



COVER SHEET

Spink, Amanda and Wolfram, Dietmar and Jansen, B.J. and Saracevik, Tefko (2001) Searching the web: the public and their queries. *Journal of the American Society for Information Science* 53(2):pp. 226-234.

Accessed from <http://eprints.qut.edu.au>

Copyright 2001 John Wiley & Sons, Inc.

SEARCHING THE WEB: THE PUBLIC AND THEIR QUERIES

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
(spink@ist.psu.edu)

Dietmar Wolfram

School of Library and Information Science
University of Wisconsin-Milwaukee
P.O. Box 413
Milwaukee, WI 53201
(dwolfram@csd.uwm.edu)

Major B. J. Jansen

Lecturer, Computer Science Program at the University of Maryland (Asian Division)
Seoul, 140-022 Korea
(jjansen@acm.org)

Tefko Saracevic

School of Communication, Information and Library Studies
Rutgers University, 4 Huntington St.
New Brunswick, NJ 08903
(tefko@scils.rutgers.edu)

ABSTRACT

In studying actual Web searching by the public at large, we analyzed over one million Web queries by users of the Excite search engine. We found that most people use few search terms, few modified queries, view few Web pages, and rarely use advanced search features. A small number of search terms are used with high frequency, and a great many terms are unique; the language of Web queries is distinctive. Queries about recreation and entertainment rank highest. Findings are compared to data from two other large studies of Web queries. This study provides an insight into the public practices and choices in Web searching.

Introduction

The Web is now a major source of information for many people worldwide. Millions of Web queries are posed daily. People can search the Web via many different search engines that use various search algorithms and techniques. The Web has attracted not only a high amount of use, but numerous studies as well. Statistics on Web use appear regularly (OCLC, [1999]). We know that single Web search engines cover less than 20% of Web sites, and cannot keep pace with Web growth (Lawrence & Giles, [1999]). Some Web search engines are more effective than others (Gordon & Pathak, [1999]). Strong regular patterns of users' Web surfing have been found (Huberman, Pirolli, Pitkow, & Lukose, [1998]). Web access spans a much broader population than non-Web-based information retrieval (IR) systems (Spink, Bateman, & Jansen, [1999]). Despite these and many other studies, as yet, we have relatively little understanding about how people actually search the Web. We understand more about how people use non-Web-based IR systems (Spink & Saracevic, [1997]) than how they use the Web.

We report findings from a large study of searching behavior by users of the Excite search engine (<http://www.excite.com>). Excite@Home Corp. is a major Internet media public company offering free Web searching and a variety of other services. The analysis covers over one million queries by over 200,000 users. Users were anonymous. We provide detailed statistics on Web searching and an analysis of query language and topics. We conclude that Web searching by the public differs significantly from searching of IR systems (such as DIALOG, Lexis-Nexis, and others) by their users.

This is a naturalistic study, involving real users in the act of searching for information on the Web. As the Web is evolving into a primary source of information for a global society, our findings, together with the findings from other similar studies, provide a clearer understanding of Web use, particularly by the broader public. In turn, this has implications for developing better design of Web interfaces and search engines.

Related Studies

This study follows in the footsteps of preceding and similar studies by the same research team, on smaller samples of data (Jansen, Spink, Bateman, & Saracevic, [1998]; Jansen, Spink, & Saracevic, [2000]). Our previous study used a sample of 51,473 queries collected on 9 March 1997. We label it the “51K study.” It is also complemented by a similar study of a large sample of public queries of the Alta Vista search engine (Silverstein, Henzinger, Marais, & Moricz, [1999]). That study involved 153,645,050 queries collected from 2 August to 13 September 1998. We label it the “Alta Vista study.” Our study reported here involves 1,025,910 queries collected on 16 September 1997. We label it the “1M study.” We could not find any other studies of similar magnitudes supported by data, even though anecdotal observations about Web queries are given in presentations and panel discussions at various conferences, but never substantiated.

A note of caution is in order. As already noted in the *Alta Vista study*, comparisons of results from various studies cannot be easily or fully achieved. Namely, while the same questions are asked, data definition and analysis differ to some extent from study to study. The metrics are not standardized; they are not necessarily the same. The basic problem starts with defining what is a “term” in a Web query. The public enters queries and the raw data are very messy. A term can be anything from words to Uniform Resource Locators (URLs) to any set of characters and symbols; a query can even be empty - no terms, and as in the *Alta Vista study* a term can also be a field-value designator. What is included and excluded as being a “term” effect the counts. A similar problem is in defining a “unique query” as we call it, or a “distinct query,” as *Alta Vista study* calls it. Thus, our comparisons should be taken more as a comparison of similarity in trends than in actual numbers. This points for a need to further develop and standardize metrics for study of Web use.

The data in the three studies were collected at different time periods; however, the difference is only a little more than a year. We compare our findings reported here with the findings from the other two studies, providing a sort of a longitudinal view of the behavior of the public in web searching in a relatively short time period, and a comparison of query characteristics from different search engines and samples. We are in the process of undertaking a study of new samples of Excite queries posed over 2 years later and consisting of 2.5 million queries.

Measured in Internet years, data used in all these studies are old, if not ancient. But not obsolete. The Internet changes fast. In contrast, people, their information needs, and behavior do not. The amount of Web use and searching is growing explosively. This does not necessarily mean that the type of use is also changing in similar ways. Longitudinal studies of Web searching can show whether people change their use of the Web, providing an insight on whether public queries are evolving and changing together with the Web.

Excite Searches

Excite searches are based on the exact terms a user enters in a query. Capitalization is disregarded in searching, with the exception of the logical operators AND, OR, and AND NOT. Stemming is not available. An on-line thesaurus and concept linking method called Intelligent Concept Extraction is used to find related terms for the terms entered. Search response is in result pages, listing URLs, and a short description of sites that match the query, ranked by a probability of relevance to the query. Various advanced search features are available (note that they may change over time). A + (plus) or - (minus) in front of a term indicates that the term must or must not appear in the result; quote marks around two or more terms indicate a search for a phrase. Relevance feedback is available to find similar sites; it is indicated with "More like this" provided with a retrieved URL. Alta Vista and other search engines have most of the same features, but they also differ in some details.

