same computer (e.g., in a public library), the possibility did not affect our identification of sessions very much. The IP address is less reliable in identifying unique users from a Web transaction log because the same user may be assigned two different IP addresses in different sessions, and two different users may share the same IP address. The intermediate proxy servers may even assign the same IP to all users, a problem known as the *AOL effect* (Pierrakos, Paliouras, Papatheodorou, & Spyropoulos, 2003). Because the use of cookies is more reliable, it was chosen as the first metric in our processing. In cases in which the cookie information was not available (e.g., disabled by the user), the IP address was used. Each user identified was assigned a unique identification (ID) in our database.

We then ordered the search queries for each user according to the timestamp in the transaction logs. Following previous

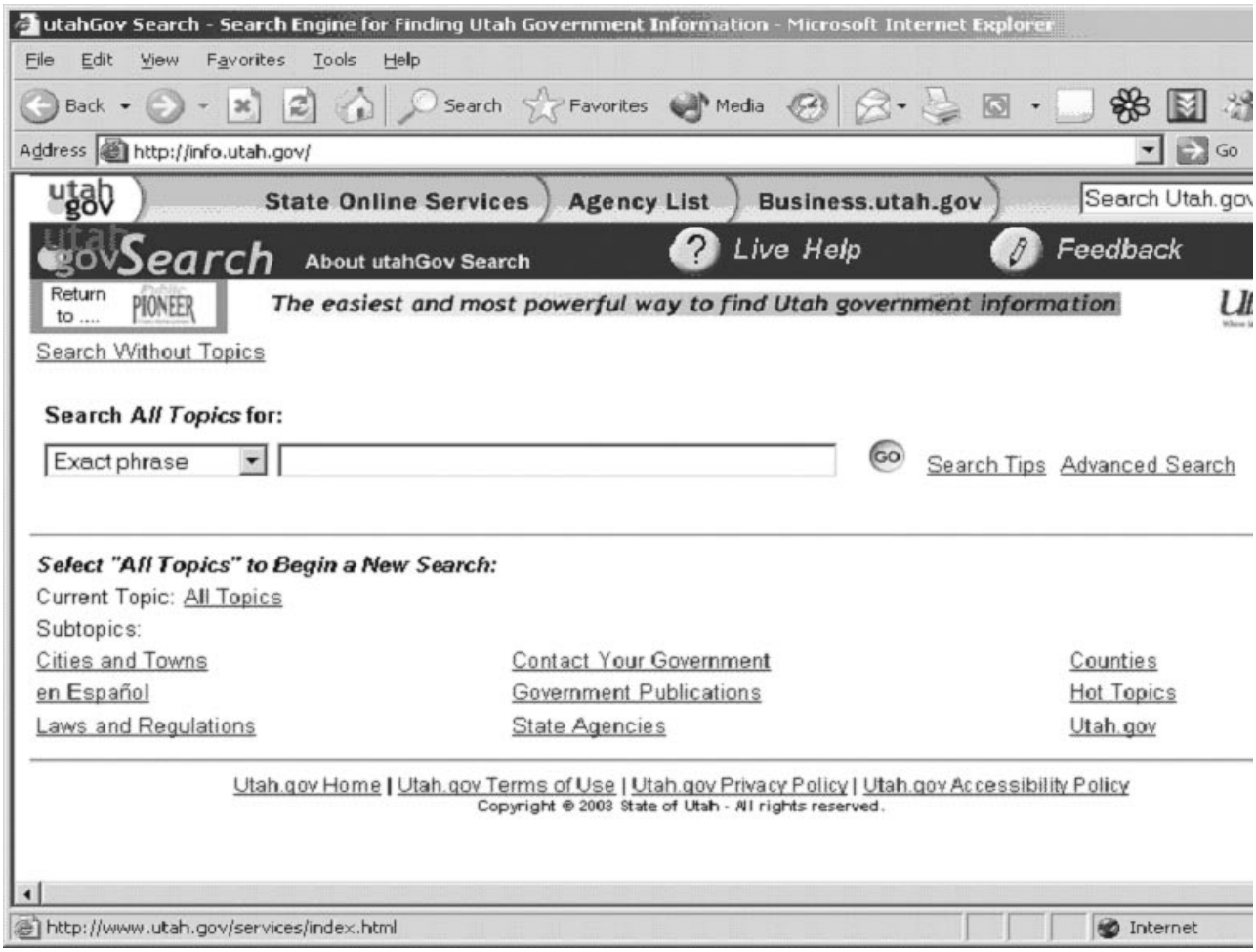

FIG. 2. The search engine of the Utah state government Web site.

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

Field	Value
Date	2003-03-01
Time	00:02:40
Client IP address	66.119.xxx.xxx (hidden for privacy reasons)
Client-to-server username	—
Server IP address	198.239.xxx.xxx (hidden for security reasons)
Server port	80
Request method	GET
URI stem	/search.asp
URI query	postFlag=1&opr1=1&val1=dmv&MSS.request.Search%20Catalog.....
Server-to-client status	200
Server-to-client bytes	0
User agent (browser)	Mozilla/4.0+(compatible;+MSIE+5.5;+Windows+NT+5.0;+T312461)
User cookie	—
Referrer URL	http://utah.gov/government/judicial.html

Note. IP, Internet Protocol; URI, Uniform Resource Identifier; URL, Universal Resource Locator.

research that suggested that queries for a single information need should be close together in terms of time (Silverstein et al., 1999), we used a cutoff of 30 minutes to identify sessions. If a user submitted a query within 30 minutes of the previous transaction, these transactions were included in the same session. On the other hand, if a user did not submit any requests to the Web server for 30 minutes, anything submitted afterward would be included in a new session. Each session was assigned a unique session ID in our database.

