# SEARCHERS, THE SUBJECTS THEY SEARCH, AND SUFFICIENCY: A STUDY OF A LARGE SAMPLE OF EXCITE SEARCHES

Major Bernard J. Jansen
Department of Electrical Engineering and Computer Science
United States Military Academy
West Point, New York 10996 USA
Tel.: (914)938-3233 Fax: 938-5956
Email: jjansen@acm.org

Amanda Spink & Judy Bateman
School of Library and Information Sciences
University of North Texas
P.O. Box 311068 Denton TX 75203
Tel.: (940) 565-2187 Fax: 565-3101
Email: spink@lis.admin.unt.edu, jbateman@unt.edu

Tefko Saracevic
School of Communication, Information and Library Studies
Rutgers Uiversity
4 Huntington Street
New Brunswick, NJ 08903 USA
Tel.: (732)932-8017 Fax: 932-6916
Email: tefko@scils.rutgers.edu

**Please Cite**: Jansen, B. J., Spink, A. & Saracevic, T. 1998. Searchers, the subjects they search, and sufficiency: A study of a large sample of Excite searches. 1998 World Conference on the WWW and Internet. Orlando, Florida.

See Other Publications

## ABSTRACT

In apparently the largest study of actual Web searches, we analyzed transaction logs of a set of 51,453 queries posed by 18,113 users of Excite, a major Internet search service. This study focuses on three areas. First, we examined the query construction, providing data on the number of terms in queries, the use of Boolean logic, use of term modifiers, and the modifications to queries by users. Second, we examine the search terms, including the most commonly used search terms and the rank/frequency distribution of search terms. Finally, in order to get an indication of the level of user satisfaction with the search engine performance and searching technique, we supplement the quantitative analysis with the results of a survey of Excite users. This indication of performance is one of the components of sufficiency, a metric that measures how well a search engine satisfies a user's information desire. Implications of the findings for Web interface design are discussed.

## INTRODUCTION

With the phenomenal increase in usage of the World Wide Web (the Web), there has been growing interest in the study of a variety of topics and issues related to use of the

Web. However, to date there has been no large-scale, quantitative or qualitative study of Web searching by users. *How do they search the Web? What do they search for on the Web? Are they satisfied with the results they receive?* These questions are addressed, as far as we can ascertain, for the first time on a large scale in this study.

In this paper we report selected results from a major and ongoing study of users' searching behavior on the Web. We examined a set of transaction logs of users' searches and responses to a survey done on Excite (http://www.excite.com), a major Web search engine. The objectives of the study reported here were to analyze (1) the size, logical structure, and modification of queries; (2) the distribution and nature of search terms, (3) to get and indication if users are satisfied with the current system. In other words, this is both a quantitative and a qualitative analysis. The study involved real users, as they interacted with Excite for searching, capturing their queries as stated – in other words this was a naturalistic case study, rather than a lab study, with all its strengths and weaknesses. This brings us to the significance of this study, which is :is:

- *"The success or failure of any interactive system and technology is contingent on the extend to which user issues, the human factors, are addressed right from the beginning to the very end, right from theory, conceptualization, and design process to development, evaluation, and to provision of services.* (Saracevic, 1997a).

Queries and searching are major human issues. The more insight we have on how these are actually done, the higher the probability that that insight may be translated into improved search engines and better IR. It can provide sharper criteria for the development and improvement of Web information retrieval (IR) systems and interfaces, as well as other IR systems. Finally, such an insight can benefit user education and training programs.