The data we analyzed consisted of a log of transaction record of 1,025,910 user queries submitted during a portion of a single day. The data set contained three fields: *Time of Day*: measured in hours, minutes, and seconds from midnight of 16 September 1997; *User Identification*: an anonymous user code assigned by the Excite server; and *Query Terms*: exactly as entered by the given user. With these three fields, we located a user's initial query and recreated the chronological series of actions by each user in a session. We analyzed the following:

1. *Term*: any unbroken string of alphanumeric characters entered by a user. Terms included words, abbreviations, numbers, and logical operators (AND, OR, NOT). URLs and e-mail addresses were treated as single terms.

2. *Query*: a set of one or more search terms; it may include advanced search features, such as logical operators and modifiers. (a) *Unique queries* are all *differing* queries entered by one user in one session; the differing queries could be modifications of the previous query or entirely new queries. (b) *Repeat queries* are all multiple occurrences of the *same* query that represent request for multipage viewing (when a user request to view a subsequent page Excite generates the same query). (c) *Zero term queries* are queries without any terms; they are generated by Excite when a user executes *More Like This* (these are considered as relevance feedback requests), or when a user enters no terms or symbols only.

3. *Session*: the entire set of queries by the same user over time. A session could be as short as one query or contain many unique and repeat queries.

4. *Result pages*: display of results for viewing. Excite presents in a single page a set of 10 Web sites ranked by estimated relevance probability. The user can choose to view only the first page or may request one at a time, the remaining pages.

Queries and Sessions

The 211,063 users posed a total of 1,025,910 queries, of which 51.8% were unique queries, 38.5% were repeat queries, and 9.7% were zero queries (Table 1).

Table 1. Summary data.

| | |
|--|-----------|
| Number of users | 211,063 |
| Number of queries (including repeat queries) | 1,025,910 |
| Number of unique queries | 531,416 |
| Number of repeat queries | 395,461 |
| Number of zero term queries | 99,033 |

| | |
|--|-----------|
| Mean number of queries per user session | 4.86 |
| Median number of queries per user session | 8 |
| Mean number of unique queries per user session | 2.52 |
| Median number of unique queries per user session | 4 |
| Total number of terms (including terms in repeat queries) | 2,216,986 |
| Total number of terms (tokens) (excluding terms in repeat queries) | 1,277,763 |
| Number of unique terms (types) | 140,279 |
| Mean number of terms per query (including repeat queries) | 2.16 |
| Median number of terms per query (including repeat queries) | 2 |
| Mean number of terms per query (excluding repeat queries) | 2.4 |
| Median number of terms per query (excluding repeat queries) | 2 |

The mean number for total queries in a session was 4.86, with a median of 8. For unique queries, the mean was 2.52, with a median of 4. In the *51K study* the mean number of queries per session was 2.8 and in the *Alta Vista study* was 2.02. To generalize: it seems that the mean number of queries per session is between 2 and 3. But the averages in this, as in all other data under study do not tell the whole story; the results are highly skewed. That is why we opted to use distributions as the basic method of analysis.

Queries per User

Some 48.4% of users submitted a single query, 20.8% two queries, and about 31% of users entered three or more unique queries (Fig. 1).

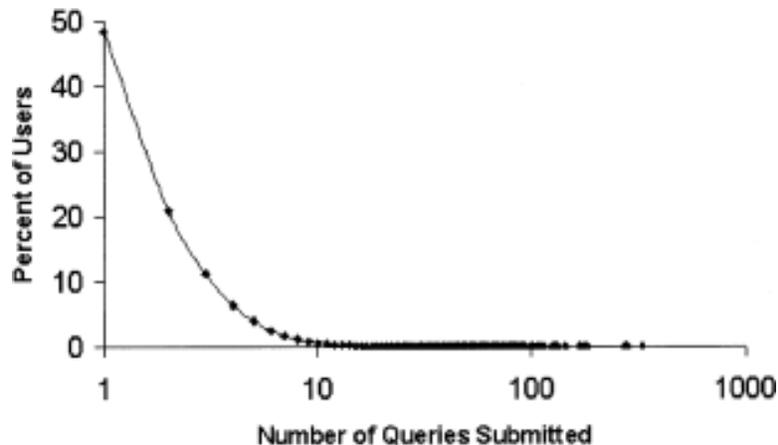


Figure 1. Number of unique (nonzero) queries submitted by each user; 4,031 users (1.9%) submitted a single zero term query and are not included.

About 1.9% of users entered nothing but a zero term query. However, the distribution is very skewed toward the lower end of the number of queries submitted, with a long tail of very few users submitting a large number of unique queries. In general, users did not enter many queries in a session, and close to half entered only one query.

In the *51K study* the percentages of users with one, two, and three queries were, respectively, 67, 19, and 7%; in the *51K study* a larger percent of users entered one query only than in the *IM study*. This statistic was not reported in the *Alta Vista study*. However, the *Alta Vista study* reports on queries per session (as they computed it, it is similar but not the same as queries per user): 77.6% of sessions had one query, 13.5% had two, and 4.4% had three. The general pattern is repeated in all studies: as to distribution, most users had one query only.

Modification of Queries

Because some 52% of users entered more than one unique query, the question arises: how were subsequent queries modified? We counted the change in the number of modified terms from a preceding to a subsequent query (Fig. 2).

