PERGAMON

Information Processing and Management 36 (2000) 207–227

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

# Real life, real users, and real needs: a study and analysis of user queries on the web

Bernard J. Jansen[a], Amanda Spink[b,*], Tefko Saracevic[c]

[a]*Department of Electrical Engineering and Computer Science, United States Military Academy, West Point, NY 10996, USA*

[b]*School of Information Sciences and Technology, The Pennsylvania State University, State College, PA 16801, USA*

[c]*School of Communication, Information and Library Studies, Rutgers University, 4 Huntington Street, New Brunswick, NJ 08903, USA*

**Abstract**

We analyzed transaction logs containing 51,473 queries posed by 18,113 users of *Excite*, a major Internet search service. We provide data on: (i) **sessions** — changes in queries during a session, number of pages viewed, and use of relevance feedback; (ii) **queries** — the number of search terms, and the use of logic and modifiers; and (iii) **terms** — their rank/frequency distribution and the most highly used search terms. We then shift the focus of analysis from the query to the user to gain insight to the characteristics of the Web user. With these characteristics as a basis, we then conducted a failure analysis, identifying trends among user mistakes. We conclude with a summary of findings and a discussion of the implications of these findings. © 2000 Elsevier Science Ltd. All rights reserved.

## 1. Introduction

A panel session at the 1997 ACM Special Interest Group on Research Issues In Information Retrieval conference entitled "Real Life Information Retrieval: Commercial Search Engines" included representatives from several Internet search services. Doug Cutting represented *Excite*, one of the major services. Graciously, he offered to make available a set of user queries as submitted to his service for research. The analysis we present here on the nature of sessions, queries, and terms resulted from this offer. Interestingly, the first two authors expressed their

* Corresponding author.

0306-4573/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0306-4573(99)00056-4

interest independently of each other, then met via email, exchanged messages and data, and conducted collaborative research exclusively through the Internet, before ever meeting in person at a Rutgers conference in February 1998, when the results were first presented. In itself, this is an example of how the Internet changed and is changing the conduct of research.

We will argue in the conclusions that real life Internet searching is changing information retrieval (IR) as well. While Internet search engines are based on IR principles, Internet searching is very different from IR searching as traditionally practised and researched in online databases, CD-ROMs and online public access catalogs (OPACS). Internet IR is a different IR, with a number of implications that could portend changes in other areas of IR as well.

With the phenomenal increase in usage of the Web, there has been a growing interest in the study of a variety of topics and issues related to use of the Web. For instance, on the hardware side, Crovella and Bestavros (1996) studied client-side traffic; and Abdulla, Fox and Abrams, (1997) analyzed server usage. On the software side, there have been many descriptive evaluations of Web search engines (e.g. Lynch, 1997). Statistics of Web use appear regularly (e.g. Kehoe, Pitkow & Morton, 1997; FIND/SVP, 1997), but as soon as they appear, they are out of date. The coverage of various Web search engine services was analyzed in several works. A recent article on this topic by Lawrence and Giles (1998) attracted a lot of attention. The pattern of Web surfing by users was analyzed as well (Huberman, Pirolli, Pitkow & Lukose, 1998). However, to date there has been no large-scale, quantitative or qualitative study of Web searching.

*How do they search the Web*? *What do they search for on the Web*? These questions are addressed in a large scale and academic manner in this study. Given the recent yearly exponential increase in the estimated number of Web users, this lack of scholarly research is surprising and disappointing. In contrast, there have been an abundance of user studies of on-line public access categories (OPAC) users. Many of these studies are reviewed in Peters (1993). Similarly, there are numerous studies of users of traditional IR systems. The combined proceedings of the International Conference on Research Issues in Information Retrieval (ACM SIGIR) present many of these studies.

In the area of Web users, however, there were only two narrow studies that we could find. One focused on the THOMAS system (Croft, Cook & Wilder, 1995) and contained some general information about users at that site. However, this study focused exclusively on the THOMAS Web site, did not attempt to characterize Web searching in a systematic way, and is devoted primarily to a description of the THOMAS system. The second paper was by Jones, Cunningham and McNab (1998) and focused again on a single Web site, the New Zealand Digital Library, which contains computer science technical reports. Given the technical nature of this site, it is questionable whether these users represent Web users in general. There is a small but growing body of Web user studies compared to the numerous studies of OPAC and IR system use.

In this paper, we report results from a major and ongoing study of users' searching behavior on the Web. We examined a set of transaction logs of users' searches from *Excite* (http://www.excite.com). This study involved real users, using real queries, with real information needs, using a real search engine. The strength of this study is that it involved a real slice of life on the Web. The weakness is that it involved only a slice — an observable artifact of what the users actually did, without any information about the users themselves or about the results

and uses. Users are anonymous, but we can identify one or a sequence of queries originating with a specific user. We know when they searched and what they searched for, but we do not know anything beyond that. We report on artifactual behavior, but without a context. However, the observation and analysis of such behavior provide for a fascinating and surprising insight into the interaction between users and the search engines on the Web. More importantly, this study provides detailed statistics currently lacking on Web user behavior. It also provides a basis for comparison with similar studies of user searching of more traditional IR and OPAC systems.

