From E-Sex to E-Commerce: Web Search Changes

Amanda Spink, Pennsylvania State University
Bernard J. Jansen, US Army War College
Dietmar Wolfram, University of Wisconsin-Milwaukee
Tefko Saracevic, Rutgers University

The Web has become a worldwide source of information and a mainstream business tool. Are human information needs and searching behaviors evolving along with Web content? As part of a body of research studying this question, we have analyzed three data sets culled from more than one million queries submitted by more than 200,000 users of the Excite Web search engine, collected in September 1997, December 1999, and May 2001.

This longitudinal benchmark study shows that public Web searching is evolving in certain directions. Specifically, search topics have shifted from entertainment and sex to commerce and people, but there is little change in query lengths or frequency per user.

SCOPE OF STUDY

Excite (http://www.excite.com) is a major Internet media company offering Web searching and a personalization portal. Its searches are based on the exact terms a user enters in a query. Capitalization is disregarded with the exception of logical commands AND, OR, and AND NOT. There is no stemming. Our study is limited to analysis of users' queries, as we had no access to data on the Web sites they accessed.

Obtaining large-scale query logs from commercial Web search engines is not an easy task. While using data only from Excite is a limitation on our study, our ongoing analysis of this search engine over a four-year period nevertheless provides a baseline for insights into Web searching trends.

Search topics have shifted, but there is little change in user search behaviors.

Each Excite query log record contained three fields:

- *Identification*—anonymous code assigned by the Excite server to a user machine.
- *Time of day*—in hours, minutes, and seconds.
- *Query*—user terms as entered.

We analyzed the following user data:

- *Sessions*—entire query sequence by a user.
- *Queries*—one or more entered terms.
- *Terms*—any string of characters bounded by white space.

Table 1 summarizes the three data sets. Note that we removed duplicate queries from the initial results, so values calculated in Tables 2 and 3 reflect the distilled set of distinct queries.

RESULTS

Table 2 shows little change over the four-year period in terms per query, queries per user, or pages per query—although queries per user took a dip in 1999.

More than 50 percent of 2001 users submitted a single short query, about 20 percent submitted two queries, and another 29 percent entered three or more unique queries. Users typically do not add or delete many terms in their subsequent queries.

A fluctuating percent of users modify queries, with 52 percent modifying queries in 1997, declining to 39.6 percent in 1999, and increasing to 44.6 percent in 2001. Users tend to move from broad to narrow terms when modifying their queries by changing some individual terms. However, the total number of terms often remains the same.

The continuing use of one short simple query suggests that information content providers can expect to reach Web users by targeting specific high-frequency words such as "free," "sex," "games," "weather," and "maps."

Fewer results per query

The trend since 1997 shows users viewing fewer pages of results per query. An Excite results page contains 10 ranked Web sites, and the percentage of Excite users who examined only one page of results per query increased from 28.6 percent in 1997 to 50.5 percent in 2001. By 2001, more than 70

Table 1. Excite data sets for 1997, 1999, and 2001.

Data set	Sessions	Queries	Terms
1997[1]	211,063	1,025,908	1,277,763
1999[2]	325,711	1,025,910	1,500,500
2001	262,025	1,025,910	1,538,120

1. B.J. Jansen, A. Spink, and T. Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," *Information Processing and Management*, vol. 86, no. 2, 2000, pp. 207-227.
2. A. Spink et al., "Searching the Web: The Public and Their Queries," *J. Am. Soc. Information Science and Technology*, vol. 53, no. 2, 2001, pp. 226-234.

Table 2. Comparative statistics for Excite Web query data sets—one million queries per study.

Variables	1997	1999	2001
Mean terms per query	2.4	2.4	2.6
Terms per query			
1 term	26.3%	29.8%	26.9%
2 terms	31.5%	33.8%	30.5%
3+ terms	43.1%	36.4%	42.6%
Mean queries per user	2.5	1.9	2.3
Mean pages viewed per query	1.7	1.6	1.7
Pages viewed per query			
1 page	28.6%	42.7%	50.5%
2 pages	19.5%	21.2%	20.3%
3+ pages	51.9%	36.1%	29.2%
Users modifying queries	52.0%	39.6%	44.6%
Session size			
1 query	48.4%	20.8%	30.8%
2 queries	60.4%	19.8%	19.8%
3+ queries	55.4%	19.3%	25.3%
Boolean queries	5.0%	5.0%	10.0%
Terms not repeated in the data set	57.1%	61.6%	61.7%
Use of 100 most frequently occurring query terms	17.9%	19.3%	22.0%

percent of Excite users looked at two pages or fewer.

Was it that users were satisfied with the results and had no need to view more pages? The trend toward viewing fewer results, combined with the small number of pages viewed and queries per session, suggests that Excite users want more relevant Web sites per total number of sites retrieved. Some users continue to have low tolerance for wading through large retrievals.

The continuing low and declining level of user interactivity is a challenge for users and Web search engine designers alike. Counter to these trends toward greater simplicity, the use of Boolean operators increased from 5 percent to 10 percent from 1997 to 2001. A recent longitudinal study of 20,000 Internet users reported little change in Web searching session times (A.L. Montgomery and C. Faloutsos, "Identifying Web Browsing Trends and Patterns," *Computer*, July 2001, pp. 94-95).

Despite some high-frequency terms, an unusually large number of terms either are never repeated or are used with low frequency. These include personal names, spelling errors, non-English terms, and Web-specific terms such as URLs.

