[Information Research, Volume 6 No. 1 October 2000]
Selected results from a large study of Web searching: the Excite study, by Amanda Spink and Jack L. Xu Location: http://www.shef.ac.uk/~is/publications/infres/paper90.html
Last updated: 6th September 2000

RESEARCH NOTE

Selected results from a large study of Web searching: the Excite study

Amanda Spink and Jack L. Xu
School of Information Sciences & Technology, The Pennsylvania State University, University Park, PA 16801, USA
Excite@Home Corporation, USA

Abstract

This paper reports selected findings from an ongoing series of studies analyzing large-scale data sets containing queries posed by Excite users, a major Internet search service. The findings presented report on: (1) queries length and frequency, (2) Boolean queries, (3) query reformulation, (4) phrase searching, (5) search term distribution, (6) relevance feedback, (7) viewing pages of results, (8) successive searching, (9) sexually-related searching, (10) image queries and (11) multi-lingual aspects. Further research is discussed.

Introduction

People are spending increasing amounts of time working with electronic information. Web searching services are now everyday tools for information seeking. However, many Web interactions are often frustrating and constrained. A growing body of large-scale, quantitative or qualitative studies is exploring the effectiveness of Web search engines (Lawrence & Giles, 1998) and how users search the Web (Silverstein, *et al.*, 1999). To support human information behaviours we are seeing the development of a new generation of Web tools, such as Web meta-search engines, to help users persist in electronic information seeking is needed to help people resolve their information problems. Our paper reports selected results from a large-scale and ongoing series of studies of Web users' searching behaviour on the Excite search engine by a diverse range of information and computer scientists. This paper reported selected results from studies of three sets of Excite transaction logs containing: (1) 30 billion queries, (2) 51,473 queries, and (3) 1.2 million queries.

1.**Excite Data Set 1 - 30 Billion Queries**: transaction log data collected and statistically analyzed by Excite researchers from 1996 to 1999 (Xu, 1999). As Excite currently processes over 30 million queries a day, the data analyzed included nearly 30 billion queries.

2. **Excite Data Sets 2 & 3 - 51K and 1 Million+ Queries**: transaction log analysis by Spink, *et al.*, (in press) of over 51,000 and 1,025,910 queries (by 211,063 Excite users, containing 2,216,986 terms) collected on 9 March 1997.

Excite users were anonymous and could not be identified in any way. But, we could identify each user's sequence of queries. Excite searches are based on the exact terms that a user enters in the query; however, capitalization is disregarded, with the exception of logical commands AND, OR, and AND NOT. Stemming is not available. Search results are provided in ranked relevance order. A number of advanced search features are available.

We focused on three levels of analysis - *sessions, queries* and *terms*. Each transaction record contained three fields. With these three fields, researchers were able to locate an Excite user's initial query and recreate the chronological series of actions by each user in a session: *Time of Day:* measured in hours, minutes, and seconds; *User*

Identification: an anonymous user code assigned by the Excite server; *Query Terms*: exactly as entered by the given user. This large-scale study of Web searching provides insights into Web searching with implications for developing better search engines and services.

Selected Findings

Selected findings are summarized below that provide interesting insights into public Web searching, including: (1) queries, (2) Boolean queries, (3) query reformulation, (4) phrase searching, (5) search terms: distribution, (6) relevance feedback, (7) viewing results, (8) successive searching, (9) sexually-related searching, (10) image queries and (11) multi-lingual aspects.

1. Queries

The mean length of Excite queries increased steadily for the years 1st May 1996 to 2nd June 1999 and the mean number of terms in unique queries was 2.4. The mean query length for US, UK, and European users in 1996 was 1.5. In 1999 the figures were - US and UK users 2.6 and European users 1.9. English language queries increased in length more quickly than European language queries. Jansen, *et al.*, (2000) report that Web queries are short and most users did not enter many queries for each search. The mean number of queries for a user was 2.8 in 1997. However, a sizable percentage of users did go on either to modify their original query or to view subsequent results. On average, a query contained 2.21 terms in 1997. About one in three queries had one term only, two in three had one or two terms, and four in five had one, two or three terms. Fewer than 4% of the queries comprised more than six terms.

2. Boolean Operators

The use of Boolean operators (AND, OR, NOT, +, -) increased from 22% of queries in 1997 to 28% of queries in 1999. From the 1996-1999 data set, approximately 8% of searches included proximity searching. Jansen, *et al.*, (2000) found that Boolean operators were seldom used. One in 18 users used any Boolean capabilities, and of the users employing them, every second user made a mistake, as defined by Excite rules. The ‘+’ and ‘-‘ modifiers that specify the mandatory presence or absence of a term were used more than Boolean operators. About 1 in 12 users employed them. About 1 in 11 queries incorporated a ‘+’ or ‘-‘ modifier, but a majority (about two out of three) of these uses were mistakes .