Because the Utah state government Web site search engine employs the Active Server Pages (ASP) method in their Web transactions, each query had to be parsed from a set of parameters from the URI Query field in the transaction log before it could be further processed. A simple program was written in Java to perform this task. All queries were stored in lowercase letters. The terms, Boolean operators, and other search features in the query were also identified for each query and stored in the database. The data consist of 1,115,388 transactions in total, in which 792,103 transactions are search queries and 323,285 are requests for actual documents in the search result. An overview of our data is presented in Table 2.

TABLE 2. Overview of the search log data.

Total number of search queries	792,103
Total number of requests for actual documents in the search result	323,285
Total number of unique users	161,042
Total number of sessions	458,962

TABLE 3. Statistics on queries.

Total number of search queries	792,103
Total number of unique queries	575,389
Total number of repeat queries	98,418
Total number of empty queries	118,296
Mean number of queries per session	1.73
Median number of queries per session	1
Mean number of unique queries per session	1.25
Median number of unique queries per session	1

Analysis Results

In this section we present the characteristics of our query data logs and the results of our analysis. We first generate descriptive statistics about the queries and search sessions of users and then compare them with the results from other studies. We also analyze how users utilize the advanced search features provided by the Web site search engine. Finally, we perform text analysis on the queries submitted by the users to identify the most common query words used and the associations between search terms.

Sessions, Queries, and Topics

Sessions and queries. As discussed earlier, there are 792,103 queries in total, submitted by a total of 161,042 unique users in 458,962 sessions. We classified the 792,103 queries into three groups, namely, unique queries, repeat queries, and empty queries (Spink et al., 2001). Unique queries are all differing queries entered by one user in one session. The differing queries can be new queries or modifications of the previous query. Repeat queries are the repeat occurrences of any query that appears previously in a session. Such repeat occurrences of a query also result from the viewing of subsequent result pages by the users. Empty queries are queries that contain no query terms. In our study, we found that a large number of users clicked on the Go button on the home page of the Utah state government Web site (see Figure 1) without changing the text "Search Utah.gov" in the search box. The result was a search query of "Search Utah.gov" submitted to the search engine. Because the users did not enter any search term in such cases, we also consider these as empty queries.

Out of the 792,103 queries, 575,389 (72.6%) are unique queries, 98,418 (12.4%) are repeat queries, and 118,296 (14.9%) are empty queries. The mean number of queries in each session is 1.73 (with a median of 1), and the mean number of unique queries in each session is 1.25 (with a median of 1). This number is much lower than the number of 2.52 reported in the Excite study (Spink et al., 2001) and 2.02 reported in the AltaVista study (Silverstein et al., 1999). The results are summarized in Table 3.

To study the number of queries per session in more detail, we look at the distribution of the numbers. In out study, 73.0% of sessions contain only one search query; 12.4% of sessions contain only two (see Figure 3). As can be seen, the distribution is skewed toward the lower end in terms of the number of queries submitted. About 96.1% of users submitted four or fewer queries in a session. This finding is consistent with those of other studies, in which most sessions are very short and contain only one or two queries (Spink et al., 2001; Silverstein et al., 1999).

There are two possible reasons for the lower number of queries submitted to our search engine per session when compared with results reported in previous studies. First, when compared with the Excite study, our study has a different definition of sessions involving time-out and cookie information, which would result in a larger number of sessions. Second, it is possible that many users were able to find the results they wanted in the first query and therefore did not need to modify their queries to perform a new search. This reult may indicate a special characteristic of Web site search engines because they need to index only a limited number of pages; higher precision and recall in the search results may be achieved more easily. At the same time, users may have a better idea of what they want specifically when using a Web site search engine, so they may be able to formulate better queries in the first step. Therefore, there is a smaller need for users to modify their search queries. It is also possible that people use general-purpose search engines and Web site search engines differently for other reasons. As previous Web site search engine studies do not include session information (Croft et al., 1995; Jones et al., 1998; Wang et al., 2003), further research will be needed.

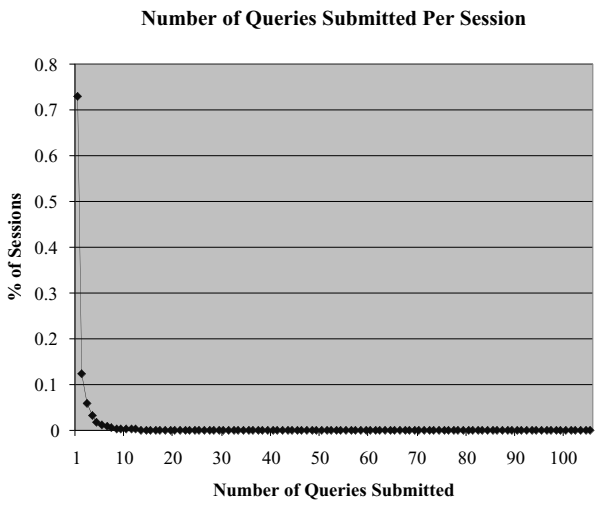

FIG. 3. Number of queries submitted per session.

TABLE 4. Comparison of top 25 queries with AltaVista and Knoxville.

