Use of query reformulation and relevance feedback by Excite users

Amanda Spink
Bernard J. Jansen and
H. Cenk Ozmultu

The authors

Amanda Spink is Associate Professor in the School of Information Sciences and Technology, Pennsylvania State University, University Park, Pennsylvania, USA.
Bernard J. Jansen is Lecturer in the Computer Science Program at University of Maryland (Asian Division), Seoul, Korea.
H. Cenk Ozmultu is a PhD candidate at the Department of Industrial Engineering, Pennsylvania State University, University Park, Pennsylvania, USA.

Keywords

Internet, World Wide Web, Excite, Search engines, Human-computer interaction, Information retrieval

Abstract

Examines the use of query reformulation, and particularly the use of relevance feedback by users of the Excite Web search engine. A total of 985 user search sessions from a data set of 18,113 user search sessions containing 51,473 queries were examined. Includes a qualitative and quantitative analysis of 191 user sessions including more than one query, to examine patterns of user query reformulation; and second, all 804 user sessions including relevance feedback were examined. Results show limited use of query reformulation and relevance feedback by Excite users – only one in five users reformulated queries. Most relevance feedback sessions were successful. Identifies the most common pattern of searching and discusses implications for Web search system design.

Electronic access

The current issue and full text archive of this journal is available at
http://www.emerald-library.com

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · pp. 317–328
 · ISSN 1066-2243

Introduction

We investigated the use of query reformulation techniques by users of the Excite Web search engine, including their use of relevance feedback as a form of query reformulation during Web searching. Previous studies show that users of IR systems, such as online databases, reformulate their queries by adding or subtracting terms during their IR search interactions to differing degrees (Efthimiadis, 1996). Relevance feedback has been a major and active IR research area and reported to be a successful feature of many IR systems (Harman, 1992; Spink and Losee, 1996).

Relevance feedback is now used extensively on the Web to reformulate a query based on Web sites identified by the user as relevant. For example, if a user enters the word "recipes" into the Excite search engine, many Web sites would be retrieved. When judging the relevance of the first ten Web sites on recipes, the user identifies a Web site discussing "cookie recipes" as more relevant to their information need. They click the "More Like This" button next to a Web site about "cookie recipes" and the Excite search engine begins to search for more Web sites inlcuding the words "cookie recipes".

In this paper we report findings from the analysis of a large data set of user queries to the Excite search engine. Our analysis shows that few users reformulate queries or use relevance feedback. This research contributes to the growing body of studies seeking to describe the dimensions of Web searching. Few studies have investigated the use of query reformulation and relevance feedback options by users of Web search engines. The study of users' interaction with Web search engines is an important and emerging area of research with implications for the development of

The authors gratefully acknowledge the assistance of Graham Spencer, Doug Cutting, Amy Smith and Catherine Yip and Jack Xu 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 assistance of Carol Chang and Agnes Goz from the University of North Texas, the generous support of our institutions for this research and the comments of the anonymous reviewers.

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

more effective Web-based human-computer interaction models, search engines and interfaces. Our study builds on a number of related Web studies examined in the next section of the paper.

Related studies

A growing body of studies is investigating many aspects of users' interactions with the Web. Experimental and comparative studies show little overlap in the results retrieved by different search engines based on the same queries (Ding and Marchionini, 1996; Gordan and Pathak, 1999; Lawrence and Giles, 1998). Many differences in search engine features and performance (Chu and Rosenthal, 1996) and users' Web searching behavior (Tomaiuolo and Packer, 1996) have been identified. Studies comparing novice and expert Web searchers show regular patterns in Web users' surfing behavior (Huberman *et al.*, 1998). Many surveys of Web users have been conducted; either library based (Tillotson *et al.*, 1995) or distributed via newsgroups. Spink *et al.*'s (1999) survey of Excite users shows that many users conduct several related searches of the Web on the same topic over time.

In a recent large-scale study, Jansen *et al.* (2000) examined 51,473 Excite queries from 18,113 users, including 113,776 terms. Findings from the study provide a picture of user interaction with a Web search engine. Most users did not use many queries per search, with a mean of 2.8 queries per search. Most users searched with one query only and did not follow with successive queries. Web queries are short, with a mean of 2.21 terms. We conducted an in-depth qualitative analysis of a sub-set of the same Excite data set of 51,473 queries to specifically examine the use and patterns of query reformulation and relevance feedback by Web users. Few studies have examined patterns of users' query reformulation and use of relevance feedback when using Web search engines. Studies of user interaction with Web search engines are important for developing models of user behavior to improve Web system design.