## CONCEPTUAL FRAMEWORK

The conceptual and theoretical frameworks and models for human-computer interaction in general, but particularly as pertaining to IR queries and searching are not yet well developed or universally accepted. Aside from algorithms, most of the work in these areas is empirical, with little or no theoretical orientation or basis. Thus, we cannot simply refer to a pre-established framework and let it go at that. Synopsis of three relevant frameworks follows. The first is derived from a model of human-computer interaction as adapted to IR, called the stratified model. In this model, the IR interaction is viewed as a dialogue between participants, *user* and *'computer' (system)* through an interface at the *surface level.* (see Saracevic, 1997b). The second, and related, framework is derived from viewing human-computer interaction, including hypertext navigation, as a conversation (Gray 1993). The approach is derived from ethnomethodology and social construction of knowledge. ). For a further discussion of how these two frameworks relate to Web searching see (Jansen, Spink, Bateman, and Saracevic 1998). The last theory is one that is the incipient stage of development that we shall refer to as sufficiency. The theory relates on a mathematical level to the early work of Cooper

(1968) and more recently to that of Dunlop (1997). Traditionally, one evaluates IR systems using the measurements of recall and precision. However, these measurements neglect the user, with all the user's experiences, skills, limitations, and expectations. We shall develop this theory of sufficiency in other works. For this study, we are interested in the user's perception of the search engine performance. In other words, w as the user satisfied with the information retrieved?

## RELATED STUDIES

In this paper, we concentrate on users' *queries, search terms, and subjective opinions* as key variables in IR interaction on the Web. While there are many papers that discuss some or other aspect aspects of Web searching, most of those are descriptive, prescriptive, commentary and the like. We could not find any studies of Web searching similar to this one, containing data on searches; , thus we have nothing to compare. However, there were several studies that included data on searching of existing, mostly commercial IR systems, and we culled data from those to provide for some comparison. Fenichel (1981) did a pioneering study in this area. Hsieh-yee (1993) compared the search term use and search tactics of novices searchers and expert searchers. Bates et. al. (1993) studied the search terminology used by humanities. Spink and Saracevic (1997) analyzed searches done by professional searchers in interaction with users. These studies are hard to compare. Still, each of them had data on the number of search terms used by searchers under study. A picture emerges showing that searches of these various populations contain a range of some 7 to 15 terms. As will be discussed below, this is a considerably higher range than the mean number of terms found in this study that concentrated on Web searches.

## BACKGROUND ON EXCITE AND DATA

Founded in 1994, Excite, Inc. is a major Internet media company which offers free Web searching and a variety of other online services. "Excite Search," according to the description in its site " the Internet's most comprehensive search tool, lets you search more than 50 million Web pages, 140,000 Web site listings, and thousands of Usenet postings." According to an independent study "during a 28-day period from Sept. 29, 1997 to Oct. 26, 1997, there were a total of 11,793,000 unique visitors to the Excite Network" (press release by Excite Inc., November 17, 1997). While this includes all the visits, in addition to searches, it is safe to assume that the overwhelming number of Excite visits are searches. This provides a picture of the huge size of the traffic on Excite. We provide only a brief description of Excite search capabilities. More details are available at their Web site. Those search features that pertain to our results are described here:

- 
    - Up to ten search terms are allowed in a query. The default option is OR. So, if a users enters a query of two or more terms, the Excite search engine ORs the terms together by default.

- A set of terms enclosed in quotation marks returns pages with the terms as a phrase in exact order. No space between quotation marks and terms.
- Boolean operators AND, AND NOT, OR can be used, but these operators must appear in ALL CAPS and with a space on each side in order to work. When using Boolean operators the ICE (concept-based search mechanism) is turned off. , Parentheses are also available for nested Boolean logic.

At SIGIR 97, Excite representatives offered to make available a set of their searches for analysis to anybody who asked. We took their offer and downloaded 51,453 queries (log transactions) from 18,113 users that were available. A lengthy and more detailed analysis of this data appears in (Jansen, Spink, Bateman, and Saracevic 1998). A more concise version of the analysis appears in this paper. The queries examined are a random subset of Excite searches on 10 March 1997. Each transaction record contained three fields:

| Time of Day | User Identification | Query Terms |
|---|---|---|

The first field was time of day measured in hours, minutes, and seconds from midnight of 9 March 1997. The next field was user identification assigned by the Excite server, and the third field was the actual query. With these three fields, we were able to locate a user's initial query and recreate the chronological series of queries by each user. This allowed us to examine both the particular actions of individual users and compare actions among all users. We examined this data in the specific areas of the query construction and query terms.