	This study		AltaVista study		Knoxville study	
Rank	Query	Frequency	Query	Frequency	Query	Frequency
1	dmv	3,794	sex	1,551,477	career services	9,587
2	tax forms	2,532	applet	1,169,031	grades	5,727
3	sex offenders	2,173	porno	712,790	tuition	4,837
4	forms	2,036	mp3	613,902	housing	4,203
5	jobs	1,587	chat	406,014	timetable	4,097
6	divorce	1,400	warez	398,953	bookstore	3,453
7	unemployment	1,359	yahoo	377,025	Rocky Top	2,582
8	employment	1,257	playboy	356,556	transcripts	2,340
9	notary	1,061	xxx	324,923	Daily Beacon	2,312
10	secretary of state	1,053	hotmail	321,267	employment	2,156
11	sales tax	941	[non-ASCII]	263,760	cheerleading	1,985
12	map	921	pamela anderson	256,559	band	1,914
13	sex offender	890	p**** [vulgarity]	234,037	registration	1,683
14	dopl	884	sexo	226,705	scholarships	1,537
15	ors	879	porn	212,161	jobs	1,488
16	maps	849	nude	190,641	football tickets	1,465
17	real estate	798	lolita	179,629	career	1,407
18	taxes	763	games	166,781	marching band	1,397
19	tax	748	spice girls	162,272	cheerleaders	1,377
20	birth certificates	741	beastiality	152,143	resume	1,375
21	birth certificate	729	animal sex	150,786	financial aid	1,331
22	drivers license	717	SEX	150,699	webmail	1,317
23	medicaid	714	gay	142,761	tickets	1,225
24	child support	699	titanic	140,963	transcript	1,211
25	"dmv"	656	Bestiality	136,578	catalog	1,187

Search Queries and Search Topics. To investigate the search topics of the users, we identified the top 25 queries submitted to the Utah search engine and compared them with that of the AltaVista study (Silverstein et al., 1999) and the Knoxville study (Wang et al., 2003), which analyzed the user queries submitted to the search engine of the University of Tennessee, Knoxville. The data are shown in Table 4.

The top three topics identified in our study are "dmv," "tax forms," and "sex offenders." From the table, it can be seen that the top queries are quite different across the three search engines. The top AltaVista queries are mostly related to sexual information, software, music, and entertainment, and the queries submitted to the Knoxville search engine are mostly related to academic matters. The top queries in our study, however, are government-related. In other words, search queries submitted to the Web site search engines are generally relevant to the corresponding domain. The results suggested that general-purpose search engines and Web site search engines have to be designed differently according to users' different information needs and query characteristics.

Seasonal effect of search topics. In their longitudinal analysis of a Web query log that covered a period of 4 years, Wang and coworkers (2003) showed that a seasonal effect exists in an academic Web site search engine. They found that the query "career services" occurred mostly in February, March, September, and October, and the query "football tickets" appeared mostly in August and September. This interesting finding had not been identified in previous general-purpose search engine research, in which the data were often limited to a short period (e.g., 43 days in the AltaVista study and 1 day in the Excite study). The data in our study covered a period of about 5.5 months (168 days). Though the data did not cover a whole calendar year, it is also interesting to see whether any particular seasonal patterns exist in our data. To this end, we took the top three queries, namely, "dmv," "tax forms," and "sex offenders," and analyzed their daily search frequencies. The results are plotted in Figure 4.

No apparent seasonal patterns were found for the queries "dmv" and "sex offenders" (see Figures 4a and 4c). However, it is interesting to note that there are slightly more requests for "sex offenders" in March. This increase is probably not related to seasonal effect; as we suggested, it may be caused by the fact that the U.S. Supreme Court ruled on March 5, 2003, that it is legal for state governments to put pictures of convicted sex offenders on the Internet (CBS News, 2003). This news was widely covered in the media, and that coverage may have drawn the public to look for the sex offender listing in the Utah state site and for other relevant information on the Web site.

On the other hand, it can be easily seen that the "tax forms" query demonstrated a very strong seasonal effect in our data (see Figure 4b). The number of search queries for "tax forms" had been steady since the beginning of March (or possibly earlier but we do not have the data to verify). The

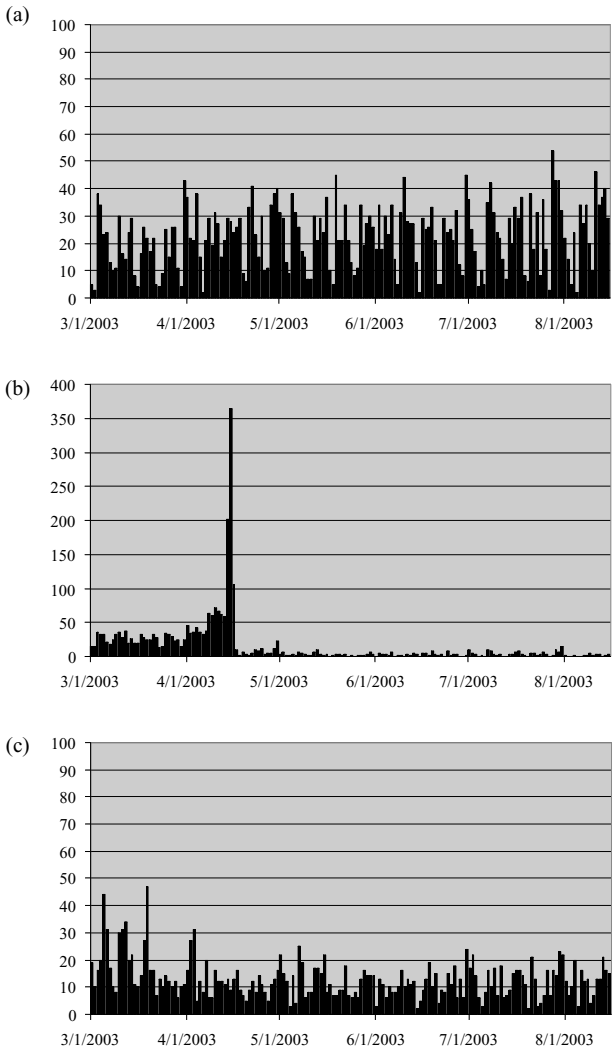

FIG. 4. Analysis of daily frequencies of the top queries: (a) "dmv," (b) "tax forms," (c) "sex offenders."

number of requests peaked on April 15, the deadline for filing individual tax returns in the United States, with 364 requests on a single day. The number dropped dramatically after the deadline had passed, with an average of fewer than 10 requests per day. This finding corroborates the seasonal effect identified by Wang and associates (2003) discussed earlier.