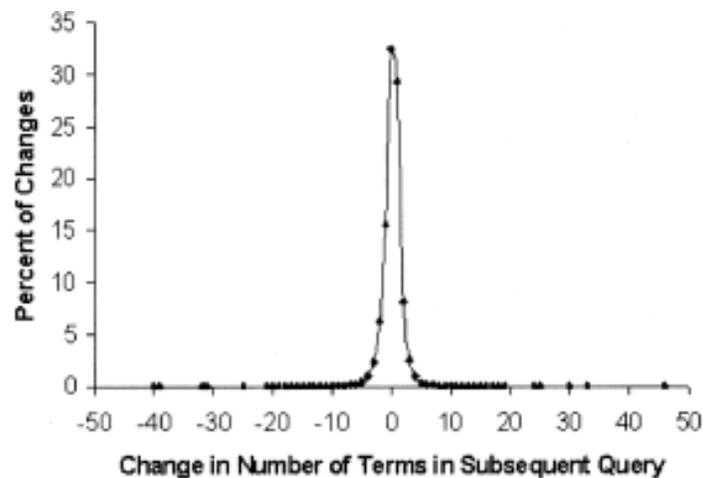


Figure 2. Changes in the number of terms in subsequent unique queries by users who submitted multiple queries.

Zero change means that the user modified one or more terms in a subsequent query, but the total number of terms in both was the same. An increase or decrease of one term means that one term was added to or subtracted from the preceding query. Percentages in this section are based on the number of queries in relation to all modified queries excluding zero term queries.

In 32.5% of modified queries, there was a modification in one or more terms, but there was no change in the number of terms in the query. That is, about one in every three modified queries had the same number of terms as the preceding query. In the remaining subsequent queries, where terms were either added or subtracted, 41.6% added terms, and 25.9% subtracted terms. Web users are more likely to add than delete a term. Users typically do not add or delete many terms in their subsequent queries. Some 99.2% of subsequent queries represented additions or subtractions of five terms or less. Modifications to queries are done in small increments over a

few queries. About 29.3% of modified queries have one more term than the preceding query, and about 15.5% have one less term. Assuming that addition of terms signifies narrowing of a query for higher precision, then Web users tend to go more often from broad to narrow formulations in queries, because the most common query modification is to add terms.

In the *51K study* 33% of queries were modified (the nature of modification was analyzed in a separate paper by Spink, Jansen, & Ozmultu, [2000]). In 34.76% of modified queries terms were changed, but the number of terms remained the same; in 19.03% of modified queries a term was added and in 16.33% a term was subtracted; in 9% two terms were added and in 8.33% two terms were subtracted. In the *Alta Vista study* the statistics were calculated somewhat differently. No statistic was given for queries that were modified but had the same number of terms; 7.1% of queries had added terms, and 3.1% had deleted terms; more specifically, 5.4% of queries had one term added and 2.1% had one term deleted; 1.4% of queries had modified operators. The *IM study* shows a significantly higher percent of modified queries than the other two, indicating a possible difference of user behavior in respective studies, or more likely, a difference in ways of counting from the log transaction. In general, a high percent of users do not modify queries to any extent, and when they do modify, they change some terms, but the total terms remain the same. Assuming that modifications are done by more sophisticated users, a concentrated study of these modifications can shed further light on the behavior of the more search-savvy part of the public.

Result Pages Viewed

Figure 3 shows the distribution of result pages examined per user.

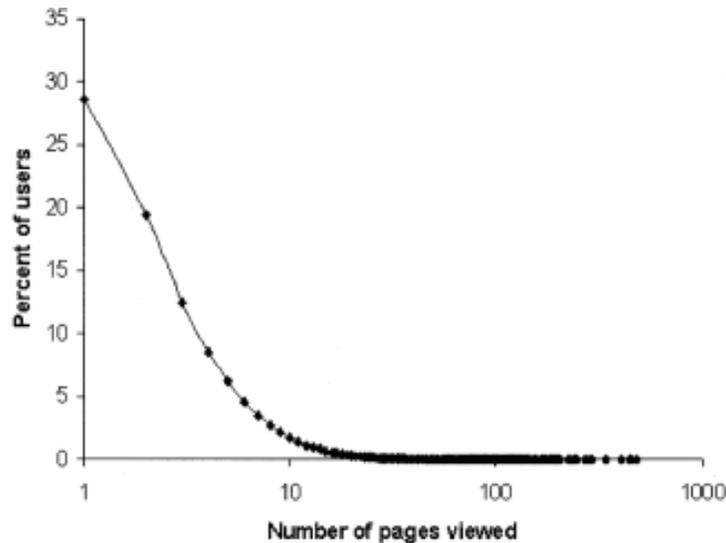


Figure 3. Number of pages viewed per user.

The median was eight pages viewed per user. However, 28.6% of users examined only one page of results, i.e., because a page contains 10 ranked Web sites, about one in every four users looked at 10 or less sites. Another 19% looked at two pages only. That is, close to half of the users looked at two or less pages. Were users so satisfied with the results that they did not need to view more pages? Were a few answers good enough? Is the precision of Web search engines that high? Are the users after precision? (Precision is calculated as the number of relevant Web sites retrieved over the total number of Web sites retrieved; relative precision can be calculated in relation to first X sites retrieved). What proportion was relevant in relation to the X sites? Or did they just give up? Using only transaction log analysis, we cannot determine the answers to these questions. However, this percentage, combined with the small number of queries per session, may illustrate a need for high precision in Web IR algorithms.

In the *51K study* the percent of users looking at one, two or three pages were, respectively, 58, 19, and 9%; in the *Alta Vista study*, the respective percentages were 85.2, 7.5, and 3.0%. Again, while the percentages do not coincide with this study, they show the same effect that a large

percent of users do not go beyond the first page. This is quite remarkable in light of the generally large number of retrievals. The public has a low tolerance of going in depth through what is retrieved.

Use of Advanced Search Features

Less than 5% of all queries used any Boolean operators (Table 2).