The Web has a number of search engines. The approaches to searching, including algorithms, displays, modes of interaction and so on, vary from one search engine to another. Still, all Web search engines are IR tools for searching highly diverse and distributed information resources as found on the Web. But by the nature of the Web resources, they are faced with different issues requiring different solutions than the search engines found in well organized systems, such as in DIALOG, or in lab experiments, such as in the Text Retrieval Conference (TREC) (Sparck Jones, 1995). Moreover, from all that we know, Web users span a vastly broader and thus probably different population of users (Spink, Bateman & Jansen, 1999) and information needs, which may greatly affect the queries, searches, and interactions. Thus, it is of considerable interest to examine the similarities and/or differences in Web searching compared to traditional IR systems. In either case, it is potentially a very different IR.

The significance of this study is the same as all other related studies of IR interaction, queries and searching. By axiom and from lessons learned from experience and numerous studies:

"The success or failure of any interactive system and technology is contingent on the extent 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, 1997).

## 2. Related IR studies

In this paper, we concentrate on users' *sessions, queries*, and *terms* as key variables in IR interaction on the Web. While there are many papers that discuss many aspects of Web searching, most of those are descriptive, prescriptive, or commentary. Other than the two mentioned previously, we could not find any similar studies of Web searching. However, there were several studies that included data on searching of existing, mostly commercial, IR systems, and we culled data from those to provide a basis for comparison between searches as done on the Web and those as done on IR systems outside the Web. A representative sample of such studies is reviewed.

The studies cited below concentrated on different aspects and variables related to searching, using different methodologies and are difficult to compare. Still, each of them had data on the *mean number of search terms* in queries constructed by the searchers under study as follows:

- Fenichel (1981): Novice searchers: 7.9. Moderately experienced: 9.6. Experienced: 14.4

- Hsieh-yee (1993): Familiar topics: Novices: 8.77. Experienced: 7.28. Non-familiar topics: Novices: 9.67. Experienced: 9.00
- Bates, Wilde and Siegfried (1993): Humanities scholars: 14.95
- Spink and Saracevic (1997): Experienced searchers: 14.8.

The studies indicated that searches by various populations contain a range of some 7–15 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 from the *Excite* search engine.

## 3. Background on *Excite* and data

Founded in 1994, Excite Inc. is a major Internet media public company which offers free Web searching and a variety of other services. The company and its services are described at its Web site (http://www.excite.com), thus not repeated here. Only the search capabilities relevant to out results are summarized.

*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. An online thesaurus and concept linking method called Intelligent Concept Extraction (ICE) is used, to find related terms in addition to terms entered. Search results are provided in ranked relevance order. A number of advanced search features are available. Those that pertain to our results are described here:

- As to search logic, Boolean operators AND, OR, AND NOT, and parentheses can be used, but these operators must appear in ALL CAPS and with a space on each side. When using Boolean operator ICE (concept-based search mechanism) is turned off.
- A set of terms enclosed in quotation marks (no space between quotation marks and terms) returns web sites with the terms as a phrase in the exact order they were entered.
- A + (plus) sign before a term (no space) requires that the term must be in an answer.
- A − (minus) sign before a term (no space) requires that the term must NOT be in an answer. We denote plus and minus signs, and quotation marks, as modifiers.
- A page of search results contains ten answers at a time ranked as to relevance. For each site provided is the title, URL (Web site address), and a summary of its contents. Results can also be displayed by site and titles only. A user can click on the title to go to the Web site. A user can also click for the next page of ten answers. In addition, there is a clickable option *More Like This*, which is a relevance feedback mechanism to find similar sites.
- When *More Like This* is clicked, *Excite* enters and counts this as a query with zero terms.

Each transaction record contained three fields. With these three fields, we were able to locate a user's initial query and recreate the chronological series of actions by each user in a session:

1. **Time of Day**: measured in hours, minutes, and seconds from midnight of 9 March 1997.
2. **User Identification**: an anonymous user code assigned by the *Excite* server.
3. **Query Terms**: exactly as entered by the given user.

Focusing on our three levels of analysis — sessions, queries, and terms — we defined our variables in the following way.

1. **Session**: A session is the entire series of queries by a user over a number of minutes or hours. A session could be as short as one query or contain many queries.
2. **Query**: A query consists of one or more search terms, and possibly includes logical operators and modifiers.
3. **Term**: A term is any unbroken string of characters (i.e. a series of characters with no space between any of the characters). The characters in terms included everything — letters, numbers, and symbols. Terms were words, abbreviations, numbers, symbols, URLs, and any combination thereof. We counted logical operators in capitals as terms. However, in a separate analysis we isolated them as commands, not terms.