To analyze further the seasonal effect of tax-related queries, we analyzed the daily requests for a broader set of tax-related queries, including all queries that contain the terms "tax," "irs," or "internal revenue." The result is shown in Figure 5. It can be seen that the pattern is similar to that in Figure 4b, in which the number reached the peak on April 15 and decreased quickly afterward.

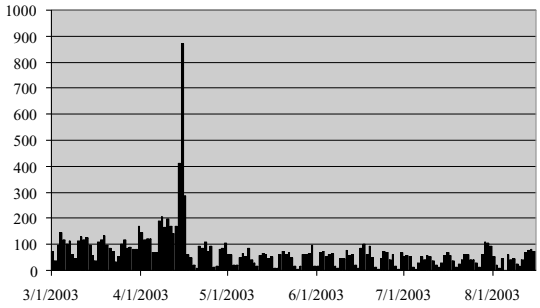

FIG. 5. Analysis of daily frequencies of the tax-related queries.

Search Terms

Search terms in queries. Similarly to other Web search studies, we analyze the search terms in each query submitted to the Web site search engine. Analysis of the search terms can reveal how users formulate their search queries and what the search topics are. The basic statistics are shown in Table 5. In total, there are 1,518,984 terms in the queries. Out of these terms, 67,958 are unique terms. The longest query consists of 40 terms. The mean number of terms in a query is 2.25, with a median of 2. This finding is consistent with the result in the AltaVista study and the Excite study, which found the mean number of terms in a query to be 2.35 and 2.21, respectively. The finding reiterates that Web queries are much shorter than those of traditional information retrieval systems. The distribution of the number of terms per query is shown in Figure 6. Some 30.7% of queries contain only a single term, 37.0% contain two, and 19.2% contain three. About 97.6% of queries contain five or fewer terms. The results show that Web users are likely to use short queries, whether they are using general-purpose search engines or Web site search engines.

Search term distribution. Several previous studies on Web search analysis have suggested that the distribution of terms used in Web search engines largely follows the Zipf distribution (Jansen et al., 2000; Spink et al., 2001). In a Zipf distribution, the quantity of interest is inversely proportional to its rank, and the Zipf distribution represents the distribution of terms in long English texts. To see whether the same pattern is observed in our data, a double-log rank-frequency chart, as shown in Figure 7, was used. The plot should be close to a straight line for a Zipf distribution.

It can be seen that the plot follows the Zipf distribution, with some discrepancies for the high- and low-ranking

TABLE 5. Statistics on terms.

Total number of terms	1,518,984
Total number of unique terms	67,958
Mean number of terms per query	2.25
Median number of terms per query	2
Largest number of terms per query	40
Percentage of nonempty queries with one term	30.7%
Percentage of nonempty queries with two terms	37.0%
Percentage of nonempty queries with three terms	19.2%
Percentage of nonempty queries with four terms	7.6%
Percentage of nonempty queries with five terms	3.2%
Percentage of nonempty queries with six or more terms	2.4%

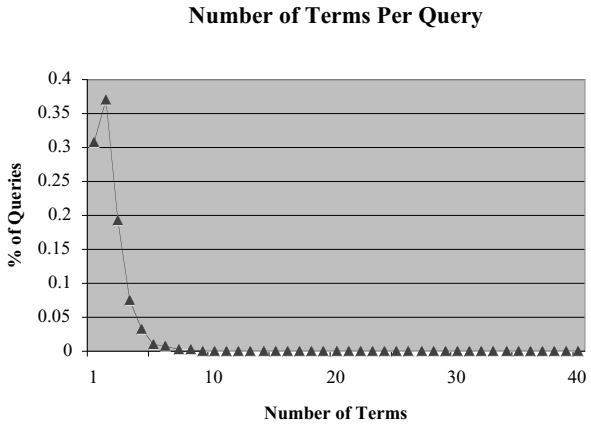

FIG. 6. Number of terms per query.

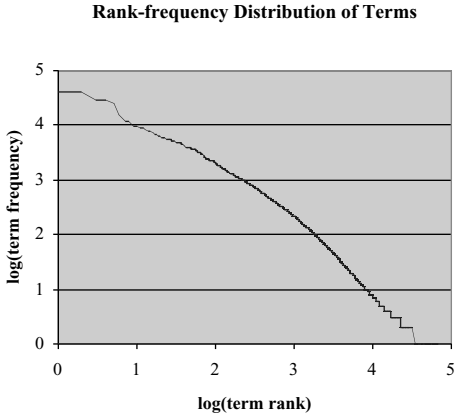

FIG. 7. Double-log rank-frequency graph for search terms.

terms. The slope of our plot is −0.9533, which is close to the theoretical value of −1 for a Zipf distribution. When compared with the distribution reported in the Excite study (Jansen et al., 2000; Spink et al., 2001), the distribution in the present study corresponds better to the Zipf distribution, especially for the lower end of the curve (terms with low frequency). Although a more sophisticated model may be needed, the preliminary finding suggests that there is a smaller percentage of infrequently used terms in Web site search engines than in general-purpose search engines. One possible explanation is that although the users of both types of search engines use a diverse set of terms, Web site search engines are restricted to particular domains and the proportion of unique or low-frequency terms is therefore also restricted.