Table 2. Use of advanced search features in queries (number of queries = 1,025,910).

| Feature | Number of queries | Percent of queries |
|---------------------|--------------------------|---------------------------|
| AND/and/And | 29,146 | 3% |
| OR/or/Or | 1,149 | 1% |
| NOT/AND NOT | 307 | 0.0003% |
| + plus (correct) | 17,028 | 2% |
| + plus (incorrect) | 27,292 | 3% |
| + plus total | 44,320 | 5% |
| - minus (correct) | 1,656 | 0.001% |
| - minus (incorrect) | 20,295 | 2% |
| - minus total | 21,951 | 2% |
| " " (quotations) | 52,354 | 5% |
| '.' (periods) | 51,804 | 5% |
| ':' (colons) | 1,459 | 1% |
| & | 3,342 | 3% |
| Relevance feedback | 99,033 | 9.7% |

Of these, AND was used most. A smaller percentage of queries used OR and a minuscule percentage AND NOT. The + (plus) and - (minus) modifiers (requiring that a term must be present or absent in the answer) were used slightly more than Boolean operators. Together, + and - were used in 7% of all queries. The ability to create phrases (terms enclosed by quotation marks) was used in only 5% of all queries.

Similar results were found in the *51K study*, where less than 10% queries had a Boolean operator, 9% had modifiers + or -, and 6% had phrases. In the *Alta Vista study* 20.4% of queries had any kind of operator or modifier; 9.7% had one operator in a query, 6% had two, 2.6% had three, and 2.1% more than three. A few users account for these more sophisticated queries. For an overwhelming number of Web users, the advanced search features do not exist. The low use of advanced search features raises questions of their usability, functionality, and even desirability, as currently presented in search engines.

However, many users that did use Boolean operators made mistakes. The most common mistake was not capitalizing the Boolean operator, as required by the Excite search engine. In this analysis, the Boolean operator AND presented a special problem because of various forms, so we did a further analysis. Some form of AND (as “AND”, “And”, and “and”) was used in 29,146 Queries; some queries had more than one AND. If considered as Boolean operators, “And”, and “and” were mistakes. Most of them were, but not all. In a number of queries “and” was used as a conjunction, for example, as in the query “College and university harassment policy”. We could not distinguish the intended use of “and” as a conjunction from that as a mistake for Boolean operator, thus our count of AND mistakes are on the high end. But the users may not be able to distinguish this either.

There was a similarly high percentage of mistakes in the use of plus + and minus - operators. The queries were checked for conformity with the Excite searching rules concerned the use of + and -. The queries that did not conform to the rules were counted as mistakes. It seems that when users are using an advanced search feature, it is as likely that they will use it correctly (as required in system instruction) as incorrectly.

Many queries incorporated searching techniques that Excite does not support. These failures can be classified as a carryover from experiences with other Web search engines, on-line public

access catalogs, and IR systems. For example, there were 914 occurrences of the operator SEARCH and 1,459 uses of the symbol “:” (colon) as a separator for terms. The symbol “.” (period) was used 51,804 times, either as a separator or as a part of URL and email addresses. The symbol “&” was used in lieu of the Boolean AND some 3,342 times. However, similar to the use of And, we cannot tell what the searcher meant. These symbols are common in many other search engines.

The usage of Boolean operators in this study was significantly lower than those reported for Web-based digital library users (Jones, Cunningham, & McNab, [1998]) and significantly lower than studies of searches by professional searchers in IR systems (Spink & Saracevic, [1997]). This may reflect a highly simplified type of searching by the broad public, in comparison of more complex searching by more sophisticated users and professionals that use these other systems.

Use of Relevance Feedback

As mentioned, when a user clicks on a link *More Like This* at a bottom of a retrieved site, the Excite transaction log counts that as a query, but a query with zero terms. Clicking on *More Like This* is, in fact, entering a command for relevance feedback requesting a set of similar sites. Assuming that all 99,033 queries with zero terms were for relevance feedback (i.e., including possible user mistakes when entering a query with no terms), only at most 9.7% of all queries used that feature. This is a small use of the relevance feedback capability. This indicates that users either did not find many relevant sites, did not care to pursue further searching for similar sites, or are unfamiliar with the capabilities of this feature. Alternatively, it could indicate that they were simply satisfied with results.

In the *51K study*, 5% of users used relevance feedback. The *Alta Vista study* did not contain data on this aspect. In the study of IR searching by professionals, it was found that some 11% of search terms come from relevance feedback (Spink & Saracevic, [1997]). Although this is a different type of feedback in terms of results, it is still an action involving relevance feedback. In IR, the use of feedback is double that in Web searching, but both uses seem relatively small. Relevance feedback, although intuitively highly desirable, in practice is simply not used much at all.

Search Terms and Topics

Terms per Query

The mean number of terms in unique queries was 2.4. Figure 4 shows the frequency for unique queries by number of terms.

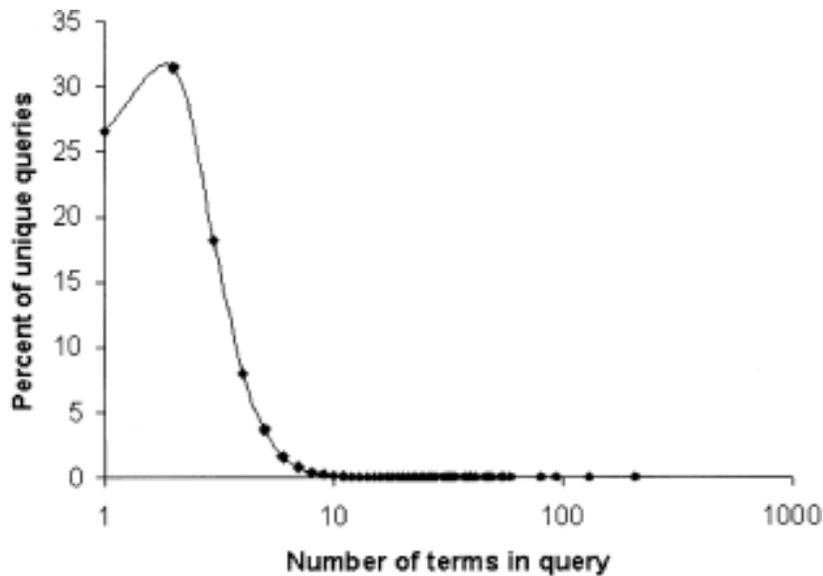


Figure 4. Number of terms appearing in each unique query. The figure does not include queries containing zero terms, which represent 9.7% of all queries.