The raw data collected are very messy. Users entered terms, commands and modifiers in all kinds of ways, including many misspellings and mistakes. In many cases, *Excite* conventions were not followed. We count these deviations as mistakes and report them in the failure analysis portion of the paper. For the most part, we took the data 'as is,' i.e., we did not 'clean' the data in any way — these queries represent real searches by real users. The only normalization we undertook in one of the counts (unique terms — not case sensitive) was to disregard capitalization, because *Excite* disregards it as well. (i.e. *TOPIC, topic* and *Topic* retrieve the same answers). *Excite* does not offer automatic stemming, thus *topic* and *topics* count as two unique terms, and '?' or '*' as stemming commands at the end of terms are mistakes, but when used counted as separate terms. We also analyzed a cleaned set of terms, that is we removed term modifiers such as the + or − signs. We took great care in the derivation of counts due to the 'messiness' of data. This paper extends findings from (Jansen, Spink, Bateman & Saracevic, 1998a,b; Jansen, Spink & Saracevic, 1998c).

## 4. Results

First, what is the pattern of user queries? We looked at the number of queries by each specific user and how successive queries differed from other queries by the same user. We classified the 51,474 queries as to *unique, modified*, or *identical* as shown in Table 1.

A unique query was the first query by a user (this represents the number of users). A modified query is a subsequent query in succession (second, third ...) by the same user with terms added to, removed from, or both added to and removed from the unique query. Unique and modified queries together represent those queries where the user did something with terms. Identical queries are queries by the same user that are identical to the query preceding it. They

Table 1
Unique, modified, and identical queries

| Query type | Number | Percent of all queries |
|---|---|---|
| Unique | 18,098 | 35% |
| Modified | 11,249 | 22% |
| Identical | 22,127 | 43% |
| Total | | 100% |

can come about in two ways. The first possibility is that the user retyped the query. Studies have shown that users often do this (Peters, 1993). The second possibility is that the query was generated by *Excite*. When a user views the second and further pages (i.e., a page is a group of 10 results) of results from the same query, *Excite* provides another query, but a query that is identical to the preceding one. Our analysis did not allow disambiguation of these two causes of identical queries.

The unique plus modified queries (where users actively entered or modified terms) amounted to 29,437 queries or 57% of all queries. If we assume that all identical queries were generated as a request to view subsequent pages, then 43% of queries come as a result of a desire to view more pages after the first one. Modifications and viewing are further elaborated in the next two tables.

### 4.1. Modifications

Some users used only one query in their session, others used a number of successive queries. The average session, including all three query types, was 2.84 queries per session. This means that a number of users went on to either modify their query, view subsequent results, or both. The average session length, ignoring identical queries, was 1.6 queries per user. Table 2 lists the number of queries per user.

This analysis includes only the 29,337 unique and modified queries. We ignored the identical queries because as stated above, it was impossible to interpret them meaningfully in this context, in order to concentrate only on those queries where users themselves did something to the queries. A substantial majority of users (67%) did not go beyond their first and only query.

Table 2
Number of queries per user

| Queries per user | Number of users | Percent of users |
|---|---|---|
| 1 | 12,068 | 67 |
| 2 | 3501 | 19 |
| 3 | 1321 | 7 |
| 4 | 583 | 3 |
| 5 | 287 | 1.6 |
| 6 | 144 | 0.80 |
| 7 | 79 | 0.44 |
| 8 | 32 | 0.18 |
| 9 | 36 | 0.20 |
| 10 | 17 | 0.09 |
| 11 | 7 | 0.04 |
| 12 | 8 | 0.04 |
| 13 | 15 | 0.08 |
| 14 | 2 | 0.01 |
| 15 | 2 | 0.01 |
| 17 | 1 | 0.01 |
| 25 | 1 | 0.01 |

Query modification was not a typical occurrence. This finding is contrary to experiences in searching regular IR systems, where modification of queries is much more common. Having said this, however, 33% of the users did go beyond their first query. Approximately 14% of users had entered three or more queries. These percentages of 33% and 14% are not insignificant proportions of system users. It suggests that a substantial percentage of Web users do not fit the stereotypical naïve Web user. These sub-populations of users should receive further study. They could represent sub-populations of Web users with more experience or higher motivation who perform query modification on the Web.

We also examined how user modified their queries. These results are display in Table 3.

Here we concentrate on the 11,247 queries that were modified by either an increase or a decrease in the number of terms from one user's query to that user's next query (i.e., successive queries by the same user at time T and T + 1). Zero change means that the user modified one or more terms in a query, but did not change the number of terms in the successive query. Increase or decrease of one means that one term was added to or subtracted from the preceding query. Percent is based on the number of queries in relation to all modified (11,247) queries.