Search terms and search topics. To study the search topics of the users further, we identified the top 50 terms in the queries in our study and compared with those of the Excite study (Spink et al., 2001) and the top 20 terms reported in the Knoxville study (Wang et al., 2003). The results are shown in Table 6. The general pattern is similar to that of the query analysis presented in Table 4. First, the terms used in Web site search engines are very different from those used in general-purpose search engines. Although some functional words are common across all three studies (such as "and," "of," and "for"), the semantic words are very different. The top Excite terms are mostly related to sexual information; the query terms submitted to the Web site search engines are mostly relevant to the corresponding domain (e.g., "utah," "tax," and "state" for the Utah state government Web site and

TABLE 6. Comparison of top 50 query terms with Excite and Knoxville.

	This study		Excite study		Knoxville study
Rank	Term	Frequency	Term	Frequency	Term
1	utah	40,425	and	21,385	of
2	of	40,072	of	12,731	services
3	and	29,208	sex	10,757	career
4	tax	28,457	free	9,710	student
5	state	25,681	the	8,013	and
6	license	15,099	nude	7,047	grades
7	county	12,115	pictures	5,939	school
8	forms	12,049	in	5,196	tuition
9	department	9,647	university	4,383	housing
10	lake	9,398	pics	3,815	football
11	business	8,866	chat	3,515	timetable
12	for	8,816	for	3,431	schedule
13	search	8,612	adult	3,385	center
14	sex	7,626	women	3,211	office
15	form	7,560	new	3,109	band
16	in	7,298	xxx	3,010	for
17	city	6,727	girls	2,732	department
18	salt	6,649	music	2,490	UT
19	services	6,383	porn	2,400	Tennessee
20	laws	6,188	to	2,265	graduate
21	registration	6,016	gay	2,187	
22	child	5,929	school	2,176	
23	dmv	5,825	home	2,150	
24	code	5,661	college	1,043	
25	insurance	5,545	state	2,010	
26	sales	5,545	naked	1,968	
27	health	5,439	american	1,961	
28	public	5,289	stories	1,958	
29	the	5,049	software	1,908	
30	vehicle	5,036	games	1,904	
31	division	4,974	diana	1,885	
32	jobs	4,840	p****	1,876	
33	records	4,752	black	1,823	
34	water	4,730	on	1,813	
35	income	4,706	photos	1,799	
36	offenders	4,659	jobs	1,735	
37	a	4,648	world	1,734	
38	map	4,373	a	1,711	
39	property	4,348	magazine	1,690	
40	application	4,238	nudes	1,690	
41	unemployment	4,213	news	1,687	
42	office	4,020	football	1,627	
43	motor	4,012	page	1,591	
44	commission	3,931	computer	1,533	
45	notary	3,929	princess	1,461	
46	court	3,873	airlines	1,409	
47	to	3,872	download	1,381	
48	employment	3,835	real	1,381	
49	divorce	3,747	education	1,376	
50	marriage	3,678	art	1,374	

"career," "student," and "grades" in the academic domain). Other popular search terms identified in our study include "license," "country," "forms," "department," and "laws" (see Table 6). Again, the results suggested that users have different information needs when using general-purpose search engines and Web site search engines. It is also interesting to note that the term "sex" also ranks 14th in our study. After looking into the queries containing the term, we found that most of these queries were submitted to search for information concerning the list of sex offenders in the state of Utah. By contrast, in the Excite data a large percentage of queries are related to sex and pornography (Spink et al., 2002; Spink et al., 2004).

When analyzing the functional words, we found that both "and" and "of" appear in the top five in all three studies; "and" is frequently used because it can be used as a Boolean operator as well as a term on its own. On the other hand, it is interesting to note that although the word "of" is ignored by many search engines unless clearly specified by the user (e.g., Google), it is still frequently used by Web searchers. As pointed out by Wang and colleagues (2003), general-purpose search engines such as Excite appear to have more functional words in the top terms, and Web site search engines have more semantic terms. As can be seen from Table 6, Excite has 6 functional words in the top 20 ("and," "of," "the," "in," "for," "to"), but the Knoxville data has only 3 ("of," "and," "for"), and this study only has 4 ("of," "and," "for," "in"). One possible reason is that the search queries in general-purpose search engines are more diverse, such that fewer semantic words appear frequently enough to appear in the top 20 list. However, for Web site search engines, the search queries in limited domains include more semantic words on the same topics that are more frequently used and thus have a higher rank.

To analyze further the topics submitted to the Utah state government Web site search engine, we also studied which terms were used more frequently together. Terms used together are often more informative in identifying which topics are frequently searched by users. The top 50 term pairs are shown in Table 7. Many frequently searched topics can be identified from the data, e.g., "tax forms," "sales tax," "state tax," "sex offenders," "business license," and "birth certificate."

TABLE 7. Top 50 term pairs.