Research questions

Our study addressed the following three research questions:

(1) What are the frequency and patterns of query reformulation by Excite users?
(2) What are the frequency and patterns of relevance feedback use by Excite users?
(3) What are the differences between relevance feedback users and the overall population of Excite users in terms of searching characteristics?

Research design

Excite data corpus

Excite, Inc. is a major Internet media public company that provides free Web searching and a variety of other services. The company and its services are described at its Web site (http://www.excite.com) and not repeated here. The Excite Web search engine capabilities relevant to our study are summarized. Excite searches are based on the exact terms that a user enters in the query, but capitalization is disregarded, with the exception of logical commands AND, OR, and AND NOT. Stemming is not available. Search results are provided in a ranked relevance order.

At the time the data were collected, Excite provided a hypertext link "More Like This", as a relevance feedback mechanism to find similar sites. The option is still available on Excite, but the wording has changed slightly. If the user finds a site that is relevant, the user need only "click" on this hypertext link, implementing the relevance feedback option. When the relevance feedback link is clicked, Excite enters and counts this into a transaction log as a query with zero terms. Use of the feature does not appear to be any more difficult than normal Web navigation. In fact, one could say that the implementation of relevance feedback is one of the simplest IR techniques available on Excite.

The basic statistics of the Excite data set related to queries and search terms are given in Table I.

The Excite transaction log record of 51,473 queries from 18,113 users, contained three fields of data on each query from 9 March 1997. Using these three fields, we located a

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

Table I Numbers of users (sessions), queries, and terms

Number of users (sessions)	18,113
Number of queries	51,473
Number of terms	113,793
Mean queries per user	2.84
Mean terms per query	2.21
Mean number of pages of results viewed	2.35

user's initial query and recreated the chronological series of actions by each user in a session:

- *Time of day*: measured in hours, minutes and seconds from midnight (US time).
- *User identification*: an anonymous user code assigned by the Excite server.
- *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.

- *Session*: a session is the entire series of queries by a user over time. A session could be as short as one query or contain many queries.
- *Query*: a query consists of one or more search terms, and possibly includes logical operators and modifiers.
- *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.

Data analysis

Our study specifically examined the frequency and patterns of query reformulation, and the use of relevance feedback by Excite users. Specific analysis conducted included:

- All 18,113 user sessions were quantitatively analyzed.
- A subset of 191 sessions (1 per cent of all 18,113 user sessions) or 33 per cent of all sessions that included query reformulation (more than one query) were qualitatively analyzed to examine user patterns of query reformulation.
- All 804 user sessions (4 per cent of all 18,113 user sessions) that included relevance feedback were qualitatively analyzed. Of the 51,473 queries, only about 6 per cent (2,543) could have been from Excite's relevance feedback option. This is a surprisingly small percentage of the queries relative to usage in IR systems. From our analysis, we identified queries resulting from the relevance feedback option and isolated the sessions (i.e. sequence of queries by a user over time) that contained relevance feedback queries.
- Working with these user sessions, we qualitatively classified each query in the session by query type. We then identified patterns in these sessions composed of transitions from query type to another query type. By classifying these session patterns we hope to gain insight in the current use of query reformulation and relevance feedback on the Web.

Results

This paper is part of a large project analyzing the Excite data set and extends findings reported in Jansen *et al.* (1999) and Spink *et al.* (1999). In the first section of the paper we examine query reformulation by Excite users.

Query reformulation analysis

We examined the number of queries per user search session in the complete data set.

Queries per search session

Table II shows the distribution of the number of queries per user session.

The mean session length for Excite users was 2.84 queries in length with two in three users entering only one query. One in three users entered multiple queries with query modification, viewing subsequent results, or both.

Changes in number of terms over successive queries

These results of the analysis are displayed in Table III.