We can see that users typically do not add or delete much in respect to the number of terms

Table 3
Changes in number of terms in successive queries

| Increase in terms | Number | Percent |
|---|---|---|
| 0 | 3909 | 34.76 |
| 1 | 2140 | 19.03 |
| 2 | 1068 | 9.50 |
| 3 | 367 | 3.26 |
| 4 | 155 | 1.38 |
| 5 | 70 | 0.62 |
| 6 | 22 | 0.20 |
| 7 | 6 | 0.05 |
| 8 | 10 | 0.09 |
| 9 | 1 | 0.01 |
| 10 | 4 | 0.04 |
| **Decrease in terms** | **Number** | **Percent** |
| −1 | 1837 | 16.33 |
| −2 | 937 | 8.33 |
| −3 | 388 | 3.45 |
| −4 | 181 | 1.61 |
| −5 | 76 | 0.68 |
| −6 | 46 | 0.41 |
| −7 | 14 | 0.12 |
| −8 | 8 | 0.07 |
| −9 | 2 | 0.02 |
| −10 | 6 | 0.05 |

in their successive queries. Modifications to queries are done in small increments, if at all. The most common modification is to change a term. This number is reflected in the queries with zero (0) increase or decrease in terms. *About one in every three queries that is modified still had the same number of terms as the preceding one*. In the remaining 7338 successive queries where terms were either added or subtracted about equal number had terms added as subtracted (52–48%) — thus users go both ways in increasing and decreasing number of terms in queries. *About one in five queries that is modified has one more term than the preceding one, and about one in six has one less term.*

### 4.2. Viewing of results

*Excite* displays query results in groups of ten. Each time that a user accesses another group of 10, which we term another page, an identical query is generated. We analyzed the number of pages each user viewed and the percentage that this represented based on the total number of users. The results are shown in Table 4.

The mean number of pages examined per user was 2.35. Most users, 58% of them, did not access any results past the first page. Were they so satisfied with the results that they did not need to view more? Were a few answers good enough? Is the precision that high? Are the users after precision? Or did they just give up and get tired of viewing results? Using only transaction logs, we cannot provide answers to these questions. But in any case, this result combined with the small number of queries per session has interesting implications for recall and may illustrate a need for high precision in Web IR algorithms. For example, using a classical measurement of precision, any search result beyond the tenth position in the list would be meaningless for 58% of Web users. Another possible interpretation is that people use partially relevant items in the first page to avoid further searching through subsequent pages. Given the hypertext nature of the Web, partially relevant items (Spink, Greisdorf & Bateman, 1998) in the top ten maybe used as a jumping off point to find relevant items. For example, a user looking for a faculty member's homepage at a university does not retrieve the faculty's homepage in the top ten but gets the university homepage. Rather than continue searching, the user starts browsing, beginning with the university page.

### 4.3. Queries

From the session level of analysis, we then moved to the query level. The basic statistics related to queries and search terms are given in Table 5.

We analyzed the queries based on length (i.e. number of terms), structure (use of Boolean operators and modifiers), and failure analysis (deviations from published rules of query construction). We also identified the number of users of Boolean logic and modifiers.

#### 4.3.1. Length

On the average, a query contained 2.21 terms. Table 6 shows the ranking of all queries by number of terms.

Percent is the percentage of queries containing that number of terms relative to the total number of queries. Web queries are short. About 62% of all queries contained one or two

Table 4
Number of pages viewed per user

| Pages viewed | Number of users | Percent of all users |
|---|---|---|
| 1 | 10,474 | 58 |
| 2 | 3363 | 19 |
| 3 | 1563 | 9 |
| 4 | 896 | 5 |
| 5 | 530 | 3 |
| 6 | 354 | 2 |
| 7 | 252 | 1 |
| 8 | 153 | 0.85 |
| 9 | 109 | 0.60 |
| 10 | 85 | 0.47 |
| 11 | 75 | 0.41 |
| 12 | 47 | 0.26 |
| 13 | 31 | 0.17 |
| 14 | 29 | 0.16 |
| 15 | 25 | 0.14 |
| 16 | 28 | 0.15 |
| 17 | 13 | 0.07 |
| 18 | 4 | 0.02 |
| 19 | 14 | 0.08 |
| 20 | 9 | 0.05 |
| 21 | 3 | 0.02 |
| 22 | 4 | 0.02 |
| 23 | 5 | 0.03 |
| 24 | 7 | 0.04 |
| 25 | 4 | 0.02 |
| 26 | 7 | 0.04 |
| 27 | 2 | 0.01 |
| 28 | 3 | 0.02 |
| 29 | 1 | 0.01 |
| 32 | 4 | 0.02 |
| 33 | 1 | 0.01 |
| 40 | 1 | 0.01 |
| 43 | 1 | 0.01 |
| 49 | 1 | 0.01 |
| 50 | 2 | 0.01 |
| 55 | 1 | 0.01 |

Table 5
Numbers of users, queries, and terms

| No. of users | Total no. of queries | Non-unique terms | Mean no. of terms per query (range) | Unique terms with case sensitive | Unique terms without case sensitive |
|---|---|---|---|---|---|
| 18,113 | 51,473 | 113,793 | 2.21 (0–10) | 27,459 | 21,862 |