The logarithmic scaling does not allow us to include the 9.7% of all queries submitted with zero terms. Web queries are generally short. Some 26.6% of queries had one term only, 31.5% had two terms, and 18.2% had 3 terms. Thus, close to 60% of all queries had one or two terms, with most of them having the “magical” search length of two terms. Less than 1.8% of the queries had more than seven terms.

In the *51K* and the *Alta Vista study* the respective mean number of terms per query was 2.32 and 2.35. In the *51K* (and *Alta Vista*) study 31% (25.8%) of queries had one term only, 31% (26.0%) two, and 18% (15%) three terms; comparing these results with a study of usage of a digital

library (Jones et al., [1998]), we find a similar query length. However, these results deviate significantly from results of IR searching studies that show the mean number of search terms when searching IR systems ranges from about 7 to 15 (data from four studies as reviewed in Jansen, Spink, Bateman, & Saracevic, [1998]). This is about three to seven magnitudes higher than found in the three studies of Web searching by the public, as reported here.

Distribution of Terms

Figure 5 shows a graph of the size-frequency distribution of all terms used in unique queries.

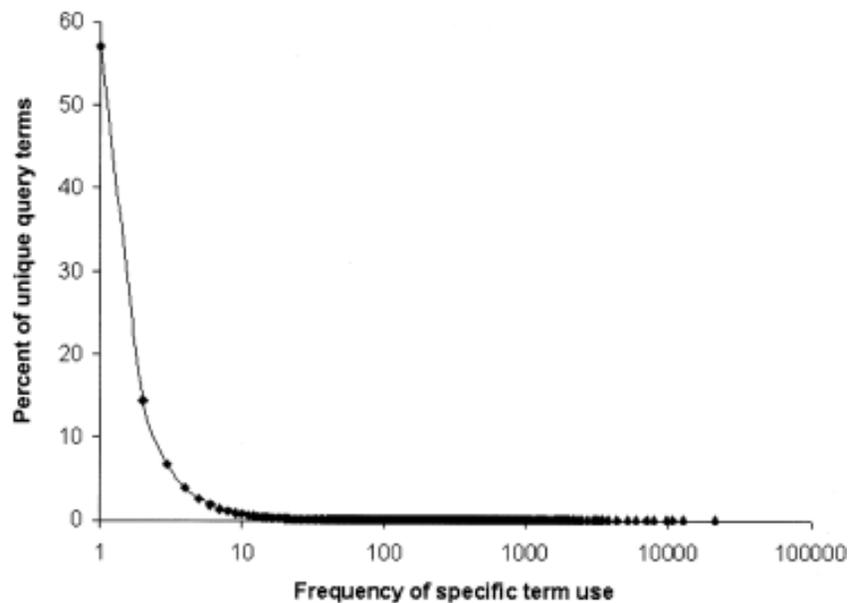


Figure 5. Term distribution within unique queries - ordered by frequency of terms in relation to percent of unique terms.

Of the 140,279 unique terms, some 57.1% were used only once, 14.5% twice, and 6.7% three times, i.e., some 78.3% of unique terms were used three times or less. The Web query language is highly varied. An unusually large number of unique terms is used with a low frequency; contributing to this are, among others, a high number of spelling errors, terms in languages other than English, and Web specific terms, such as URLs. On the other end, an unusually small number of unique terms are used with a very high frequency.

A double log rank-frequency plot, often used to determine the accordance with a Zipf distribution, appears in Figure 6.

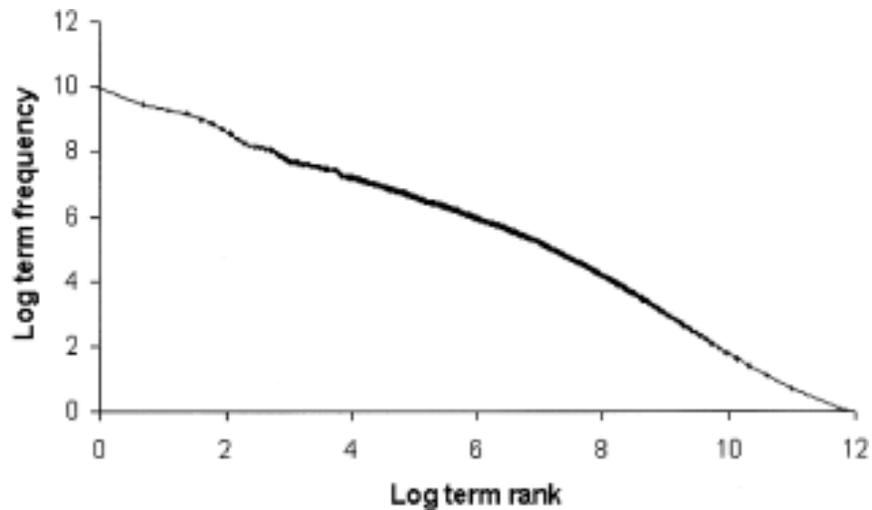


Figure 6. Rank-frequency distribution of terms used within unique queries, ordered by rank using a log-log transformation.

Traditionally, a Zipf distribution has been applied to extensive textual passages, but has also been investigated for database contents in bibliographic and full text databases (Zipf, [1949]).

