By MICHAEL CHAU, XIAO FANG, and OLIVIA R. LIU SHENG

WHAT ARE PEOPLE SEARCHING ON GOVERNMENT WEB SITES?

A study of search activity on the Utah.gov Web site.

The U.S. government provides a large amount of information to the public on the Web. While the Freedom of Information Act requires the government and federal agencies to disclose a great deal of information to the public, the Paperwork Reduction Act allows these agencies to maintain and provide information through the Internet. Due in part to this Act, large amounts of government information have been put online and made publicly accessible. The provision of information was generally considered harmless before Sept. 11, 2001. After the terrorist attacks that day, the U.S. government and agencies became more concerned about what information was put online, and a lot of sensitive information was removed from the Web shortly thereafter. This includes reports on vulnerabilities related to plants and control measures, maps of power plants or water systems, emergency response plans, and so forth [4]. Terrorist access to such information can potentially facilitate terrorist attacks and become harmful to the public.

However, arbitrarily removing information from the government Web sites is not the best solution to the problem. If too much information is removed, some legitimate requests for information may not return results, causing inconveniences to the public. For example, members of the general public also need to know the location of nuclear plants for the

sake of their own safety and concerns. It is important to study what information people are looking for on these sites in order to determine what information should be made more easily accessible (or otherwise) to the general public. For example, if Web log analyses indicate the public is interested in some sensitive security information (such as nuclear plant information) but government agencies decide such information should only be accessed through legitimate requests and remove it from their Web sites, they should provide other secure means for the public to access this information (for example, after proper authentication and authorization at physical locations).

Number of search queries	792,103
Number of non-empty queries	673,807
Number of unique users	161,042
Number of sessions	458,962
Mean number of queries per session	1.73
Median number of queries per session	1

Table 1. Overview of the search log data.

One way to study people's information needs is to analyze the logs of search engines. Search engine log analysis has been conducted on many general-purpose search engines. The most popular one is the study of the Excite search log. Three single days of search logs (sampled in 1997, 1999, and 2001) of the Excite search engine (www.excite.com) were made available to researchers and many studies have been reported [8, 10, 11]. These analyses have provided much information about the information needs and searching behavior of search engine users, including their search topics and search characteristics. Wang et al. also reported their study of the information needs of the users of an academic search engine [12]. The results of these analyses, however, are not specific to government Web sites.

There are several reasons for analyzing the logs of government search engines. For example, it has been shown that users' behavior in Web site search engines can be quite different from that of general-purpose search engines [2, 12]. Search topics and other query characteristics could be significantly different and research is needed to explore such unique behavior on government Web sites. Studying the search logs can reveal the public's information needs on government Web sites, which allows governments to better manage the content on their Web sites and support their online visitors [5, 6]. Another reason to study the search logs is because government Web sites may store information relevant to public safety and national security, the study can possibly reveal if there is any suspicious activity on their Web sites.

In this article, we report our study on analyzing the search logs of the Utah State Government Web site Utah.gov. The Utah.gov site is one of the most advanced government Web sites and was named the best state government Web portal in the U.S. by the Center for Digital Government in 2003 [1]. A large variety of information and e-services are provided on this Web site. Therefore, it was chosen as the basis for our study. This project is also part of the Utah State's Center of Excellence Program funded in part by the NSF. As discussed, our goal is to study people's search patterns for government information, such as top queries, query term distribution, and session analysis. Such analysis will help government agencies to better understand the public's information needs, improve the design of their Web sites, and possibly uncover suspicious activities occurring on their Web sites. Here, we describe the data collection process, the characteristics of the data, and the analysis method as well as the query analysis results.