Table 6
Number of terms in queries ($N$ queries = 51,473)

| Terms in query | Number of queries | Percent of all queries |
|---|---|---|
| 10 | 185 | 0.36 |
| 9 | 125 | 0.24 |
| 8 | 224 | 0.44 |
| 7 | 484 | 0.94 |
| 6 | 617 | 1 |
| 5 | 2158 | 4 |
| 4 | 3789 | 7 |
| 3 | 9242 | 18 |
| 2 | 16,191 | 31 |
| 1 | 15,874 | 31 |
| 0 | 2584 | 5 |

terms. Fewer than 4% of the queries had more than 6 terms. As mentioned, we could not find any other data on Web searches from a major Web search engine, thus, the only comparisons are with the two smaller studies by Croft et al. (1995) and Jones et al. (1998). The query length observed in our research is similar to results from these two studies. This deviates significantly from traditional IR searching. As shown above, the mean number of search terms used in searching of regular IR systems ranged from about 7 to 15. This is about three to seven times higher than the mean number of times found in this study, and our count is on the high side, because we counted operators as well. Admittedly, the circumstances and context between searches done by users of IR systems such as DIALOG, and searches of the Web, done by the general Internet population, may be vastly different. Thus this comparison may have little meaning.

*4.3.2. Relevance feedback*

A note should be made on queries with zero terms (last row of Table 6). As mentioned, when a user enters a command for relevance feedback (*More Like This*), the *Excite* transaction log counts that as a query, but a query with zero terms. Thus, the last row represents the largest possible number of queries that used relevance feedback, or a combination of those and queries where users made some mistake that triggered this result. Assuming they were all relevance feedback, only 5% of queries used that feature — a small use of relevance feedback capability. In comparison, a study involving IR searches conducted by professional searchers as they interact with users found that some 11% of search terms came from relevance feedback (Spink & Saracevic, 1997), albeit this study looked at human initiated relevance feedback. Thus, in these two studies, relevance feedback on the Web is used half as much as in traditional IR searches. This in itself warrants further study, particularly given the low use of this potentially highly useful and certainly highly vaunted feature.

*4.3.3. Structure*

Next, we examined the structure of queries, focusing first on how many of the 51,473 queries explicitly utilized Boolean operators or modifiers (see Table 7).

Table 7
Use of Boolean operators and modifiers in queries ($N$ queries = 51,473)

| Operator or modifier | Number of queries | Percent of all queries | Incorrect | Percent incorrect |
|---|---|---|---|---|
| AND | 4094 | 8 | 1309 | 32 |
| OR | 177 | 0.34 | 46 | 26 |
| AND NOT | 105 | 0.20 | 39 | 37 |
| ( ) | 273 | 0.53 | 0 | 0 |
| + (plus) | 3010 | 6 | 1182 | 39 |
| − (minus) | 1766 | 3 | 1678 | 95 |
| " " | 3282 | 6 | 179 | 5 |

The Number column lists the number of queries that contained that particular Boolean operator or modifier. The next column is the percentage that number represents of all queries. Incorrect means the number of queries containing a specific operator or modifier that was constructed not following *Excite* rules — they could be considered as mistakes. The last column is the percentage of queries containing a given operator or modifier that were incorrectly constructed. We discuss the failures in a later section.

From Table 7, at least one thing is obvious — Boolean operators were not used much, with AND receiving the greatest use. These numbers were significantly lower than those reported by Jones et al. (1998), and significantly lower than studies of searches from IR systems and OPAC systems [Croft et al. (1995) did not report this information]. Modifiers were used a little more often, with the '+' and "" (i.e., phrase searching) being used the most. For example, based on what we reviewed so far in this paper, we have a large set of queries that are extremely short, seldom modified, and very simple in structure. Yet, the vast majority of users never viewed anything beyond the first 10 results. Is the recall and precision rate of *Excite* that good? Is something else at work here? One interpretation may be that users only glance at the first page to see how poorly they performed their search. Rather than taking time to learn the detailed procedures of *Excite*, they try anything (trial and error) and then try to judge from the hits what they did wrong.

Table 8
Use of logic and modifiers by users ($N$ users = 18,113)

| Operator or modifier | Number of users using it | Percent of all users | Incorrect | Percent incorrect |
|---|---|---|---|---|
| AND | 832 | 5 | 418 | 50 |
| OR | 39 | 0 | 11 | 28 |
| AND NOT | 47 | 0 | 9 | 19 |
| ( ) | 120 | 1 | 0 | 0 |
| + (plus) | 826 | 5 | 303 | 30 |
| − (minus) | 508 | 3 | 362 | 38 |
| " " | 1019 | 6 | 32 | 0 |

*4.3.4. Number of users*

In Table 8, we examine how many of the 18,113 users, opposed to the number of queries, used any Boolean logic (first four rows) or modifiers (last three rows) in their queries (regardless of how many queries they had).