## SEARCH STATISTICS

The basic statistics related to search terms and queries are given in Table 1. Queries consist of one or more search terms, and possibly includes logical operators and requirements. A term is any unbroken string of characters (i.e., unbroken in a sense that there is no space between characters). The characters in terms included everything – alphabet, numbers, and symbols. Terms were words, abbreviations, numbers, symbols, URLs, and any combination thereof. Logical operators are also counted as search terms when standing alone. The data is raw and messy – users entered terms in all kinds of ways and combinations, majority correct, but also including many with abbreviations, misspellings, errors and the like. We took the data ‘as is,’ i.e. we did not ‘clean’ the data in any way – because these queries represent real searches by real users. We took great care in derivation of counts, but because of the messiness of data there still may be errors – we estimate not more than 1%.

We provide three statistics: (1) Non-unique terms: sum of all terms over all queries making also a distinction for capitalization i.e. case sensitive. (2) Unique terms with case sensitive: count of unique terms where *Topic, TOPIC,* and *topic* are counted as three terms. (3) Unique terms with case non-sensitive: the three capitalization forms of *topic* are counted as one term.

| No. of users | No. of queries | Non-unique terms | Mean of terms<br><br>And Range | Unique terms with case sensitive | Unique terms without case sensitive |
|---|---|---|---|---|---|
| 18,113 | 51,453 | 113,776 | 2.21 0-10 | 27,459 | 21,837 |

**TABLE 1. Numbers of users, queries, and terms**

*There were on the average 2.8 queries per user*, meaning that a number of users went on and refined in some way their query. *On the average, a query contained 2.21 terms.* As mentioned, we could not find any data on Web searches, thus, we can not compare this average to other Web searching. However, as we showed above, the mean number of search terms in searching of regular IR systems ranged from about 7 to 15. This is from about three to seven magnitudes higher than found in this study. E, and even this is on the high side, because we counted operators as well.

As mentioned, Excite can handle queries of 1 to 10 terms. Table 2 shows the ranking of all queries by number of terms. The column Terms is the number of terms in the query. Percent is the percentage of queries containing that number of terms relative to the total number of queries.

| Terms in query | Number of queries | Percent |
|---|---|---|
| >= 5 | 3,793 | 7.06 |
| 4 | 3,789 | 7.20 |
| 3 | 9,242 | 17.79 |
| 2 | 16,191 | 31.65 |
| 1 | 15,854 | 31.81 |
| 0 | 2,584 | 5.02 |

**TABLE 2: Number of Terms in Queries.**

Web queries are short. *One in about every three queries had one term only, two out of three had one or two terms, and four out of five had one, two or three terms.* Less than 8% of the queries were 5 terms or more. A note should be made on queries with zero terms (last row). When a user enters a command for relevance feedback *(More Like This)*, Excite counts that as zero terms. Thus the last row represents the potential largest number of queries that used relevance feedback, or a combination of those and queries where user made some mistake that triggered this result. *If we take that all of them are relevance feedback queries then only one in about twenty queries used that feature – a small use of relevance feedback capability*. Again, in comparison with professionally-assisted IR

searching from the same study by Spink & Saracevic (1997), w. With the same caveat, i where it was found that some 11% of search terms came from relevance feedback. T, the relevance feedback is used half as much on the Web. It is surprising that the users use very little this potentially highly useful and certainly highly vaunted feature.

