Real life information retrieval: a study of user queries on the Web

Major Bernard J. Jansen
Department of Electrical Engineering and Computer Science,
United States Military Academy
West Point, New York 10996 USA
jjansen@acm.org

Amanda Spink & **Judy Bateman**
School of Library and Information Sciences,
University of North Texas
P.O. Box 311068, Denton TX 75203 USA
spink@lis.admin.unt.edu, jbateman@unt.edu

Tefko Saracevic
School of Communication, Information and Library Studies,
Rutgers University
4 Huntington Street, New Brunswick, NJ 08903 USA
tefko@scils.rutgers.edu

Please Cite: Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. 1998. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, Vol. 32. No. 1., pp. 5 -17.

See Other Publications

ABSTRACT. We analyzed transaction logs of a set of 51,473 queries posed by 18,113 users of *Excite*, a major Internet search service. We provide data on: (i) **queries** - the number of search terms, and the use of logic and modifiers, (ii) **sessions** – changes in queries during a session, number of pages viewed, and use of relevance feedback, and (iii) **terms** - their rank/frequency distribution and the most highly used search terms. Common mistakes are also observed. Implications are discussed**.**

INTRODUCTION

A panel session at the 1997 SIGIR 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 queries, sessions, and terms resulted from this offer. Interestingly, the authors expressed their interest independently of each other, then met, 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 presented first. In itself, this is an example how the Internet changed conduct of research. We will argue in the conclusions that real life Internet searching is changing IR as well. While Internet search engines are based on IR principles, Internet searching is very different then IR searching as traditionally practiced and researched. Internet IR is a different IR, with a number of implications.

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 & Besravros (1996) studied client-side traffic; and Abdulla, et. al., (1997) analyzed server use. 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 et. al. 1997, FIND/SVP, 1997), but as soon as they appear they are out of date. The coverage of various Web search services was analyzed in several works - a recent article on the topic by Lawrence & Giles (1998) attracted a lot of attention. The pattern of Web surfing by users was analyzed as well (Huberman et al. 1998). 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? 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 from *Excite* (http://www.excite.com). The study involved real users, and their queries as they searched *Excite*. The strength of the 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. The users are anonymous. 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, albeit limited, behavior still provides for a fascinating, and somewhat surprising insight about the interaction between users and the search engines on the Web, similarly as was found in many studies of searching of more traditional IR 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. They all follow the basic principles of IR and human-computer interaction. 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). Moreover, from all that we know, Web users spans a vastly broader (and thus probably different) population of users and use, which may greatly reflect on 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 IR, but 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 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, 1997).

RELATED IR STUDIES

In this paper, we concentrate on users' *queries, sessions,* and *terms* as key variables in IR interaction on the Web. While there are many papers that discuss aspects of Web searching, most of those are descriptive, prescriptive, or commentary. 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 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, thus, they are hard 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 (1993): Humanity scholars: 14.95
- Spink & Saracevic (1997): Experienced searchers: 14.8.

The studies indicated that searches by 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 public company which offers free Web searching and a variety of other services. The company and its services are described at its Web site, thus not repeated here. Only search capabilities relevant to 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. Up to ten search terms are allowed in a query. 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 a 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 operators ICE (concept-based search mechanism) is turned off.
- A set of terms enclosed in quotation marks (no space between quotation marks and terms) returns answers with the terms as a phrase in exact order.
- 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.

Here is how things were counted. A query consists of one or more search terms, and possibly includes logical operators and modifiers. A term is any unbroken string of characters (i.e. no space between 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 is 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 – these we count as mistakes. 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 without case sensitive) was to disregard capitalization, because *Excit*e 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 took great care in derivation of counts, but because of the 'messiness' of data there still may be errors – we estimate at less than 1%.

QUERIES

The basic statistics related to queries and search terms are given in Table 1. We provide three statistics: (1) *Non-unique terms:* sum of all terms over all queries with 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, and (3) *Unique terms with case non-sensitive:* the three capitalization forms of *topic* are counted as one term.

No. of users	Total no. of queries	Non-unique terms	Mean of terms 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 1. Numbers of users, queries, and terms

There were on the average 2.84 queries per user, meaning that a number of users went on and refined in some way their query. On the average, a query contained 2.35 terms. As mentioned, we could not find any data on Web searches, thus, we can not compare this average to other Web searching. However, some comparison with IR searching can be made. As we showed above, the mean number of search terms in searching of regular IR systems ranged from about 7 to 15. This is about three to seven magnitudes higher than found in this study, and even this 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 are vastly different, thus this comparison may have little meaning. But still, it is interesting to make the comparison.

As mentioned, *Excite* accepts queries from 1 to 10 terms. Table 2 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. Less than 4% of the queries had more than 6 terms. A note should be made on queries with zero terms (last row). As mentioned, when a user enters a command for relevance feedback *(More Like This*), *Excite* counts that as a query, but a query with zero terms. Thus, the last row represents the potentially largest number of queries that used relevance feedback, or a combination of those and queries where user made some mistake that triggered this result. 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). Thus, the relevance feedback on the Web is used half as much as in traditional IR searches. But it is surprising that in either case the users use relatively very little this potentially highly useful and certainly highly vaunted feature.