We relate these numbers to the number of queries. Incorrect means the number of users committing mistakes by not following *Excite* rules as stated in instructions for use of these operators and modifiers. Percent incorrect is the proportion of those users using a given operator or modifier incorrectly or as a mistake. The user population that incorporated Boolean operators was very small. Only 6% of the 18,113 users used any of the Boolean capabilities, and these were used in less than 10% of the 51,473 queries. A minuscule percentage of users and queries used OR or AND NOT. Only about 1% of users and $\frac{1}{2}$% of queries used nested logic as expressed by a use of parentheses. The '+' and '−' modifiers were used by about the same number of people that used Boolean operators. Together '+' and '−' were used by 1334 or 7% of users in 4776 (9%) queries. The ability to create phrases (terms enclosed by quotation marks) was also seldom used — only 6% of users and 6% of queries used them. From this, it appears that a small number of users account for the occurrences of the more sophisticated queries, indicating that there is little experimentation by users during their sessions. About 5% of the users account for the 8.5% of queries that contained Boolean operators. We discuss the ramifications of this finding for system design later in the paper.

## 5. Failure analysis

Next, we turn to a discussion of the surprisingly high number of incorrect uses or mistakes. When they used it, 50% of users made a mistake in the use of the Boolean AND; 28% made an error in uses of OR, and only 19% used AND NOT incorrectly, but only 47 users, a negligible percent, used AND NOT at all. The most common mistake was not capitalizing the Boolean operator, as required by the *Excite* search engine. For example, a correct query would be: information AND processing. The most common mistake would be: information and processing.

When we look at queries, 32% contained an incorrect use of AND, 26% of OR, and 37% of AND NOT. 'AND' presents a special problem, so we did a further analysis. We had 4094 queries that used AND in some form (as 'AND,' ''And, or 'and'). Some queries had more than one AND. Altogether, there were 4828 appearances of all forms of AND: 3067 as 'AND', 41 as 'And,' and 1720 as 'and.' If considered as Boolean operators, the last two or 1761 instances were mistakes. Most of them were, but not all. In a number of queries 'and' was used as conjunction e.g. as in query *College and university harassment policy*. Unfortunately, we could not distinguish the intended use of 'and' as a conjunction from that as a mistake, thus our count of AND mistakes is on the high end.

There was a similarly high percentage of mistakes in the use of plus and minus operators — respectively 30% and 38%. Most of the time, spaces were used incorrectly. Minus presents an especially vexing problem, because it is also used in phrases such as *pre-teen*. Thus, our count of mistakes is at the high end. It is easy to see that Web users are not up to Boolean, and even less to follow searching rules. At the very least, system redesign seems to be in order. The most

common mistake was stringing all the terms of the query together, as in a mathematical formula. For example, a correct query would be: +information +processing. The most common mistake would be: +information+processing (with no space between information and the next +). Consistent spacing rules between Boolean operators and term modifier may solve this problem. In the use of Boolean operator, a space between the operator and the term is required. With the use of term modifiers, the space must not be there.

There were also a large number of queries that incorporated searching techniques that *Excite* does not support. These failures can be classified as a carry over from user learning associated with other search engines, including those from other Web, OPAC, and IR systems. For example, there were 26 occurrences of the proximity operator NEAR. There were 79 uses of the ':' as a separator for terms. There were numerous occurrences of '.' used as a term separator. The symbol '&' was used in-lieu of the Boolean AND over 200 times. These symbols are common in many other search engines.

## 6. Terms

We also analyzed user queries according to the terms they included. A term was any series of characters bounded by white space. There were 113,793 terms (all terms from all queries). After eliminating duplicate terms, there were 21,862 unique terms that were non-case sensitive (in other words, all upper cases are here reduced to lower case). In this distribution logical operators AND, OR, NOT were also treated as terms, because they were used not only as operators but also as conjunctions (we already discussed the case of '*and*.' and presented the figures for various forms of the term, thus subtraction can be easily done). We discuss terms from the perspective of their occurrence, their fit with known distributions, and classification into some broader subject headings.

### 6.1. Occurrences

We constructed a complete rank-frequency table for all 113,793 terms. Out of the complete rank-frequency-table we took the top terms, i.e. those that appeared 100 times or more, as presented in Table 9.

The 74 terms that were used 100 or more times across all queries appeared a total of 20,698 times as search terms. They represent only 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 9121 occurrences of 11 common terms that do not carry any content by themselves (and, of, the, in, for, +, on, to, or, &, a), we are left with 63 subject terms that have a total frequency of 11,577 occurrences — that is 0.29% of unique subject terms account for 10.3% of all terms in all queries. The high appearance of '+' represents a probable mistake — the inclusion of space between the sign and a term, as required by *Excite* rules.