Rank	Utah.gov			AltaVista		
	Search query	Frequency	Percentage	Search query	Frequency	Percentage
1	dmv	3,794	0.56%	sex	1,551,477	0.27%
2	tax forms	2,532	0.38%	applet	1,169,031	0.20%
3	sex offenders	2,173	0.32%	porno	712,790	0.12%
4	forms	2,036	0.30%	mp3	613,902	0.11%
5	jobs	1,587	0.24%	chat	406,014	0.07%
6	divorce	1,400	0.21%	warez	398,953	0.07%
7	unemployment	1,359	0.20%	yahoo	377,025	0.07%
8	employment	1,257	0.19%	playboy	356,556	0.06%
9	notary	1,061	0.16%	xxx	324,923	0.06%
10	secretary of state	1,053	0.16%	hotmail	321,267	0.06%

Table 2. Comparison of Top 10 queries with AltaVista.

The Utah State Government Web Site

We collected more than one million search queries submitted to Utah.gov from March 1, 2003 to August 15, 2003 [2]. Utah.gov has a Web site search engine accessible from the main page of the site. Site visitors can enter search queries in the text

USERS OF GENERAL-PURPOSE SEARCH ENGINES HAVE MUCH BROADER INFORMATION NEEDS THAN WHAT IS PROVIDED BY GOVERNMENT SEARCH ENGINES.

box provided and submit the queries to the Web site search engine. Our search log contains a total of 1,895,680 records. Each record represents a request that 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. Each record consists of 14 fields, such as timestamp, IP address, the type of request submitted, and other parameters of the request.

We extracted the search queries from the data and used information on cookies and IP addresses to identify users from the data. Following previous research, the sessions were identified from the user data using the widely applied rule of thumb, in which the maximal session length should be less than 30 minutes [3]. Each session was assigned a unique session id in our database; there are 792,103 queries in total, submitted by a total of 161,042 unique users in 458,962 sessions. Each session has on average 1.73 queries, or 1.25 unique queries. The latter number is much lower than the number 2.52 reported in the Excite study [11] and 2.02 reported in the AltaVista study [9]; the results are summarized in Table 1.

Suspicious Term	Number of Queries Containing the Term	Number of Sessions Containing the Term
terrorism	119	83
sars	118	92
nuclear	116	79
west nile virus	95	77
water system	62	29
emergency AND plan	61	35
power plant	58	28
radioactive	39	19
smallpox	37	27
disease control	15	12
nuclear AND map	3	1
pipeline AND map	1	1
anthrax	1	1

Table 3. Frequencies of terms potentially related to terrorism.

Analysis and Discussion

When we compared our data with that of previous search log analysis for general-purpose search engines, we found that Web users behave similarly when using a government Web site search engine and a general-purpose search engine regarding the average number of terms per query and the average number of result pages viewed per sessions [2]. However, they show a lower number of queries per session and a different set of terms and topics used in their queries. In our study, we identified the top 25 most frequent queries in our data and compared the queries with that of AltaVista [9]; here, we correlate the top 10 of the search results between Utah.gov and AltaVista in Table 2. As one can expect, the top queries submitted to the government Web site search engines are different from those in general-purpose search engines. The government Web site queries are mostly related to people's general information needs from the government, such as the Department of Motor Vehicles and tax forms. The results suggest the public frequently relies on the government Web site to obtain useful information relevant to their daily activities. It is also interesting to note the different distribution of the queries in the two studies. In addition, we found that the top 25 queries in our data accounted for 4.48% of all non-empty queries, while the top 25 in AltaVista only represented 1.56% of their data. The difference indicates the queries in government Web search engines (and other Web site search engines) are less diverse. In other words, users of general-purpose search engines have much broader information needs than what is provided by government search engines. It is therefore desirable to customize the design of government Web sites and their search engines by making some prominent links to popular informa-

THE TOP QUERIES AND SEASONAL EFFECT ANALYSIS REVEAL THE INFORMATION NEEDS OF THE GENERAL PUBLIC ON GOVERNMENT WEB SITES.

tion in order to better cater to users' needs.