Research has shown that a traditional Zipf model does not adequately fit term distributions, but are better represented with more sophisticated models (Nelson, [1989]; Wolfram, [1992]). A double log rank-frequency plot, often used to determine the accordance with a Zipf distribution, appears in Figure 6. To correspond to a Zipf distribution, the double log plot should be close to a straight line. The resulting distribution is slightly unbalanced for the high and low ranking terms, indicating that, just as with database term distributions, a query term distribution may require a more sophisticated model to describe the relationship between the selection of terms and their frequency of appearance within queries.

The public language of Web queries has its own and unique characteristics. The public “talks” in Web searches in its own way. This warrants further study of both ends of the rank-frequency distribution, and of other linguistic characteristic of Web queries so that user query language can be anticipated and supported.

High-Frequency Terms

Table 3 lists the top 75 terms, occurring more than 1110 times in unique queries.

Table 3. Listing of 75 most frequently occurring terms within the 531,416 unique queries (excite treats everything as lower case).

| Term | Frequency | Term | Frequency | Term | Frequency |
|-------------|------------------|-------------|------------------|-------------|------------------|
| and | 21385 | naked | 1968 | web | 1366 |
| of | 12731 | american | 1961 | history | 1359 |
| sex | 10757 | stories | 1958 | video | 1356 |
| free | 9710 | software | 1908 | sports | 1351 |
| the | 8013 | games | 1904 | california | 1345 |
| nude | 7047 | diana | 1885 | men | 1327 |
| pictures | 5939 | p***** | 1876 | national | 1306 |
| in | 5196 | black | 1823 | big | 1290 |
| university | 4383 | on | 1813 | york | 1277 |
| pics | 3815 | photos | 1799 | texas | 1276 |
| chat | 3515 | jobs | 1735 | porno | 1263 |
| for | 3431 | world | 1734 | maps | 1256 |
| adult | 3385 | a | 1711 | employment | 1234 |
| women | 3211 | magazine | 1690 | city | 1222 |
| new | 3109 | nudes | 1690 | canada | 1204 |
| xxx | 3010 | news | 1687 | playboy | 1197 |
| girls | 2732 | football | 1627 | car | 1195 |
| music | 2490 | page | 1591 | erotic | 1189 |
| porn | 2400 | computer | 1533 | weather | 1184 |
| to | 2265 | princess | 1461 | map | 1159 |

| | | | | | |
|---------|------|-----------|------|---------------|------|
| gay | 2187 | airlines | 1409 | internet | 1156 |
| school | 2176 | download | 1381 | international | 1113 |
| home | 2150 | real | 1381 | high | 1113 |
| college | 2043 | education | 1376 | star | 1110 |
| state | 2010 | art | 1374 | asian | 1110 |

P***** = expletive.

The top 75 terms in frequency represent only 0.05% of all unique terms, yet they account for 9% of all 1,277,763 search terms in all unique queries. We then deleted the eight common terms without content by themselves (and, of, the, in, for, +, on, to, or, &, a) in 56,545 occurrences. We were left with 67 subject terms or 0.04% of unique terms (types) that account for 11.5% of all terms used in all queries (tokens). The subjects represented by the top terms are interesting by themselves. For instance, there are a number of terms that represent sexuality. Also the high rank of term “Diana” reflects the interest of the time related to Princess Diana death. However, from this list of terms, we cannot derive the variety of topics of Web queries, beyond inference from terms used. Thus, we undertook a different analysis, as reported in the next two sections.

The following were the 25 highest-ranking subject terms in the *51K study*: *sex, nude, free, pictures, new, university, women, chat, gay, girls, xxx, music, software, pics, ncaa, home, stories, p***** (expletive), college, naked, adult, state, big, basketball, men.*

The *Alta Vista study* reports on “the 25 most popular queries,” with a different method for identification, where query frequency rather than term frequency was analyzed, but the results are comparable to term frequencies: *sex, applet, porno, mp3, chat, warez, yahoo, playboy, xxx, hotmail, [non-ASCII query], pamela anderson, p***** (expletive), sexo, porn, nude, lolita, games, spice girls, bestiality, animal sex, SEX, gay, titanic, bestiality.*

In the *51K study* the most frequent 64 subject terms represented 0.29 of unique terms, yet they account for 18.2% of all terms in all queries. This is similar to what we found in the *IM study*. But *Alta Vista* distribution differs: “the 25 most common queries asked form fully 1.5% of the

total number of queries asked . . . despite being only a 0.00000016% of the unique queries. The term “applet” was of high frequency; “examination of logs shows that almost all queries containing the term were submitted by a robot” - showing another unsuspecting aspect for further analysis. It is not explained how to distinguish robot queries; we could not find any method for doing this.

Clearly, all studies show a high degree of usage of most frequent terms, way out of their proportion to total number of terms. Some of these high-frequency terms reflect interest in current events, others the perennial human preoccupation with matters of sex; still others hint at a number of other topics. But, in another way this also indicates that there are great many terms total, especially in the long tail of infrequently used terms. The Web query vocabulary contains a very large number of different terms - much more than found in large English texts in general. There are few comprehensive studies of what terms people use, the distribution of those terms, and the modification of those terms during Web searching. The potential benefit of such studies to IR system developers, users, and Web site classifiers and designers could be high.

Co-occurrence of Terms

What types of information were people seeking on the Web? What were the query topics? A simple interpretation of the most frequent terms as listed above provides some answer by inference, but that is not at all indicative of the range of topics searched. For instance, the list shows a high usage of sexual terms, but also of contemporary interest terms, and terms that indicate other topics. To seek an answer to these questions we undertook two further analyses. The first one is quantitative, concentrating on study of co-occurrences of terms. The second one is qualitative, using a classification approach.

An in-depth analysis of term pairs in the *IM study* is reported in Ross and Wolfram ([2000]). The analysis covers term pairs in unique queries only. Taken from that analysis are the 50 most frequently occurring term pairs, as presented in Table 4.