Similarly, '&'' was used often as a part of an abbreviation, such as in AT&T, but also as a substitute for logical AND, as in Ontario & map. In the latter case, it is a mistake and would appear as a separate term. On the other end of the distribution we have 9790 terms that appeared only once. These terms amounted to 44.78% of all unique terms and 8.6% of all

Table 9
Listing of terms occurring more than 100 times (**** = expletive)

| Term | Frequency | Term | Frequency | Term | Frequency |
|---|---|---|---|---|---|
| and (incl. 'AND', & 'And') | 4828 | & | 188 | estate | 123 |
| Of | 1266 | stories | 186 | magazine | 123 |
| The | 791 | p**** | 182 | computer | 122 |
| Sex | 763 | college | 180 | news | 121 |
| Nude | 647 | naked | 180 | texas | 119 |
| Free | 610 | adult | 179 | games | 118 |
| In | 593 | state | 176 | war | 117 |
| Pictures | 457 | big | 170 | john | 115 |
| For | 340 | basketball | 166 | de | 113 |
| New | 334 | men | 163 | internet | 111 |
| + | 330 | employment | 157 | car | 110 |
| University | 291 | school | 156 | wrestling | 110 |
| Women | 262 | jobs | 155 | high | 109 |
| Chat | 256 | american | 153 | company | 108 |
| On | 252 | real | 153 | florida | 108 |
| Gay | 234 | world | 152 | business | 107 |
| Girls | 223 | black | 150 | service | 106 |
| Xxx | 222 | porn | 147 | video | 105 |
| To | 218 | photos | 142 | anal | 104 |
| Or | 213 | york | 140 | erotic | 104 |
| Music | 209 | A | 132 | stock | 102 |
| Software | 204 | Young | 132 | art | 101 |
| Pics | 202 | History | 131 | city | 100 |
| Ncaa | 201 | Page | 131 | porno | 100 |
| Home | 196 | Celebrities | 129 | | |

terms in all queries. The tail end of unique terms is very long and warrants in itself a linguistic investigation. In fact, the whole area of query language needs further investigation. There are no comprehensive studies of terms, the distribution of those terms, the modification of those terms, etc., of Web queries. Such studies have potential to benefit IR system and Web site development.

### 6.2. Term categories

In order to ascertain some broad subjects of searching, we classified the 63 top subject 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. These subjects are displayed in Table 10.

About 25% of the highest used terms apparently dealt with some or other sexual topic. However, that represents fewer than 3% of all terms. Of course, if one classifies additional terms further down the distribution (such as those listed in the "Gender" category as *Sexual*) the percent will be higher. We perused the rest of the terms and came to the conclusion than no more than some two dozen of the other terms will unmistakably fall into that category. If we added them all together, the frequency of terms in *Sexual* will increase but not that much,

Table 10
Subject categories for terms appearing more than 100 times

| Category | Terms selected from 63 terms with frequency of 100 and higher | Frequency for category | Percent of frequency -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 |
| Fine arts | *music, art* | 310 | 2.68 | 0.72 |

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 the highest frequency of all categories, but still three out of every four terms of the 63 highest frequency terms are not sexual; if extended to the frequency of use of all terms we estimate that 39 out of 40 of all terms were not sexual.

While the category *Sexual* is certainly big, in comparison to all other categories in no way does it dominate searching. Interest in other categories is high. Of the 63 highest terms, 16% are modifiers (free, new, big ...), 10% deal with places (state, american ...), 8% with economics (employment, jobs ...), and the rest with social activities, education, sports, computing, and arts. In other words, Web searching does cover a gamut of human interests. It is very diverse. In light of this, the stereotypical view of the Web user searching primarily for sexual information may not be valid.

There are two other groupings not listed in the table that should be noted. First, there were 1398 queries for various uniform resource locators (URL). Although no one URL made the top of the list, if lumped together as a category, it was one of, if not the largest query category. The second group was searching for multimedia documents (e.g. images, videos, and audio files). There were 708 queries for these multimedia files, with many of these terms looking for specific formats.

### 6.3. Distribution of terms

We constructed a graph of rank–frequency distribution of all terms. This graph is shown in Fig. 1.

The resulting distribution seems to be unbalanced at the ends of the graph, the high and low ranking terms. In the center and lower regions, the graph follows the traditional slope of a Zipf distribution representing the distribution of words in long English texts. At the beginning, it falls off very gently, and toward the end it shows discontinuities and an unusually long tail, representing terms with a frequency of one. A trend line is plotted on the figure with the corresponding equation. The trend line is approximately that of the Zipf distribution. A proper Zipf distribution would be a straight line with slope of $-1$. The trend line does not plot well for the higher frequency terms due to the large number of terms occurring only once or twice.