## QUERY CONSTRUCTION

As mentioned, Excite can handle Boolean logical operators, AND, OR, and AND NOT. In addition, parentheses ( ) allow for use of nested logic. We examined how many queries explicitly utilized Boolean operators, including nesting, as presented in Table 3. Boolean operators must be upper case. Additionally, to receive the correct result NOT must be used with AND. The column Incorrect displays the number of queries containing a specific Boolean operator that was constructed incorrectly. The last column is the percentage of queries containing Boolean operators that were incorrectly constructed. In a sense the last two columns pertain to failure analysis.

| Operator | Number of queries | Percent of all queries | Incorrect | Percent incorrect |
|---|---|---|---|---|
| AND | 4,488 | 8.44 | 1262 | 28.11 |
| OR | 132 | 0.26 | 46 | 34.85 |
| AND NOT | 120 | 0.23 | 79 | 65.83 |
| ( ) | 273 | 0.53 | 88 | 32.23 |

**TABLE 3: Use of Boolean Operators in Queries.**

Boolean operators were used very sparingly. Only 5,013 queries, *or about one in every 10 queries contained a Boolean operator*, and in those AND was used by far the most. A minuscule percentage of queries used OR or AND NOT. Only 273 of the total number of 5,013 queries with operators used nested logic – i.e. *only one in about eighteen Boolean searches placed some of the terms with operators in parentheses.* Of those queries that used the Boolean operators, 1262 or about 28% were incorrect. *About one in every three queries that used Boolean operators or parentheses was not entered as required by Excite.* The very small use of Boolean operators and the very large percentage of mistakes when they are used shows that the Web searchers are not up to Boolean. Redesign seems to be in order.

## CONTENT AND DISTRIBUTION OF TERMS

The 51,474 queries contained 21,862 unique terms that were non-case sensitive. As mentioned, for counting purposes of these unique terms we normalized all terms to be in lower case. We started with creation of a mega-table that contained distribution of all unique terms ranked from highest to lowest frequency of appearance (lowest frequency being one), but this mega-table is way to large to present. The 74 terms that were used

100 or more times in all queries had a frequency of 20,698 appearances as search terms in all queries. They represent 0.34 % of all unique terms, yet they account for 18.2 % of all 113,776 search terms in all queries. If we delete the 11 common terms that do not carry any content by themselves (*and, of, the, in, for, +,on, to, or, &, a*) that altogether had 9,121 occurrences, we are left with 63 subject terms that have a frequency of 11,577 occurrences – that is 0.29% of unique subject terms account for 10.3% of all terms in all queries. Interestingly, the high appearance of '+' represents also a probable mistake – a space between the sign and a required term. On the other end of the distribution we have 9.790 terms that appeared only once. These terms with frequency of one amounted to 44.78% of all unique terms and 8.6% of all terms in all queries. *In other words, about one in every ten subject terms used in all queries comes from a list of 64 terms. Close to a half of unique terms appeared only once.*

In order to ascertain some broad subjects of searching, we classified the 64 top terms into a set of common themes. Admittedly, such a classification is arbitrary and each reader can use his/her own criteria. Still a rough picture emerges as shown in Table 11. The first Percent column refers to percent of the frequency in the category in relation to the total frequency of 11,577 for all 63 terms; the second Percent column refers to percent of the frequency in the category in relation to the total of 113,776 terms in all queries.

| **Category** | **Terms selected from 63 terms with frequency of 100 and higher** | **Frequency for category** | **Percent of freq. -63 terms** | **Percent of all terms** |
|---|---|---|---|---|
| Sexual | *sex, nude, gay, xxx, pussy, naked, adult, porn, anal, erotic, porno* | 2862 | 24.72 | 2.51 |
| Modifiers | *free, new, big, real, black, young, de, high, page* | 1902 | 16.42 | 1.67 |
| Place | *state, american, home, world, york, texas, florida, city* | 1144 | 9.88 | 1.01 |
| Economic | *employment, jobs, company, business, service, stock, estate, car* | 968 | 8.36 | 0.85 |
| Pictures | *pictures, pics, photos, video* | 906 | 7.82 | 0.80 |
| Social | *chat, stories, celebrities, games, john* | 804 | 6.94 | 0.71 |
| Education | *university, college, school, history* | 758 | 6.54 | 0.67 |
| Gender | *women, girls, men* | 648 | 5.59 | 0.60 |
| Sports | *ncaa, basketball, wrestling* | 477 | 4.12 | 0.42 |
| Computing | *software, computer, internet* | 437 | 3.77 | 0.38 |

| | | | | |
|---|---|---|---|---|
| News | *magazine, news, war* | 361 | 3.12 | 0.32 |
| Art | *music, art* | 310 | 2.68 | 0.72 |