Rank	Term pair		Frequency
1	department	of	7,622
2	state	utah	6,997
3	of	utah	6,988
4	lake	salt	6,335
5	tax	forms	5,469
6	of	state	4,471
7	sales	tax	4,372
8	state	tax	4,333
9	offenders	sex	4,145
10	division	of	3,734
11	in	utah	3,658
12	income	tax	3,596
13	estate	real	3,463
14	tax	utah	2,908
15	code	utah	2,667
16	city	lake	2,587
17	city	salt	2,549
18	offender	sex	2,401
19	of	services	2,313
20	drivers	license	2,194
21	and	of	2,156
22	commission	tax	2,057
23	of	office	2,006
24	child	support	1,921
25	of	the	1,812
26	county	lake	1,791
27	county	utah	1,772
28	form	tax	1,772
29	county	salt	1,762
30	notary	public	1,683
31	commerce	of	1,680
32	of	secretary	1,647
33	secretary	state	1,641
34	motor	vehicles	1,640
35	and	utah	1,633
36	motor	vehicle	1,630
37	return	tax	1,501
38	motor	of	1,495
39	recovery	services	1,462
40	articles	of	1,368
41	business	license	1,326
42	park	state	1,302
43	parks	state	1,299
44	birth	certificate	1,293
45	security	social	1,284
46	application	for	1,274
47	department	utah	1,234
48	board	of	1,227
49	of	vehicles	1,225
50	child	care	1,203

As can be seen from Table 7, many of term pairs appear to be part of a group of three or more terms that appear frequently together. For example, the pairs "state-utah" and "of-utah" are likely to be part of the phrase "state of utah," and the pairs "department-of," "division-of," "motor-vehicle," "motor-vehicles," and "of-vehicles" are possibly part of the phrase "department of motor vehicles" and its variations. In order to identify these topics, we analyzed our data again and identified the groups of three and four terms that appear most frequently in the queries. The most frequent term groups with three terms are listed in Table 8 and those with four terms are listed in Table 9.

In Table 8, search topics such as "Salt Lake City," "Utah State Tax," "State of Utah," and "Secretary of State" can be easily identified. The table also reveals that some term groups still appear to be part of an even longer phrase, e.g., "motor-of-vehicles," "department-of-motors," etc. In Table 9, we identified the top three four-term groups that co-occurred most frequently in the queries. One can easily see that these represent three search topics frequently requested by users, namely, "Office of Recovery Services," "Department of Motor Vehicles," and "Utah State Tax Commission." The results show that such information is frequently requested by users and suggests that Web site designers should allow users to access it more easily

TABLE 8. Top term groups with three terms.

Rank	Term groups			Frequency
1	salt	lake	city	2,524
2	salt	lake	county	1,749
3	utah	state	tax	1,613
4	state	of	utah	1,611
5	secretary	of	state	1,610
6	motor	of	vehicles	1,168
7	department	of	utah	1,082
8	of	office	recovery	1,039
9	bill	of	sale	1,005
10	state	tax	commission	1,000
11	department	of	motor	989
12	department	of	commerce	980
13	of	office	services	944
14	of	recovery	services	922
15	office	recovery	services	917
16	department	of	services	870
17	department	motor	vehicles	822
18	department	of	vehicles	812
19	utah	tax	commission	811
20	state	tax	forms	774
21	income	tax	forms	769
22	state	income	tax	755
23	utah	state	commission	734

TABLE 9. Top term groups with four terms.

Rank	Term group				Frequency
1	office	of	recovery	services	894
2	department	of	motor	vehicles	805
3	utah	state	tax	commission	702

(e.g., through links from the main page of the Utah state government Web site).

Result Pages and Actual Documents Viewed

Search result pages. During a search session with the Utah state government Web site, a user first submits the query to the search engine and the first result page listing the top 25 search results will be shown. If there are more than 25 hits for the search, the user can request subsequent result pages (sometimes known as *result screens*). As mentioned earlier, these are often counted as repeat queries in Web search studies. On average, each user views 1.47 pages in the search results in our study (see Table 10).

This result is comparable to the one reported in the AltaVista study, in which the average number of result pages viewed in a session was 1.39. The Excite study also found that 28.6% of users examined only the first page of the search results. Although the Utah search engine provides 25 search results in the first page, we do not see a big difference in the number of result pages viewed per user. In general, all results suggested that most Web users browse only a small number of result pages.

TABLE 10. Result pages and actual documents viewed.

Number of result pages (25 items each) viewed per session	1.47
Number of actual documents in the result list viewed per session	0.70
Percentage of sessions viewing zero document in the search results	59.4%
Percentage of sessions viewing one document in the search results	28.3%
Percentage of sessions viewing two documents in the search results	6.9%
Percentage of sessions viewing three or more documents in the search results	5.4%

Actual documents viewed. When browsing through the result pages, if the user sees an item in the search result of interest, he or she can click on the link and view the actual content of the searched document. The Utah state government Web site site search engine has been designed in such a way that when such an action is performed, it is logged in the transaction. In total, there are 323,285 records of such requests. On average, only 0.70 document is viewed in each session, with a median of 0. Only 0.56 document is viewed for each unique query, and 0.48 was viewed for each result page viewed by the user. In 59.4% of sessions, no single document was viewed by the user; in 28.3% of sessions, only one document was viewed. The distribution is shown in Figure 8. As can be seen, once again the distribution is highly skewed toward the lower end, suggesting that most users viewed only a very small number of result pages during a search session.