We wondered if the number of modifiers (e.g. '+', '–', '', etc.) and the number of queries with all terms strung together (e.g. +information+processing+journal) could be affecting the rank–frequency distribution, that is, if the number of modifiers, stray characters, and run-on terms, were creating such a long tail of single occurrence terms. Therefore, we decided to clean all terms and re-plot the rank–frequency graph. In cleaning terms, we removed all modifiers and separated all terms that were obviously strung together. Due to the varying nature of the terms, this could not be done automatically. For example, one could not just remove '+' from all terms because, for example, with c++ (the programming language), the '+' is part of a valid term. In the cleaning process, all 113,793 terms were qualitatively examined. In most cases, a decision would clearly be made on whether or not to clean the term. In cases were there was doubt, the term was not modified. Once clean, we again generated a rank–frequency (log) plot. This rank–frequency plot is shown in Fig. 2.

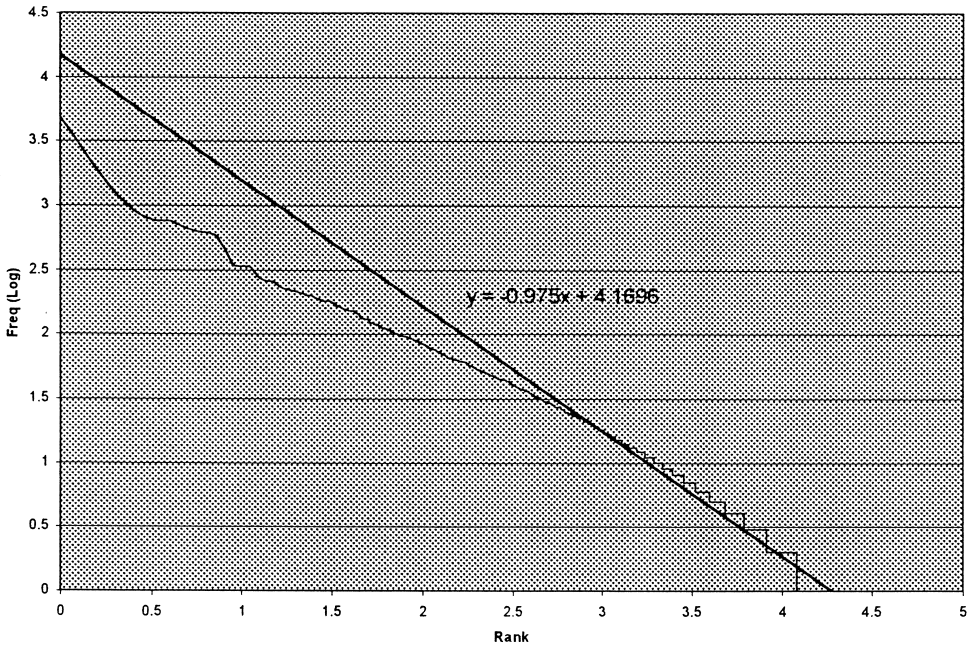


Fig. 1. Rank vs frequency (log) of all terms.

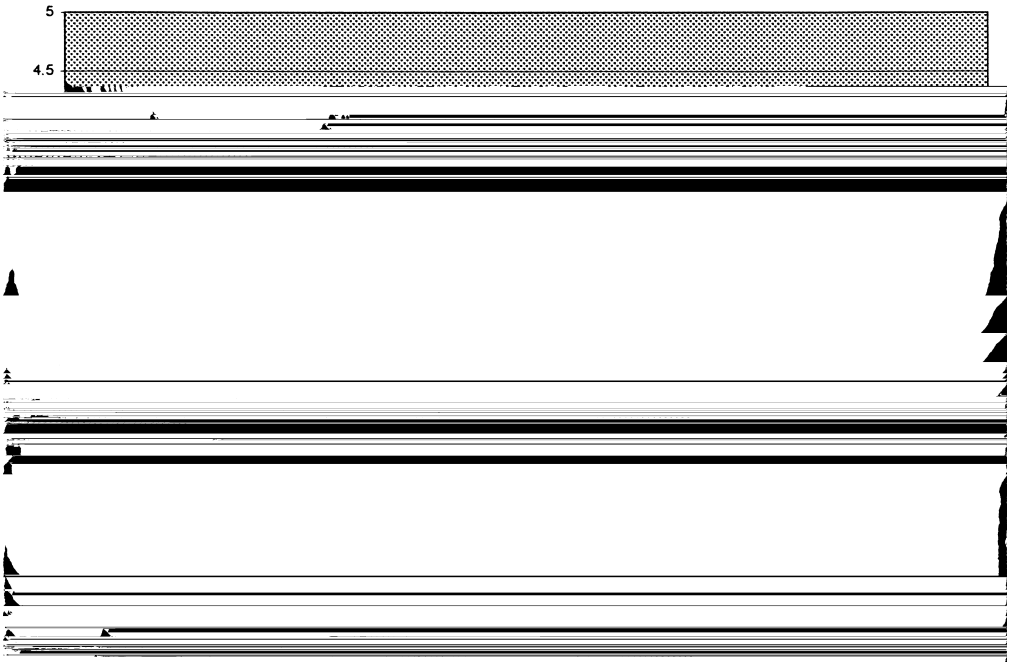


Fig. 2. Rank (log) vs frequency (log) of cleaned terms.