We also revealed some interesting seasonal patterns in our query logs. Seasonal effect has been demonstrated in other search engines such as university Web search engines [12]. For example, the query "career services" occurred more frequently in February, March, September, and October than in other months. Similarly, the query "football tickets" appeared most often in August and September. In our data, we found some similar patterns for terms related to information needs that are "seasonal." An example of the search for tax-related queries (all queries that contain the terms "tax," "irs," or "internal revenue") is shown in Figure 1. The number reached its peak on April 15 (the deadline for filing individual tax returns in the U.S.) and dropped quickly afterward.

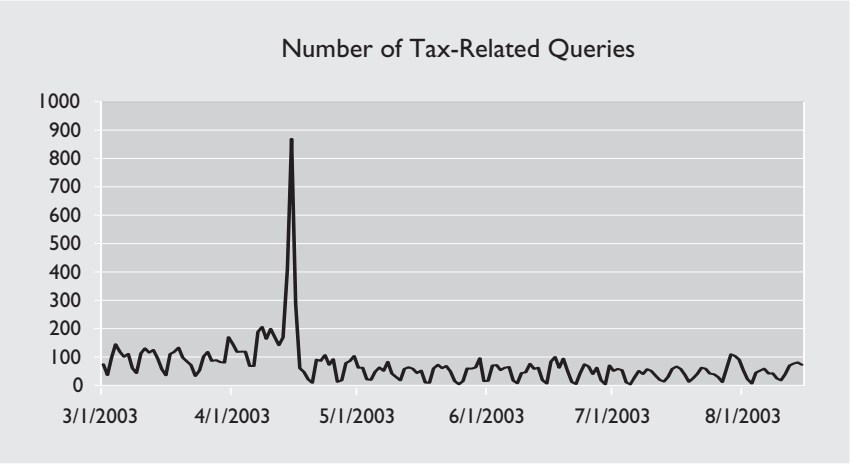

Figure 1. Seasonal effect of tax-related queries.

The top queries and seasonal effect analysis reveal the information needs of the general public on government Web sites. However, they do not show the queries of people with specific purposes (such as terrorists) because these queries would not appear frequently in the data. In order to study the search of security-related information on government Web sites, we identified a set of "suspicious" terms with the help of a researcher specializing in terrorist research in the U.S. and analyzing the occurrences of these terms in the search log data. The list of these terms and their frequencies is shown in Table 3.

All the 13 suspicious terms exist in the search logs, indicating there are users who entered these queries into the search engine to look for information relevant to these queries. While we have no way to determine the real intent of these users—whether they are searching for such information for legitimate use or for other purposes, it is important to investigate what information they are looking for.

The first few of these queries are related to nuclear and radioactive plants and substances. The U.S. government was widely criticized for putting detailed nuclear plant maps and nuclear substance transportation routes on government Web sites during the aftermath of the Sept. 11 attacks because such information potentially allows terrorists to easily target such facilities for attack. When we took a closer look at these queries, we found that some of them may not be submitted by the general public, such as "radioactive waste storage" and "nuclear waste transportation route map."

Because a lot of sensitive security information has been removed from government Web sites, a search on the Utah.gov Web site with these queries did not return any sensitive security information. While our manual analysis showed that the Utah.gov Web site currently does not contain any sensitive security information, we are unsure of the situation of other government Web sites. The U.S. will be vulnerable if terrorists could find details of such information easily.

Some other queries are related to the water system in state of Utah. One particular user searched for "map of pipeline." Details of water systems and pipeline maps are also important for terrorism activities because terrorists can easily poison water supplies according to such information. It is believed that Al Qaeda had plans to poison U.S. water supplies and possessed some important information about the country's water system [7]. Other queries searched for diseases that could be used in terrorism attacks, such as SARS, West Nile virus, anthrax, and smallpox. While it is possible that these queries were submitted by terrorists, we should note that it is equally possible these queries were submitted by some general citizens who wanted to know whether there were any reported cases of these diseases in the state of Utah. There are also queries that are related to disease control or emergency response planning. Again, there are two possible scenarios: terrorists are assessing the state government's ability in handling terrorism attacks (or lack thereof) or the general public is looking for relevant information in this aspect.