The Web query vocabulary contains a very large number of different terms compared with large English texts in general. The language of queries has unique characteristics, which content providers can benefit from studying.

Topic shift

How did Web searching topics change from 1997 to 2001? We classified a random sample of 2,414 queries from 1997; 2,539 queries from 1999; and 2,453 queries from 2001 into 11 nonmutually exclusive, general topic categories. Table 3 shows the results.

There is an ongoing shift in search topics. From 1997 to 2001, categories such as "Entertainment or recreation" and "Health or sciences" moved down the ranking. "Commerce, travel, employment, or economy" and "People, places, or things" moved up.

In 1997, approximately one in six Web queries was about sex. By 2001, this was down to one in 12, and many of these related to human sexuality, not pornography. By 1999, "Commerce, travel, employment, or economy," "People, places, or things," and "Computers or Internet" moved closer to the top of the list, while "Sex and pornography" and "Entertainment or recreation" moved down.

The shift to e-commerce queries coincided with changes in information distribution on the publicly indexed Web. By 1999, some 83 percent of Web servers contained commercial content. By 2001, Web searching and Web content continued to evolve from an entertainment to a business medium.

Interestingly, non-English queries and unclassifiable queries have nearly tripled since 1997. Many queries are single terms such as "naz;" numbers such as "182;" or acronyms, such as "TOF." Without additional terms, it is difficult for Web search engines to interpret such queries.

This longitudinal benchmark study extends previous research suggesting little movement toward longer or more frequent queries (D. Wolfram et al., "Vox Populi: The Public Search of

Table 3. Distribution of query samples across general topic categories.

Rank	1997 Excite data set (2,414 queries)	1999 Excite data set (2,539 queries)	2001 Excite data set (2,453 queries)
1	19.9% Entertainment or recreation	24.5% Commerce, travel, employment, or economy	24.7% Commerce, travel, employment, or economy
2	16.8% Sex and pornography	20.3% People, places, or things	19.7% People, places, or things
3	13.3% Commerce, travel, employment, or economy	10.9% Computers or Internet	11.3% Non-English or unknown
4	12.5% Computers or Internet	7.8% Health or sciences	9.6% Computers or Internet
5	9.5% Health or sciences	7.5% Sex and pornography	8.5% Sex and pornography
6	6.7% People, places, or things	7.5% Entertainment or recreation	7.5% Health or sciences
7	5.7% Society, culture, ethnicity, or religion	6.8% Non-English or unknown	6.6% Entertainment or recreation
8	5.6% Education or humanities	5.3% Education or humanities	4.5% Education or humanities
9	5.4% Performing or fine arts	4.2% Society, culture, ethnicity, or religion	3.9% Society, culture, ethnicity, or religion
10	4.1% Non-English or unknown	1.6% Government	2.0% Government
11	3.4% Government	1.1% Performing or fine arts	1.1% Performing or fine arts

the Web," *J. Am. Soc. Information Science and Technology*, vol. 52, no. 12, 2001, pp. 1073-1074).

EDUCATED QUERIES

Our results show that Web queries from a major commercial search engine continue to be simple in structure with a minority of queries incorporating advanced search features. Many queries that do contain advanced searching operators are mistakes, such as noncapitalized Boolean operators.

Despite commonly retrieving a large number of Web sites, users tend to view few result pages per query. This trend appears to be increasing with the majority of Web users not browsing beyond the first or second page of results.

The language of Web queries is rich and increasingly varied, including people's names, acronyms, and non-English terms, as the subject distribution of Web queries moves closer to increasingly commercial and international Web site content.

Many aspects of searching have remained constant even as the Web evolves in size, content, and services. People are seeking to resolve their information problems via search engines that cater to human-computer interaction on a massive scale. We need a new generation of Web searching tools based on a more thorough understanding of human information behaviors. Such tools would assist users with query construction and modification, spelling, and analytical problems that limit their ability or willingness to persist in finding the information they need.

As the Web evolves into an international economic resource, users encounter corresponding risks if they blindly trust the capabilities of Web search engines to retrieve good data from a few key words. Better Web design must be met on the user side by more effective search behaviors.

Our results provide important insights into the state of Web information and search, which may affect economic, social, and political issues in the future. Critically, we see little change in user search strategies, coupled with ongoing user frustration with the search process. These results pose challenges to Web designers and to organizations that depend on increasing Web access by business, medical, educational, and scientific users, as well as to Web users themselves.

Our studies are intended to motivate the design of more effective tools to counter the large-scale public searching behaviors that remain the same even as the Web becomes more complex. In addition, our work may encourage users to develop more effective searches. We are currently expanding our study with analysis of large-scale query logs from Fast.no, a European Web search engine. ■

Acknowledgments

We thank Doug Cutting, Jack Xu, and Soo Young Rieh from Excite@Home.com, and C. Lee Giles and Dan Lorence of Pennsylvania State University for their useful comments and suggestions.

Amanda Spink *is an associate professor of information sciences and technology at Pennsylvania State University. Contact her at spink@ist.psu.edu.*

Bernard J. Jansen *is a major in the US Army, currently in the Chief Information Office, US Army War College, Carlisle Barracks, Pa. Contact him at jjansen@acm.org.*

Dietmar Wolfram *is an associate professor at the School of Information Studies, University of Wisconsin-Milwaukee. Contact him at dwolfram@ uwm.edu.*

Tefko Saracevic *is a professor in the School of Communication, Information, and Library Studies at Rutgers University. Contact him at tefko@ scils.rutgers.edu.*