**Table 4. Fifty most frequently occurring term pairs in unique queries
(shown as: term1-term2 term pair frequency)**

| | | | | | | | | | |
|-------------------|-------|-------------------|-----|--------------------|-----|----------------------|-----|--------------------|-----|
| and-and | 6,116 | of-and | 690 | or-or | 501 | women- nude | 382 | sex-pics | 295 |
| of-the | 1,901 | pictures- of | 637 | sex- pictures | 496 | pics-nude | 380 | north- carolina | 295 |
| pics-free | 1,098 | how-to | 627 | nude- pictures | 486 | of- department | 365 | free-teen | 293 |
| university- of | 1,018 | and-the | 614 | for-sale | 467 | united- states | 361 | free- porn | 290 |
| new-york | 903 | free- pictures | 637 | and-not | 456 | of-history | 332 | and- nude | 289 |
| sex-free | 886 | high- school | 571 | and-sex | 449 | adult-free | 331 | and- pictures | 286 |
| the-in | 809 | xxx-free | 569 | the-to | 446 | of-in | 327 | for-the | 284 |
| real-estate | 787 | and-free | 545 | the-the | 419 | university- state | 324 | new- jersey | 280 |
| home-page | 752 | adult-sex | 508 | princess- diana | 410 | sex-nudes | 312 | of-free | 273 |
| free-nude | 720 | and-or | 505 | the-on | 406 | a-to | 304 | chat- rooms | 267 |

A number of term co-occurrences are not topic related, such as *and-and*, *the-to*, etc. The others are a closer representation of a topic sought. Interestingly, high up in the frequency list are nonsexual oriented topics of queries represented by term pairs such as: *university-of*; *new-york*; *real-estate*; *home-page*; *high-school*; etc.

In the *Alta Vista study* a correlation coefficient between pairs of terms was calculated, and significantly correlated terms were identified. This is not the same analysis as the calculation of

the frequency of term pairs, but it still allows for an interpretation of a query topic. In the report, the following are provided as examples of some highly correlated pairs of terms: *cindy-crawford*; *persian-kitty*; *pamela-anderson*; *visual-basic*; *buffy-slayer*; *slayer-vampire*; *buffy-vampire*.

The most highly correlated terms are constituent of phrases, and they reflect topics, for example, three of the entries represent “Buffy the Vampire Slayer”, a TV show. They found that “the strongest correlations resulted from short queries that were actually single-term phrase queries.”

They also provide a list of highly correlated terms that indicate *field = value* (defined as “boolean items of [this form]”) such as *domain = nl* - a term for all queries emanating from The Netherlands. (This further illustrates the difficulty in defining what is a “term”; in the *51K* and *IM study* we did not incorporate these field values as terms). The highly correlated terms representing field values are: *lang = ko - domain = ko*; *date = restricted - applet*; *referred = yes - sessmodlen=4+*; *referred=yes - sessmod = restart*; *the - qwords = 6+*.

Although it makes little sense to count co-occurrences of such values as terms or topics, it provides some (even trivial) explanations, such as that the users from Korea ask for result pages in Korean; and that *applet*, queried by a robot, requests a date restricted set of pages.

A further analysis in the *Alta Vista study* included “phrasifying” - seeking correlation between pairs of already correlated terms. The results include high correlations between the following (*Term A*) and (*Term B*): (*links, kitty*) and (*persian, adult*); (*www.http*) and (*http, com*); (*harvard business*) and *review*; (*used, car*) and (*used, prices*); (*bluemountain, com*) and (*www.bluemountain*); (*anderson, lee*) and (*pamela, lee*); (*ibm, video*) and (*highlander, newsgroups*); (*persian, kitty*) and (*persian, links*).

Most of these three- and four-way correlations involve phrases, describing well-recognized topics. But some, like (*ibm, video*) and (*highlander, newsgroup*) are obscure.

Classification of Queries

The second approach we undertook to answer the question about topics of queries was qualitative and thus more subjective - we used a human classification method for queries. Many search engines also apply human classification as augmentation for automatic classification or clustering that proved to have a degree of inadequacy and inaccuracy.

We took a random sample of 2,414 queries. The sample was stratified to include queries with advanced search features. From the sample, we developed, tested, and applied a classification scheme, using a grounded theory approach (similarly as used in a study of developing a Taxonomy of Value for Library and Information Services in Saracevic & Kantor, [1997]). The Web query classification was developed and applied within a class on classification in the library and information science program at Rutgers University under the leadership of Professor James D. Anderson and by Cheryl Erenberg. Eventually, the scheme, as developed, has 11 major categories, shown in Figure 7.

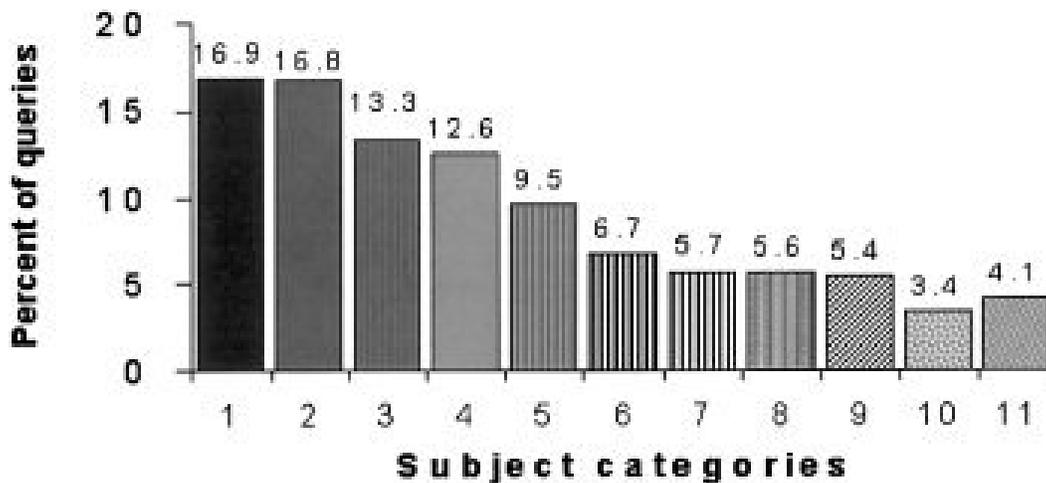


Figure 7. Distribution of a sample of queries across subject categories. (*N* sample = 2414 queries). Legend for subject categories: 1. Entertainment, recreation. 2. Sex, pornography, preferences. 3. Commerce, travel, employment, economy. 4. Computers, the Internet. 5. Health, the sciences. 6. People, places, things. 7. Society, culture, ethnicity, religion. 8. Education, the humanities. 9. The performing and fine arts. 10. Government. 11. Unknown, incomprehensible.