**TABLE 4. Subject categories for terms appearing more than 100 times**

There is no way of going around it: a lot of terms (and thus queries) dealt with some or other sexual topic. *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, or if extended to all terms in all queries then we estimate that about one in forty terms is sexual.* Of course, if one classifies some more terms in the category *Sexual* the percent will be higher. We perused the rest of the terms and came to the conclusion than no more than some dozen of other terms will unmistakably fall in that category. If we added them all together the frequency of terms in *Sexual* will increase but not that much, and particularly not in relation to thousands of terms in other categories that are widely spread across all frequencies. In other words, *as to frequency of appearance of terms among the 63 highest frequency terms those in category Sexual have highest frequency of all categories, but still three out of every four terms of 63 highest frequency terms are not sexual; if extended to the frequency of use of all terms we estimate that 39 out 40 of all terms used are not sexual.* While category *Sexual* is certainly big, in comparison to all other categories in no way does it dominate searching. We cannot say that if we categorize the frequency of appearance of all the unique terms that category *Sexual* will even remain the highest category. Considering the shear huge size of remaining terms, it probably will not. Interest in other categories is high, categories that deal with places, economics, social activities, education, sports, computing, and arts. In other words Web searching does cover a gamut of human interests. It is very diverse.

## SUFFICIENCY

Unfortunately, we could not survey the actual users of our quantitative analysis to determine if their searches had retrieved relevant information. Instead, data was gathered through an interactive eighteen (18) question survey developed by the researchers in conjunction with the staff at EXCITE, Inc. The interactive survey was made available through EXCITE's Home Page for 5 days from Friday April 11 to Tuesday April 15,1997. Only those EXCITE users who accessed EXCITE's Home Page (http://www.EXCITE.com) directly could access the survey form (http://www.unt.edu/survey/excite.html). Only a portion of the survey information is presented here. For an in-depth reporting of results see (Spink, Bateman, and Jansen 1997). From the on-line survey, most users reported retrieving relevant information from EXCITE on their current topic (Table 12).

| Users' Retrieval of Relevant Information From EXCITE on Current Topic | | |
|---|---|---|
| Retrieval Status | Number | % |

| | | |
|---|---|---|
| Yes | 206 | 72% |
| No | 80 | 28% |
| Total | 286 | 100% |

**Table 5. Users' retrieval of relevant information from EXCITE on current topic.**

At first, we thought that this was great. The search techniques were sufficient for a high percentage of users. However, after careful analysis, the number that replied *no* was surprisingly high. For reasons outlined in (Spink, Bateman, and Jansen 1997), most of the respondents to the survey were repeat users of Excite, and therefore probably the most skilled in using the Excite searching techniques. So, even with the most experienced users, almost 30% were not satisfied with the current search techniques, search performance, or the data contained within the information base. Based on this figure and the problems that users exhibit with search techniques, it further supports the need for redesign.

## SUMMARY

The analysis involved 51,473 queries from 18,113 users, having all together 113,776 terms, of which 21,862 were unique terms disregarding capitalization. We are providing the highlights of our findings:

- The users did not have many queries per search. The mean number of queries per user was 2.8.
- Web queries are short. On the average, a query contained 2.35 terms. Less than 4% of the queries were more than 6 terms. Relevance feedback was not used that much.
- Boolean operators were not frequently used. About one in 15 queries contained a Boolean operator, and in those AND was used by far the most.
- The distribution of the frequency of use of terms in queries was highly skewed. A few terms were used repeatedly and a lot of terms were used only once. On the top of the list, the 64 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 a half of unique terms.
- There is a lot of searching about sex on the Web, but all together it 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, or if extended to all terms in all queries then about one in 40 terms is sexual. But while sexual terms are high as a category, they still represent a very small proportion of all terms. A great many other subjects are searched, and the diversity of subjects searched is very high.