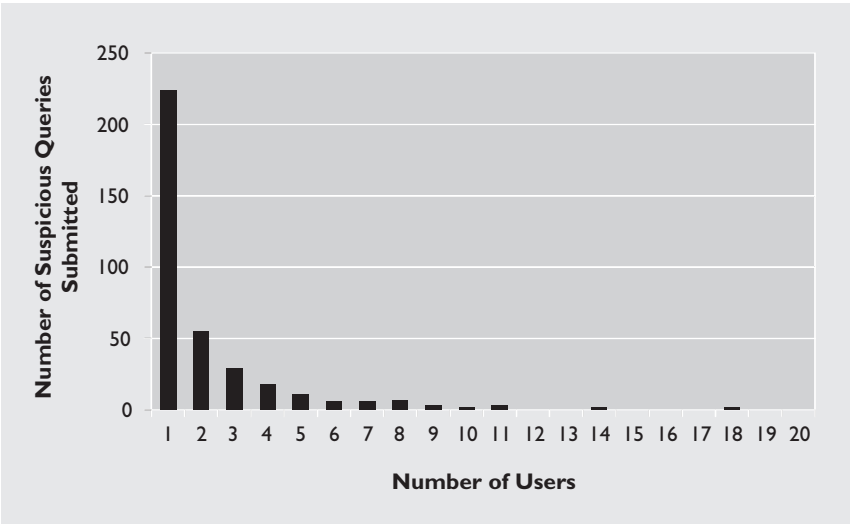

Figure 2. Distribution of the number of suspicious queries submitted by users.

As discussed earlier, we have used information on cookies and IP addresses to identify users from the data. The originators of these suspicious terms are associated with a small number of users. They were queried by 355 unique users, only 0.22% of the total population in this study. We further analyzed the distribution of the number of queries containing suspicious terms submitted by each of these users. As shown in Figure 2, each of these users submitted an average of 2.1 suspicious queries, with a maximum of 18. Although there is no extremely high number of queries containing suspicious terms submitted by any single user in this study, this kind of analysis could be useful in providing alerts by identifying users who submit a substantial number of suspicious queries.

Implications

While some of these queries were made by the general public for legitimate reasons (such as environmental protection or vaccination), it is possible that some queries were made by terrorists for planning possible attacks. We suggest that such information is desired by both parties—terrorist groups may use such information for planning attacks and the general public needs this information for better protecting the citizens. It is important for government to satisfy the information needs of legitimate users while keeping terrorists from accessing sensitive information. This issue has inspired the question of how much information should be made available for online access and how to control access to sensitive information. One possible solution is to put non-sensitive information online while keeping sensitive information accessible only after verifying identification (or security clearance if needed) at government information offices.

Another solution is to put all information on the Web but restrict access to sensitive information by measures like password control. This will allow authorized parties to retrieve information more easily, but will be more prone to security threats such as hacking. Further research will be needed in the area of information security, such as user profiling and automatic detection and alerting of suspicious activities using data mining techniques.

Our analysis also indicates that a significant proportion of people are looking for information

related to a small number of topics in government Web sites, like tax information and the Department of Motor Vehicles. For example, the term "tax" appeared in 3.59% of all queries. To allow the general public to access online government information more easily, the Web site designers can analyze the search logs or the Web access logs. This data can reveal more about users' most-wanted information resources and make the links to these resources easily accessible by users, say, by placing them prominently in the first page of the Web site. For example, the LinkSelector technique, which has been successfully applied to a university's Web site, can be used to select the most appropriate set of links on the home page of a Web site in order to maximize the efficiency and effectiveness of a Web site's usage based on log analysis [5]. Such techniques can be effectively applied in e-government projects to improve the performance of government portals.