Under each major category there are a number of subcategories, not shown here. The top category in subject of queries was *Entertainment, recreation* (16.9%), closely followed by *Sex, pornography, preferences* (16.8%). It should be noted that not all queries in this category are about pornography; many are about other aspects of sex and sexuality. Thus, in no way is pornography a major topic of Web queries, even though the top ranked terms may indicate this. Only one in about six Web queries is about sex, and, as mentioned, not even all of those are geared toward pornography. The interests of Web users and the topics of their queries are wide ranging. Commerce, including travel, employment, and a number of economic matters is also high on the list. Close to 10% of queries are about health and the sciences; this includes life sciences, medicine, mental health, physical sciences, and engineering. Admittedly, any classification, including this one, has a degree of subjectivity built in, but it is still illustrative, and moreover, such classifications have a high degree of understanding by the public. That is the reason why so many search engines use classification.

Interestingly, the distribution of topics of Web queries, as found here, does not coincide with the distribution of information on the publicly indexable Web, as reported Lawrence and Giles ([1999]). They found that about 83% of servers contain commercial content. The remaining are distributed as follows: 6% of Web servers have scientific/educational content, close to 3% are in health, about 2% each are personal, and societies; pornography was the subject of slightly more than 1% of servers. In frequency distribution, the Web content and subjects of Web queries differ considerably. What is there and what the public asks about is not exactly the same. This conclusion may very well be correct; however, it is only based on the comparison between the Web content in Feb. 1999 and Web user queries in Sept. 1997.

Discussion

We studied a log of over one million Web queries, to discern how the public searches the Web. We also compared the results with two other related studies of large query corpora. Unfortunately, these types of studies, with log data as only available data, cannot answer other very interesting questions about performance results of these queries, or performance of different

search engines. However, they do provide a snapshot for comparison of public behavior while searching, a behavior that can also serve as a clue for improvement of search engines.

We found that a great majority of Web queries posed by the public are short, not much modified, and very simple in structure. Very few queries incorporate advanced search features, and when they do half of them are mistakes. Despite getting, as a rule, a large number of Web sites as answers to their queries, Web users view few result pages; they tend not to browse beyond the first or second page of results. Web users are not much interested in relevance feedback. Overall, a small number of terms are used with very high frequency, while there are great many terms that are used only once. The language of Web queries is very rich, and even unique. The distribution of the subject of Web queries does not follow the distribution of the subject content of Web sites. The number of queries posed on the Web is huge, but searching is a very low art.

Conclusion and Further Research

People are spending more and more time creating, seeking, retrieving, and using electronic information. But their interactions with Web search engines are short and limited. To adjust to these factors and to human behavior we need a new generation of Web searching tools that work with people to help them persist in electronic information seeking to resolve their information problems. Using this study and our current analysis of a data set of 1.7 million Excite queries from 1999, we can begin to identify trends in Web searching.

Acknowledgements

We would like to thank Excite, Inc. and Jack Xu for providing the data for our research. We also thank Steve Lawrence and C. Lee Giles from NEC Research Institute, and Frank Ritter from The Pennsylvania State University for their useful suggestions. The Web query classification was developed and applied within a class on classification in the library and information science program at Rutgers University under the leadership of Professor James D. Anderson, and by Cheryl Erenberg. We acknowledge their contribution and the generous support of our institutions for this research.

References

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: The retrieval effectiveness of search engines. *Information Processing and Management* , 35(2) 141-180.

Huberman, B.A., Pirolli, P., Pitnow, J.E., & Lukose, R.M. (1998). Strong regularities in World Wide Web surfing. *Science* , 280(5360), 95-97.

Jansen, B.J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: A study of user queries on the Web. *SIGIR Forum* , 33(1), 5-17.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). A study of users queries on the Web. *Information Processing and Management* , 36 (2) (Special Issue: Web Research & IR), 207-227.

Jones, S., Cunningham, S.J., & McNab, R. (1998). Usage analysis of a digital library. *Proceedings of the Third ACM Conference on Digital Libraries* (pp. 293-294).

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

Nelson, M.J. (1989). Stochastic models for the distribution of index terms. *Journal of Documentation* , 45(3), 227-237.

OCLC Inc. (1999). Web statistics and analysis. URL: <http://www.oclc.org/oclc/research/projects/webstats/index.htm>.

Ross, N.C.M., & Wolfram, D. (2000). 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* , 51(10), 949-958.

Saracevic, T., & Kantor, P. (1997). Studying the value of library and information services. I. Establishing a theoretical framework. II. Methodology and Taxonomy. *Journal of the American Society for Information Science* , 48(6), 527-542, 543-563.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum* , 33, 3.

Spink, A., & Saracevic, T. (1997). Interaction in information retrieval: Selection and effectiveness of search terms. *Journal of the American Society for Information Science* , 48(8), 728-740.

Spink, A., Bateman, J., & Jansen, B.J. (1999). Searching the Web: Survey of Excite users. *Internet Research: Electronic Networking Applications and Policies* , 9(2) 117-128.

Spink A., Jansen, B.J., & Ozmultu. (2000). Query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policies* , 10(4), 317-328.

Wolfram, D. (1992). Applying informetric characteristics of databases to IR system file design. Part I. Informetric models. *Information Processing and Management* , 28(1), 121-133.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.