3. Query Reformulation

Spink, *et al.*, (in press) found that most users searched one query only and did not follow with successive queries. The average session, ignoring identical queries, included 1.6 queries. About 2 in 3 users submitted a single query, and 6 in 7 did not go beyond two queries.

4. Phase Searching

Phrases (terms enclosed by quotation marks) were seldom, while only 1 in 16 queries contained a phrase - but correctly used.

5. Search Terms: Distribution

Jansen, *et al.*, (2000) report the distribution of the frequency of use of terms in queries as highly skewed. A few terms were used repeatedly and a lot of terms were used only once. On the top of the list, the 63 subject terms that had a frequency of appearance of 100 or more represented only one third of one percent of all terms, but they accounted for about one of every 10 terms used in all queries. Terms that appeared only once amounted to half of the unique terms.

6. Relevance Feedback

Relevance feedback was rarely used. About one in 20 queries used the feature *More Like This.* Spink, *et al.*, (in

press) found that a third of Excite users went beyond the single query, with a smaller group using either query modification or relevance feedback, or viewing more than the first page of results. They examined the occurrence of each query type (unique, modified, relevance feedback, view a results page, etc.) in a large sample of user sessions. The distribution of query type changes as the length of the user session increase. For the user sessions of two and three queries, the relevance feedback query is dominant. As the length of the sessions increase, the occurrences of relevance feedback as a percentage of all query types decreases. 63% of relevance feedback sessions could be construed as being successful. If the partially successful user sessions are included, then more than 80% of the relevance feedback session provided some measure of success.

7. Viewing Results

Xu (1999) reported that from 1996 to 1999, for more than 70% of the time, a user only views the top ten results. On average, users viewed 2.35 pages of results (where one page equals ten hits). Over half the users did not access result beyond the first page. Jansen, *et al.*, (2000) found that more than three in four users did not go beyond viewing two pages.

8. Successive Searching

Spink, Bateman and Jansen (1999) conducted an interactive survey of over 300 Excite users and found that many had conducted two searches, or three or more related searches using the Excite search engine over time when seeking information on a particular topic. Successive searches often involved a refinement or extension of the previous searches as new databases were searched and search terms changed as the Excite users understanding and evaluation of results evolved over time from one successive search to the next.

9. Sex-Related Searching

Jansen, *et al.*, (2000) found searching about sex on Excite represents only a small proportion of all searches. When the top frequency terms are classified as to subject, the top category is Sexual. As to the frequency of appearance, about one in every four terms in the list of 63 highest used terms can be classified as sexual in nature. But while sexual terms are high as a category, they still represent a very small proportion of all terms. Many other subjects are searched and the diversity of subjects searched is very high. See Spink & Ozmultu (forthcoming).

10. Image Searching

Goodrum & Spink (1999) conducted a specific analysis of image queries within the 1.2 million queries. Provisions for image searching by Web search engines is important for users. Users seeking images input relatively few terms to specify their image information needs on the Web. Users seeking images interact iteratively during the course of a single session, but input relatively few queries overall. Most image terms are used infrequently with the top term occurring in less than 9% of queries. Jansen, *et al.*, (2000) found that many terms were unique in the large data sets, with over half of the terms used only once. Terms indicating sexual or adult content materials appear frequently in image queries. They represented a quarter of the most frequently occurring terms, but were a small percentage of the total terms.

11. Multilingual Searching

Xu's (2000) analysis of 634 million web pages shows a 28 language corpus - most of the Web is English language with an increasing amount of web sites in Japanese, German, French, Italian and Chinese, Spanish, etc. Multilingual retrieval techniques will be at the forefront of IR research for the foreseeable future.

Discussion

This ongoing study of Web searching has examined a number of large-scale Excite transaction logs. We reported selected results from three studies of the Excite query corpora. These studies, using large scale log data, can answer some interesting questions about Web searching, but cannot address the results of users' queries or assess the performance of different search engines. The analysis does provide a snapshot for comparison of public Web

searching that can help improve Web search engines and services. We conclude from our analysis that most Web queries are short, without much modification, and are simple in structure. Few queries incorporate advance search techniques, and when they are used many mistakes result. However, relevance feedback and advanced search features are growing in use. People retrieve a large number of Web sites, but view few result pages and tend not to browse beyond the first or second results pages. Overall, a small number of terms are used with high frequency and many terms are used once. Web queries are very rich in subject diversity and some unique. The subject distribution of Web queries does not seem to map to the distribution of Web sites subject content. Web searching is a huge public challenge, but an imprecise skill.