Conclusion

In this article, we have reported our research on analyzing the query log of a government Web site search engine and we found that some terrorism-related queries do exist in our data. A limitation of this study is that the analysis was only performed on one government Web site. Nevertheless, this problem must be taken seriously by governments in their information policies. On the other hand, as many countries have launched e-government (or digital government) projects, increasing numbers of government agencies are putting their information on the Web. The analysis of the search logs helps us better understand what users are seeking on government Web sites. Based on our study, we make the following suggestions to government agencies:

- Perform search log analysis to determine users' search behaviors and monitor for suspicious information requests.
- Make the most requested information more easily accessible to the public by putting it on the first page of the Web site or creating prominent navigation links.
- Develop a clear classification on what kinds of information are potentially vulnerable to the country's security and establish guidelines on what classes of information should be made accessible online. Information with mid-level or high-level sensitivity should be made available to an individual or organization only after proper authentication and/or security clearance.

While removing vast amounts of information from government Web sites is not a good solution to the problem, it is not easy to strike a balance between providing easy access to information to the public and preventing terrorists from gaining access to sensitive information. We hope the suggestions proposed here will help alleviate the problem in government Web sites. C

References

1. Center for Digital Government. Utah State Portal ranks No. 1 (2003); www.centerdigitalgov.com/center/highlightstory.phtml?docid=69811.
2. Chau, M., Fang, X., and Sheng, O.R.L. Analysis of the query logs of a Web site search engine. *Journal of the American Society for Information Science and Technology 56*, 13 (2005), 1363–1376.
3. Cooley, R., Mobasher, B., and Srivastava, J. Data preparation for mining World Wide Web browsing patterns. *Knowledge and Information Systems 1*, 1 (1999).
4. Electronic Frontier Foundation. Chilling effects of anti-terrorism: 'National security' toll on freedom of expression, (2004); www.eff.org/Privacy/Surveillance/Terrorism/antiterrorism_chill.html.
5. Fang, X. and Sheng, O.R.L. LinkSelector: A Web mining approach to hyperlink selection for Web portals. *ACM Transactions on Internet Technology 4*, 2 (2004), 209–237.
6. Fang, X. and Sheng, O.R.L. Designing a better Web portal for digital government: A Web-mining based approach. In *Proceedings of the 2005 National Conference on Digital Goverment Research (dg.o 2005)*, (Atlanta, GA, 2005).
7. Feds arrest Al Qaeda suspects with plans to poison water supplies. Fox News (July 30, 2002); www.foxnews.com/story/0,2933,59055,00.html.
8. Jansen, B.J., Spink, A., Bateman, J., and Saracevic, T. Real life information retrieval: A study of user queries on the Web. *ACM SIGIR Forum 32*, 1 (1998), 5–17.
9. Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. Analysis of a very large Web search engine query log. *ACM SIGIR Forum 33*, 1 (1999), 6–12.
10. Spink, A., Jansen, B.J., Wolfram, D., and Saracevic, T. From e-sex to e-commerce: Web search changes. *IEEE Computer 35*, 3 (Mar. 2002), 107–109.
11. Spink, A., Wolfram, D., Jansen, B.J., and Saracevic, T. Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology 52*, 3 (May 2001), 226–234.
12. Wang, P., Berry, M.W., and Yang, Y. Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology 54*, 8 (Aug. 2003), 743–758.

Michael Chau (mchau@business.hku.hk) is an assistant professor in the School of Business at the University of Hong Kong, Hong Kong.
Xiao Fang (xiao.fang@utoledo.edu) is an assistant professor in the College of Business Administration at the University of Toledo, OH.
Olivia R. Liu Sheng (olivia.sheng@business.utah.edu) is a Presidential Professor and the Emma Eccles Jones Presidential Chair of Information Systems at the University of Utah, Salt Lake City, UT.

This research was supported in part by National Science Foundation Grant No. 0410409.