The result is quite surprising, even though it is consistent with the findings of Jones and colleagues (1998), which suggested that users did not view the actual document content in 64% of the queries submitted to their Web-based digital library. We are not able to compare our findings with those of other large-scale Web search projects such as the AltaVista and the Excite studies because such data were not collected or analyzed in those studies. Were most users able to find the document they wanted on the basis of the title and snippet displayed in the result pages? Or were the results so bad that the users did not bother to click on them and look at the content? Did they give up and leave the site, or did they try to

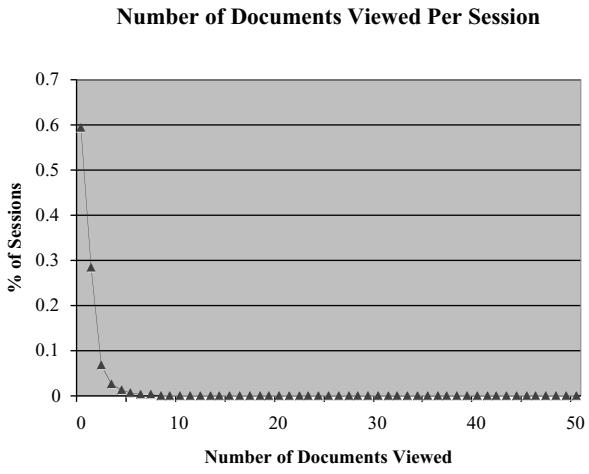

FIG. 8. Number of documents viewed per session.

locate the relevant document themselves by browsing? These questions indicate an interesting research topic for future study.

Advanced Search Features and Default Settings

As suggested by previous Web search studies, it is interesting to study how advanced search features such as Boolean operators are used when users formulate Web queries. The Utah state government Web site search engine provides two different types of advanced search features. In the first type, the user can choose among four options when performing a search: "Exact phrase," "Free-text query," "All of these words," and "Advanced query." The "Exact phrase" search is the default option, which adds a pair of quotation marks to the search query. The "Free-text query" allows users to search for documents containing any terms in the query submitted (a Boolean "OR" search). If the user chooses "All of these words," the engine searches for documents containing all of the terms in the query submitted, but not necessarily appearing as a phrase (a Boolean "AND" search). The "Advanced query" allows users freely to specify AND, OR, NOT, quotation marks, and other operators in their search queries.

To use the second type of advanced search feature, the users need to click on the "Advanced Search" link on the main search page. A new search page is displayed with a more complex search form. Using this form, the users can perform searches on different fields of documents (e.g., title, body, URL, author). The users can also specify a range of dates when the document was last updated, choose the number of results to be shown in one result page, and select how the results should be ordered.

Our findings are summarized in Table 11, with data from the Excite study (Spink et al., 2001) listed for comparison. It can be seen that the Boolean operators AND, OR, and NOT were not used much; AND was used most frequently, appearing in 2.6% of queries. This finding is consistent with the results of the general-purpose Web search engine Excite (Jansen et al., 2000; Spink et al., 2001).

About 3.4% of queries utilized the advanced search features provided in the complex search form. It is surprising to find that this percentage is larger than that of use of Boolean operators. It suggests that a notable percentage of users prefer to utilize advanced search functions, such as searching in different fields, in performing Web searches to satisfy their information needs. As other studies did not report data about this aspect, it would be interesting to study this issue further.

TABLE 11. Usage of operators and advanced search features.

	This study		Excite study	
Feature	Number of queries	Percentage of queries	Number of queries	Percentage of queries
AND	20,206	2.6%	29,146	3%
OR	657	0.08%	1,149	1%
NOT	692	0.09%	307	0.0003%
+ (plus)	154	0.0002%	44,320	5%
− (minus)	364	0.0005%	21,951	2%
" "	230,094	29.0%	52,354	5%
( )	811	0.1%	(Not reported)	(Not reported)
Advanced Search	26,744	3.4%	(Not reported)	(Not reported)
Any of the above	272,474	34.4%	(Not reported)	(Not reported)

It is also interesting to note that although three of the operators listed in Table 11 (plus, minus, and parentheses) are not supported by the search engine, they are still used by some users. The most likely reason is that these operators are supported by several popular search engines such as AltaVista, Excite, and Google (though with different interpretations). It is possible that these users just assumed the Utah search engine would support these operators and thus utilized them in their search queries. A similar finding was also reported in the Excite study, in which some 6% of users used colons and periods, which are not supported by the Excite search engine.

A surprising result in the analysis is the use of quotation marks in the search queries. The data showed that 29.0% of queries used quotation marks. This result is very different from the results reported in the Excite study, in which only 5.1% of queries made use of quotation marks. We found that the most possible cause of the large discrepancy is the default settings of the Utah search engine. When a user performs a search without making any changes to the search options, the "Exact phrase" search option is used as default and a pair of quotation marks are added to the queries. As many users do not change the default settings in information retrieval systems (Jones et al., 1998), the quotation marks are used exceptionally frequently.

In total, 34.4% of queries utilized at least one of the search features. The number is much higher than the 20.4% reported in the AltaVista study, again because of the high usage of quotation marks in the search queries.

Discussions

Key Findings

In general, we found that Web users behave similarly when using a Web site search engine and a general-purpose search engines in terms of the average number of terms per query and the average number of result pages viewed per sessions. However, the users of the Web site search engine show a lower number of queries per session and a different set of terms and topics used in their queries. We suggested the possible reason for these two differences is that users of Web site search engines have more specific information needs than those of general-purpose search engines. We also found that the search terms in a Web site search engine follow the Zipf distribution more closely than those in a general-purpose search engines, indicating that there is a smaller number of "rare terms" that are only used once or twice in our query log.

In studying users' behavior and usage patterns, we found that on average less than one document was actually viewed among the search results presented to the users. Although the

result also has been reported in another Web site search engine study (Jones et al., 1998), it has not been tested in large-scale search log study on general-purpose search engines, and the numbers (in both the study of Jones and coworkers [1998] and the present study) are much less than those reported in a monitored Web search study (Spink, 2002). Do people search differently in a real-world setting compared with an experimental environment? Further research is needed to explain the discrepancy found in these studies.

Implications for Web Site Designers

The similarities and differences between general-purpose search engines and Web site search engines can have important implications for the design of these search engines. Specifically, it is important to customize and improve Web site search engines on the basis of transaction log analysis. For example, as the transaction log in this study has revealed that many users rely on the default settings of the search engine without modification, it is important for search engine developers to ensure that the default options, such as Boolean operators, phrase search option, and the number of search results per page, are the most suitable setting for most users.