Further Research

We have provided a selected overview of results from a large-scale and ongoing series of Web searching studies. For further results and details from the analyses are reported in the papers listed below:

- *Queries characteristics* - Wolfram (2000)
- *Multimedia searching* - Jansen, *et al*. (in press)
- *Linguistic aspects of queries* - Jansen, *et al.* (forthcoming)
- *Term co-occurrence* - Ross & Wolfram (in press)
- *Queries in elicitation form* - Spink, *et al.* (forthcoming)
- *Sex-related queries* - Spink & Ozmultu (forthcoming)
- *Business-related queries* - Spink & Guner (forthcoming)
- *Analysis of 1997 1 million queries* - Spink, *et al.* (2000)

Previous results are currently being compared with results from an analysis of 1.7 million Excite queries to isolate similarities and/or differences in Web searching from 1997 to 1999. We conclude that continued research into Web user behaviour is needed to impact the development of new types of user interfaces and software agents to aid users in better Web searching.

References

- Goodrum, A. and Spink, A. (in press) "Image searching on the Excite Web search engine." *Information Processing and Management.*
- Goodrum, A. and Spink, A. (1999) "Visual information seeking: a study of image queries on the World Wide Web." *Proceedings of the 62nd Annual Meeting of the American Society for Information Science, Washington, DC, October 1999* (pp. 665-674).
- Jansen, B. J., Goodrum, A. and Spink, A. (in press) "Searching for multimedia: analysis of audio, video and image Web queries." *World Wide Web Journal.*
- Jansen, B. J. and Spink, A. (in press) "The Excite research project: a study of searching characteristics by Web users." *Bulletin of the American Society for Information Science. Invited Paper.*
- Jansen, B. J., Spink, A., Bateman, J. and Saracevic, T. (1998a) "Real life information retrieval: a study of user queries on the Web." *ACM SIGIR Forum*, 32(1), 5-17.
- Jansen, B. J., Spink, A., Bateman, J. and Saracevic, T. (1998b) "Searchers, the subjects they search, and sufficiency: a study of a large sample of Excite searches." *Proceedings of WebNet 98 Conference, Orlando, FL, November 1999.*
- Jansen, B. J., Spink, A. and Pfaff, A. (forthcoming) *The language of Web queries.*
- Jansen, B. J., Spink, A. and 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**(2), 207-227.
- Jansen, B. J., Spink, A. and Saracevic, T. (1998) "Failure analysis in query construction: Data and analysis from a large sample of Web queries." *Proceedings of the Third ACM Conference on Digital Libraries, Pittsburgh, PA.* (pp. 289-290).
- Lawrence, S. and Giles, C.L. (1998) "Searching the World Wide Web." *Science*, **280**(5360), 98-100.
- Ross, N. C. M. and Wolfram, D. (in press) "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.*
- Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999) "Analysis of a very large Web search engine query log." *ACM SIGIR Forum*, **33,** 3.
- Spink, A., Bateman, J. and Jansen, B. J. (1999) "Searching the Web: survey of Excite users." *Internet*

Research: Electronic Networking Applications and Policy, **9**(2), 117-128.

- Spink, A. and Guner, O. (2000). "Business queries on the Web." *WebNet 2000 Poster*.
- Spink, A., Jansen, J. and Ozmultu, H.C. (in press) "Use of query reformulation and relevance feedback by Excite users." *Internet Research: Electronic Networking Applications and Policy*.
- Spink, A., Milchak, S. and Sollenberger, M. (forthcoming) Elicitation queries to the Excite search engine.
- Spink, A. and Ozmultu, H. C. (forthcoming) *Sexual queries on the Web*.
- Spink, A., Wolfram, D., Jansen, B. J. and Saracevic, T. (in press) "Searching the Web: The public and their queries." *Journal of the American Society for Information Science*.
- Wolfram, D. (2000) "A query-level examination of end-user searching behaviour on the Excite search engine." *CAIS 2000: Proceedings of the 28th Annual Conference of the Canadian Association for Information Science, June 2000*. Available at: http://www.slis.ualberta.ca/cais2000/wolfram.htm [6 September 2000]
- Xu, J. (2000) "Multilingual search on the World Wide Web." *Presentation to HICSS-33, January 4-7, 2000. Maui, Hawaii.*
- Xu, J. (1999) "Internet search engines: real world IR issues and challenges." *Presentation to CIKM 99, October 31-November 4, 1999. Kansa City, MI.*

Information Research, Volume 6 No. 1 October 2000

***Selected results from a large study of Web searching: the Excite study*, by Amanda Spink and Jack L. Xu Location: http://www.shef.ac.uk/~is/publications/infres/paper90.html**

Last updated: 6th September 2000