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	2,158	4

4	3,789	7
3	9,242	18
2	16,191	31
1	15,854	31
0	2,584	5

TABLE 2: Number of terms in queries. (N queries = 51,473

In Table 3. we examine how many of the **18,113 users** used any Boolean logic (first four rows) or modifiers (last three rows) in their queries (regardless of how many queries they had). 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 proportion of those users using a given operator or modifier immortally or as a mistake.

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
" "	1,019	6	32	0

TABLE 3. Use of logic and modifiers by users. (N users = 18,113)

Next we examined how many of the **51,473 queries** explicitly utilized Boolean operators or modifiers as presented in Table 4. 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.

Operator or modifier	**Number of queries**	**Percent of all queries**	**Incorrect**	**Percent incorrect**
AND	4094	8	1,309	32
OR	177	0.34	46	26

AND NOT	105	0.20	39	37
( )	273	0.53	0	0
+ (plus	3,010	6	1,182	39
- (minus)	1,766	3	1,678	95
" "	3,282	6	179	5

TABLE 4.: Use of Boolean operators and modifiers in queries. (N queries = 51,473)

Boolean operators were used very sparingly. 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. In those, AND was used by far the most. A miniscule percentage of users and queries used OR or AND NOT. Only about 1% of users and ½% of queries used nested logic as expressed by a use of parentheses. The '+' and '-' modifiers were used about the same as Boolean operators. Together '+' and '-' were used by 1,334 or 7% of users in 4,776 (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

Next we turn to a discussion of the surprisingly high number of incorrect uses or mistakes. However, as will be seen, there are a number of judgement calls on what constitutes a mistake. When they used it, a whooping 50% of users made a mistake in use of the Boolean AND; 28% in uses of OR, and only 19% in uses of AND NOT, but only 47 users, a negligible percent, used AND NOT at all. When we look at queries, 32% contained 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 4,094 queries that used AND in some form (as 'AND,' "And, and 'and'). Some queries had more than one AND. Altogether, there were 4,828 appearances of all forms of AND: 3,067 as 'AND', 41 as 'And,' and 1,720 as 'and.' If considered as Boolean operators, the last two or 1,761 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 higher end. There was a similar high percent of mistakes in use of plus and minus operators – respectively 30% and 38%. Most of the time spaces were used incorrectly. Minus presents a specially vexing problem, because it is also used in phrases such as *pree-teen.* It is easy to see that Web users are not up to Boole, and even less to rules. Redesign seems to be in order.

SESSIONS

Next we looked 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 6. A unique query was the first query by a user (this represents the number of users, including an error). A modified query is a subsequent query in succession (second, third …) by the same user with terms added to or removed from the unique query. Unique and modified queries together represent those queries where user did something with terms. Identical queries are queries by the same user that are identical to the query previous to it. They can come about in two ways. First is that the user retyped the query. Second is generated by *Excite*: when viewing the second and further pages with the same query Excite provides a another query for this, but a query that is identical to the preceding one.

Query Type	**Number**	**Percent of all queries**
Unique	18,098	35

Modified	11,249	22
Identical	22,127	43

TABLE 5: Unique, Modified, and 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 request for viewing subsequent pages, then 43% of queries come as a result of desire to view more pages after the first one. Modifications and viewing are further elaborated in the next two tables.

Some users used only one query in their session, others used a number of successive queries. Table 6 lists the number of queries per user. This analysis includes only the 29,337 unique and modified queries, in order to concentrate only on those queries where users themselves did something to the queries. A big majority of users did not go beyond their first and only query. Some 67% of users had one and only query. Query modification was not a strong trend. This is contrary to experiences in searching of regular IR systems, where modification of queries is very much a way of doing things.

Excite displays query results in-groups of 10. 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 7.

Queries per user	**Number of users**	**Percent of users**	**Queries per user**	**Number of users**	**Percent of users**
1	12,068	67	10	17	0.09
2	3,501	19	11	7	0.04
3	1,321	7	12	8	0.04
4	583	3	13	15	0.08
5	287	29	14	2	0.01
6	144	0.80	15	2	0.01
7	79	0.44	17	1	0.01
8	32	0.18	25	1	0.01
9	36	0.20			

TABLE 6: Number of Queries Per User

Pages viewed	**Number of users**	**Percent of all users**	**Pages viewed**	**Number of users**	**Percent of all users**

1	10,474	58	21	3	0.02
2	3,363	19	22	4	0.02
3	1,563	9	23	5	0.03
4	896	5	24	7	0.04
5	530	3	25	4	0.02
6	354	2	26	7	0.04
7	252	1	27	2	0.01
8	153	0.85	28	3	0.02
9	109	0.60	29	1	0.01
10	85	0.47	32	4	0.02
11	75	0.41	33	1	0.01
12	47	0.26	40	1	0.01
13	31	0.17	43	1	0.01
14	29	0.16	49	1	0.01
15	25	0.14	50	2	0.01
16	28	0.15	55	1	0.01
17	13	0.07			
18	4	0.02			
19	14	0.08			
20	9	0.05			

TABLE 7: Number of Pages Viewed Per User.