- Most respondents of the survey reported retrieving relevant information; however, the number who reported not finding relevant information was surprisingly high.

## CONCLUSIONS

We investigated a large sample of searches on the Web, represented by logs of queries from a major Web search provider. We augmented this sample with an on-line survey. However, we consider this study just as a beginning. We already downloaded a sample of over a million Excite queries for analysis. In a way, we consider this study as a pilot for analysis of a much larger sample. Further, we are expanding the theoretical unpinning of sufficiency.

Web search users seem to differ significantly from users of traditional IR systems, such as those represented by users of DIALOG or assumed users of Text Retrieval Conference (TREC). It is still IR, but a very different IR. Web users are certainly not comfortable with Boolean operators and other advanced means of searching. They certainly do not frequently browse the results, beyond the first page or so. These facts in themselves emphasize the need to approach design of Web IR systems and search engines in a significantly different way than the design of IR systems as practiced to date. For instance:

- The low use if advanced searching techniques would seem to support the continued research into new types of user interfaces, intelligent user interfaces, or the use of software agents to aid users in a much simplified and transparent manner.
- The impact of the large number of unique terms on key term lists, thesauri, association methods, and latent semantic indexing deserves further investigation – the present methods are not attuned to the richness in the spread of terms.
- The area of browsing and relevance feedback also desires further investigation, among others the question of actual low use of these features should be addressed.
- In itself, the work on investigation and classification of a large number of highly diverse queries presents a theoretical and methodological challenge. The impact of producing a more refined classification may be reflected in making browsing easier for users and precision possibly higher – both highly desirable features.

To end with a general question. Certainly, the Web is a marvelous new technology. People have always been unpredictable in how they will use any new technology. It seems that this is the case with the Web. In the end, it all ends with the users and the use people make of the Web. Maybe they are searching the Web in ways that designers and IR researchers have not contemplated or assumed, as yet. Aren't they?

## REFERENCE

Bates, M.J., Wilde, D. N. and Siegfried, S. (1993) An analysis of search terminology used by humanities scholars: The Getty online searching project report. *Library Quarterly, 63* (1), 1-39.

Cooper, W.S. Expected Search Length: a Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems. *American Documentation*, vol. 19, January 1968.

Crovella, M. E. & Bestavros, A. (1996). Self-similarity in World Wide Web traffic evidence and possible causes. *Proceedings of. ACM SIGMETRICS*, 126-137.

Dunlop, M.D. (1997) Time, Relevance and interaction Modeling for Information Retrieval. Proceedings of the 20th Annual International ACM SIGIR Conference on Information Retrieval. 206-213.

Excite Inc. (1997). *Excite Network sustains unprecedented momentum in traffic growth. Relevant Knowledge October results place Excite Network firmly within top five.* Redwood City, CA. Press release November 17, 1997

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*, 23-32.

Gray, S.H. (1993). *Hypertext and the technology of conversation. Orderly situational choice.* Westport, CN: Greenwood.

Hsieh-yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science, 44*(3), 161-174

Lynch, C. (1997). Searching the Internet. *Scientific American, 276*. .50-56.

Saracevic, T. (1997a). Users lost: Reflections on the past, future, and limits of information science. *SIGIR Forum, 31* (2) 16-27.

Saracevic, T. (1997b). The stratified model of information retrieval interaction: Extension and applications. *Proceedings of the American Society for Information Science, 34*, 313-327.

Spink, A., Bateman, J., & Jansen, B.J. (1998). *Searching heterogeneous collections on the Web: Behavior of Excite users*. Unpublished paper. School of Library and Information Sciences, University of North Texas.

Spink, A. & Saracevic, T. (1997). Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching. *Journal of the American Society for Information Science, 48*, (8), 741-761.

Zorn, Peggy, et. al. *Advanced Searching: Tricks of the Trade*. Online, May 1996.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the assistance of Graham Spencer, Doug, Cutting, Amy Smith and Catherine Yip of Excite, Inc in providing the data and information for this research. Without the generous sharing of data by Excite Inc. this research would not be possible.