BRIEF COMMUNICATION

Vox Populi: The Public Searching of the Web

Dietmar Wolfram
School of Library and Information Science, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201. E-mail: dwolfram@csd.uwm.edu

Amanda Spink
School of Information Sciences and Technology, The Pennsylvania State University, 511 Rider I Building, 120 S. Burrowes St., University Park, PA 16801-3857. E-mail: spink@ist.psu.edu

Bernard J. Jansen
U.S. Army War College, Carlisle Barracks, PA 17013. E-mail: jjansen@acm.org

Tefko Saracevic
School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ 08903. E-mail: tefko@scils.rutgers.edu

In previous articles, we reported the state of Web searching in 1997 (Jansen, Spink, & Saracevic, 2000) and in 1999 (Spink, Wolfram, Jansen, & Saracevic, 2001). Such snapshot studies and statistics on Web use appear regularly (OCLC, 1999), but provide little information about Web searching trends. In this article, we compare and contrast results from our two previous studies of Excite queries' data sets, each containing over 1 million queries submitted by over 200,000 Excite users collected on 16 September 1997 and 20 December 1999. We examine how public Web searching changing during that 2-year time period.

As Table 1 shows, the overall structure of Web queries in some areas did not change, while in others we see change from 1997 to 1999.

Our comparison shows how Web searching changed incrementally and also dramatically. We see some moves toward greater simplicity, including shorter queries (i.e., fewer terms) and shorter sessions (i.e., fewer queries per user), with little modification (addition or deletion) of terms in subsequent queries. The trend toward shorter queries suggests that Web information content should target specific terms in order to reach Web users. Another trend was to view fewer pages of results per query. Most Excite users examined only one page of results per query, since an Excite results page contains ten ranked Web sites. Were users satisfied with the results and did not need to view more pages? It appears that the public continues to have a low tolerance of wading through retrieved sites. This decline in interactivity levels is a disturbing finding for the future of Web searching.

Queries that included Boolean operators were in the minority, but the percentage increased between the two time periods. Most Boolean use involved the AND operator with many mistakes. The use of relevance feedback almost doubled from 1997 to 1999, but overall use was still small. An unusually large number of terms were used with low frequency, such as personal names, spelling errors, non-English words, and Web-specific terms, such as URLs. Web query vocabulary contains more words than found in large English texts in general. The public language of Web queries has its own and unique characteristics.

How did Web searching topics change from 1997 to 1999? We classified a random sample of 2,414 queries from 1997 and 2,539 queries from 1999 into 11 categories (Table 2).

From 1997 to 1999, Web searching shifted from entertainment, recreation and sex, and pornography, preferences to e-commerce–related topics under commerce, travel, employment, and economy. This shift coincided with changes in information distribution on the publicly indexed Web.

Received February 22, 2001; Revised May 14, 2001; Accepted May 14, 2001

Lawrence and Giles (1999) found that by 1999 some 83% of Web servers contained commercial content. In frequency distribution, Web content and Web searching appear to be converging. In addition, not all queries in the sex, pornography, and preferences category were about pornography, but were more related to sexual health and sexuality.

Longitudinal studies of Web searching can provide valuable insights into how public Web searching is evolving, changing, and moving in certain directions. These insights can support Web design and public policy decisions. Our analyses revealed that public Web searching has both changed over time and not changed. In general, both the social context and use framework, as well as design advances, are driving Web searching. The design of Web technology is based on social and technological assumptions of human information behavior (HIB) and human–computer interaction (HCI) on a massive scale. As Web content continues to expand and evolve, and Web searching grows, the question is how changes in human information needs and behaviors can best be identified and supported by Web technologies.

TABLE 1. Comparative statistics across 1997 and 1999 Web query data sets.

Variables	1997 Excite study (>1M queries)	1999 Excite study (>1M queries)
Mean terms per query	2.4	2.4
Terms per query		
1 term	26.3%	29.8%
2 terms	31.5%	33.8%
3+ terms	43.1%	36.4%
Mean unique queries per user	2.5	1.9
Mean pages viewed per query	1.7	1.6
% of users who viewed		
1 page	28.6%	42.7%
2 pages	19.5%	21.2%
3+ pages	51.9%	36.1%
% of users who modified queries	52%	39.6%
Session size: Unique queries		
1 query	48.4%	60.4%
2 query	20.8%	19.8%
3+ queries	30.8%	19.8%
% of Boolean queries	5%	8%
% of queries: Relevance feedback	5%	9.7%
% of terms unique to the data set	57.1%	61.6%

TABLE 2. Distribution of sample of queries across subject categories: 1997 and 1999.

1997 Excite data set (2,414 queries)	1999 Excite data set (2,539 queries)
1. Entertainment, recreation, 16.9%	1. Commerce, travel, employment, & economy, 24.4%
2. Sex, pornography & preferences, 16.8%	2. People, places, & things, 20.3%
3. Commerce, travel, employment, & economy, 13.3%	3. Computer & the Internet, 10.9%
4. Computers & the Internet, 12.5%	4. Sex, pornography, & preferences, 7.5%
5. Health & the sciences, 9.5%	5. Health & the sciences, 7.8%
6. People, places, & things, 6.7%	6. Entertainment & recreation, 7.5%
7. Society, culture, ethnicity, & religion, 5.7%	7. Unknown & incomprehensible, 6.8%
8. Education & the humanities, 5.6%	8. Education & the humanities, 5.3%
9. Performing & fine arts, 5.4%	9. Society, culture, ethnicity, & religion, 4.2%
10. Government, 3.4%	10. Government, 1.6%
11. Unknown & incomprehensible, 4.1%	11. Performing & fine arts, 1.1%

References

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

Lawrence, S., & Giles, C.L. (1999). Accessibility of information on the Web. Nature, 400, 107–109.

OCLC. (1999). Web statistics and analysis, Online Computer Library Center. Available at: www.oclc.org/oclc/research/projects/webstats/index.htm.

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, 53, 226–234.