It is also important to note that many users type their search queries in the query box without carefully looking at the original content in the box. As discussed earlier, we found that many search queries contain the phrase "Search Utah.gov," which is the default text in the search box that tells users that searches can be performed. Web site designers should implement the search box carefully (e.g., by using client-side script) in order to prevent users from sending this type of query to the search engine becasue their doing so would result in poor search results and higher Web server load.

We also showed that a large proportion of people are looking for information related to a small number of topics in Web site search engines, e.g., tax and Department of Motor Vehicles. Web site designers can learn more about users' most-wanted information resources by analyzing the search logs or the Web access logs of a Web site. Web site designers should make the links to these resources easily accessible by users, e.g., by placing them prominently in the first page of the Web site.

Conclusion and Future Directions

In this article, we report our research on analyzing and mining the transaction log of a Web site search engine. The log data of the Web site search engine of the Utah state government was used as our test data. In our study of search terms and search topics, one limitation of analyzing term association on the basis of the word level is that we could not determine exactly how the words were used in the queries as phrases. Also, it is desirable to know the association between phrases rather than between terms (e.g., the association between the phrase "Department of Motor Vehicles" and other noun phrases). Future research can address this need by applying noun phrase extraction techniques, such as the Arizona Noun Phraser (Tolle & Chen, 2000), to search log data.

As discussed earlier, our study also found that on average users viewed less than one document among the search results presented to them. This number is much smaller than the one reported in Spink (2002), in which the users performed searches in a monitored environment. Important potential research topics are why users view such a small number of documents in their Web searches and how their search behaviors differ in a real-life search task versus an experimental setup. It is also interesting to study how these statistics differ across the Web site search engines in different domains.

As many countries have launched projects on e-government (or digital government), more and more government agencies are putting their information on the Web. The findings reported in this study also have important implications for e-government research by showing how governments can provide the general public with better access to their Web-based information by designing better Web site search engines. The analysis of the search logs helps us better understand what users are looking for on government Web sites. Although most users search for general information such as tax forms or governmental departments, further analysis of government search logs would be needed to investigate how users search for sensitive security information on government Web sites.

Acknowledgments

We gratefully thank the Utah state government and its provider, Utah Interactive, for kindly providing us the data and information used in this study. We also thank the two anonymous reviewers for their insightful suggestions.

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. Menlo Park, CA: AAAI Press.

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 Seventh WWW Conference. Retrieved August 24, 2000, from http://www7.scu.edu.au

CBS News. (2003, March 5). Court allows states to throw the book. CBS News, March 5, 2003. Retrieved August 30, 2004, from http://www.cbsnews.com/stories/2003/03/05/supremecourt/main542863.shtml

Center for Digital Government. (2003). Utah State portal ranks no. 1. Retrieved August 30, 2004, from http://www.centerdigitalgov.com/center/highlightstory.phtml?docid=69811

Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused crawling: A new approach to topic-specific Web resource discovery. In Proceedings of the Eighth International World Wide Web Conference. Retrieved August 30, 2004, from http://www8.org/w8-papers/5a-search-query/crawling/index.html

Chau, M., & Chen, H. (2003a). Comparison of three vertical search spiders. IEEE Computer, 36(5), 56–62.

Chau, M., & Chen, H. (2003b). Personalized and focused Web spiders. In N. Zhong, J. Liu, & Y. Yao (Eds.), Web Intelligence (pp. 197–217). Heidelberg, Germany: Springer-Verlag.

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 [Special Issue on Web Retrieval and Mining], 35(1), 167–183.

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.

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 SIGCOMM Conference on Applications, Technologies, Architectures and Protocols for Computer Communications (241–253).

Croft, W.B., Cook, R., & Wilder, D. (1995). Providing government information on the Internet: Experiences with THOMAS. In Proceedings of the Digital Libraries '95 Conference (pp. 19–24). Retrieved August 30, 2004, from http://csdl.tomu.edu/DL95/contents.html

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.

Fenstermacher, K.D., & Ginsburg, M. (2003). Client-side monitoring for Web mining. Journal of the American Society for Information Science and Technology, 54(7), 625–637.

Hurst, M. (2001). Layout and language: Challenges for table understanding on the Web. In Proceedings of the First International Workshop on Web Document Analysis (pp. 27–30). Retrieved August 30, 2004, from http://csc.liv.9c.uk/~wda2001/

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 Third ACM Conference on Digital Libraries (pp. 293–294). New York: ACM Press.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms (pp. 668–677). Philadelphia: Society for Industrial and Applied Mathematics.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., et al. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks [Special Issue on Neural Networks for Data Mining and Knowledge Discovery], 11(3), 574–585.

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

Montgomery, A.L., & Faloutsos, C. (2001). Identifying Web browsing trends and patterns. IEEE Computer, 34(7), 94–95.

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C.D. (2003). Web usage mining as a tool for personalization: A survey. User Modeling and User-Adapted Interaction, 13, 311–372.

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. (2002). A user-centered approach to evaluating human interaction with Web search engines: An exploratory study. Information Processing and Management, 38, 401–426.

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.

Tolle, K.M., & Chen, H. (2000). Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science, 51(4), 352–370.

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.

Zamir, O., & Etzioni, O. (1999). Grouper: A dynamic clustering interface to Web search results. In Proceedings of the Eighth World Wide Web Conference. Retrieved August 30, 2004, from http://www8.org/w8-papers/3a-search-query/dynamic/dynamic.html