The mean number of pages examined per user was 2.21. 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 needed to go viewing more? Is the precision that high? And are the users after precision – few answers were good enough? Or did they just give up? Who knows? But in any case, this, of course, has interesting implications for recall and may illustrate a need for high precision in Web IR algorithms.

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, AND NOT were also treated as

terms, because they were used not only as operators but also as conjunctions. We discussed already the case of '*and.*' and presented the figures for various forms of the term, thus subtraction can be easily done. Out of the complete rank-frequency-table we took the top used terms i.e. those that appeared 100 times or more, as presented in Table 8.

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		

TABLE 8. Listing of Terms Occurring More Than 100 Times. (**** = expletive)

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 – lack 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.* 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. The tail end of unique terms is very long and warrants in itself a linguistic investigation.

We constructed a graph of rank – frequency distribution of all terms, but it is not shown here because of space restrictions. The resulting distribution seem to be unbalanced indeed. The graph does not follow the traditional slope of a Zipf distribution representing the distribution of words in long English texts. At the beginning it falls of very steeply, and toward the end it shows discontinuities and an unusually long tail representing terms with frequency of one. The terms in the language of queries is distributed very differently than the terms in texts or discourse.

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. There is no way of going around it: a lot of terms, about 25% of highest used terms, dealt with some or other sexual topic, but that represents only less than 3% of all terms. Of course, if one classifies some more terms further down the distribution 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 two 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 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.

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. Queries in searching of regular IR systems are some three to seven magnitudes larger. About one in three queries had one term only, two in three had one or two terms, and four in five had one, two or three terms. Less than 4% of the queries were more than 6 terms.
- Relevance feedback was not used that much. About one in 20 queries used the feature *More Like This.* In comparison with professionally-assisted IR searching, relevance feedback is used half as much on the Web.
- Boolean operators were not frequently used. One in 18 users used any Boolean capabilities, and of those users that used them, every second user made a mistake according to *Excite* rules. As to the queries, about one in 12 queries contained a Boolean operator, and in those AND was used by far the most. About one in 190 queries used nested logic. About one in every three queries that used Boolean operators or parentheses was not entered as required by *Excite*. Web searchers are reluctant to use Boolean searches and when using they have great difficulty in getting them right
- The '+' and '-' modifiers that specify a must for presence or absence of a term were used more than Boolean operators. About 1 in 12 users used them. About one in 11 queries incorporayed a '+' or '-' modifier. But a majority of uses were mistakes: about two out of three uses of these operators was incorrect. The ability to create phrases (terms enclosed by quotation marks) was seldom used – about one in 16 queries contained a phrase, but mistakes were negligible.
- Most users searched one query only and did not follow with successive queries. About two in three users had a single query, and 6 in 7 did not go beyond two queries.
- On the average users viewed 2.21 pages. Over half of users did not access result beyond the first page. More than three in four users did not go beyond viewing two pages
- 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 63 subject terms that had a frequency of appearance of 100 or more,

represented only one third of one percent of all terms but they accounted for about one of every 10 terms used in all queries. Terms that appeared only once amounted to a half of unique terms**.** The graph of the rank-frequency distribution does not show a regularity as found in studies of distribution of words in texts.

- 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. 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.

CONCLUSIONS

We investigated a large sample of searches on the Web, represented by logs of queries from *Excite*, a major Web search provider. However, we consider this study just as a beginning. We have begun the analysis of a new sample of over 1 million queries. In a way, we consider this study as a pilot for analysis of a much larger sample.

While Web search engines follow the basic principles of IR, Web search users seem to differ significantly from users of traditional IR systems, such as those represented by users of DIALOG or assumed (and highly artificial) users of 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 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 relevance feedback also desires further investigation. Among others, the question of actual low use of this feature should be addressed in contrast to assumptions about high usefulness of this in IR research. If users use it so little, what is the impetus for testing, such as in TREC, on relevance feedback in the present form? This is one of the examples where users are voting with their fingers, and research is going the other way.
- 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 as well. 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?

ACKNOWLEDGMENTS

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. We also acknowledge the generous support of our institutions for this research.

REFERENCE

Abdulla, G., Fox E.A., & Abrams, M. (1997). Shared User Behavior on the World Wide Web. *Proceedings of the WebNet'97*, 54-59.

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.

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

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.

FIND/SVP (1997) *The 1997 American internet user survey*. http://www.cyberdialogue.com/isg/internet

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

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

Lawrence, S. Giles, C.L. (1998). Searching the World Wide Web. *Science, 280* (5360), 98-100.

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

Kehoe, C., Pitkow, J., & Morton, K. (1997). *GVU's 8th WWW user survey.* Atlanta, GA: Graphic, Visualization, and Usability Center, Gergia Tech Research Center. Http://www.gvu.gatech.edu/user_surveys

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

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.