We focus on the 11,249 queries in the data set 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. related or

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

Table II Number of queries per user session

Queries per user session	Number of user sessions	Per cent of user sessions
1	12,068	67.3
2	3,501	19
3	1,321	7
4	583	3
5	287	1.7
6	144	0.8
7	79	0.5
8	32	0.22
9	36	0.2
10	17	0.08
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

Table III Changes in number of terms in successive queries

	Number	Percent
Increase in terms		
0	3,909	34.76
1	2,140	19.03
2	1,068	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		
–1	1,837	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

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. Per cent is based on the number of queries in relation to all modified (11,249) queries. Query modification was not a typical occurrence. This finding is contrary to experiences in searching other IR systems, where modification of queries is more common (Spink and Saracevic, 1997). Overall, we found that:

- *A total of 33 per cent of the users did go beyond their first query*. Approximately 14 per cent of users entered three or more queries. These percentages of 33 per cent and 14 per cent are significant 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, as they could represent sub-populations of Web users with more experience or higher motivation who perform query modification on the Web. We can see that users typically do not add or delete much in respect to the number of terms in their successive queries. Query modifications were 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 modified still had the same number of terms as the preceding one*. In the remaining 7,338 successive queries where terms were either added or subtracted, about equal numbers had terms added as subtracted (52 per cent to 48 per cent) – 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*.

Classification of queries

An analysis of 181 user sessions shows that 33 per cent of all user sessions in the data set included two or more queries. Table IV shows

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

Table IV Basic data

Number of user sessions	Number of queries	Number of terms	Mean terms and range
191	1,369	3,015	2.19 0-10

the basic data for the query classification analysis. We qualitatively analyzed all 191 user sessions to examine how successive queries differed from other queries by the same user during the same session. Each query was first classified as:

- Unique query (U): unique query by a user; or
- Modified query (M): 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; or
- Next page (P): when a user views the second and further pages (i.e. a page is a group of ten results) of results from the same query, Excite provides another query, but a query that is identical to the preceding one; or
- Relevance feedback(R): when a user enters a command for relevance feedback (More Like This), the Excite transaction log shows that as a query with zero terms.

Table V shows the number of occurrences of each type of query:

- The mean number of queries per user was 2.19 with a range from two to ten.
- Overall, users performed limited query modification as one in five queries were modified queries.
- One in two queries were requests for the next page of results.
- Less than one in ten queries were relevance feedback.

Table V Number of occurrences of each query type

Query type	Number of queries	Percentage of queries
Unique query (U)	340	25
Modified query (M)	274	20
Next page (P)	642	46
Relevance feedback (R)	123	9
Total	1,379	100

Query patterns during sessions

We next examined how the succession of queries differed amongst users. For example, if a user enters a unique query (U) followed by three modified queries (M) and finally a next page (P), this is represented as the shift pattern UMP. The user shifted from a unique query, to modified queries, to looking at the second ten retrieved Web sites. Table VI shows the number of occurrences of each session shift pattern in the data set.

Overall, our analysis of users' query patterns during sessions shows:

- that the most common user session (one in four) was UP – a unique query followed by a request to view the next page of results with no query modification;
- that one in two users viewed the next page of results before modifying their query;
- that one in two user sessions included query modification;
- that there is not a lot of subject change – 73 per cent of user sessions included one topic and 27 per cent included two topics;

Table VI Query patterns during user sessions

Pattern	Number of user sessions	Percentage of user sessions
UP	52	28
UM	18	10
UPM	8	4.2
UPU	8	4.2
UMP	8	4.2
UPMP	7	3.6
UPR	5	2.6
UR	4	2
UPMPM	2	1
UPMPMP	2	1
UMPMP	2	1
UMUM	2	1
RU	2	1
Other (unique – single occurrence)	69	31
Total	191	100

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

- that one in three user sessions were unique in their query patterns, e.g. only one user entered the pattern UMPRMP – a unique query followed by a modified query, then a next page, a relevance feedback, followed by a modified query and a next page; and
- that relevance feedback was not used extensively during the user sessions examined.

Distributions of the session length

We analyzed the length of the 191 sessions that included more than one query, including the number of queries in user sessions. A user could repeat the same type of query consecutively. For example in one instance, a user repeated Next page (P) 37 consecutive times. The distribution of the user session lengths is shown in Figure 1.

As in most of the human behavior models, exponential distribution seems to be the most logical choice of distribution. The shape of the distribution in Figure 1 supports the exponential distribution option as the distribution of the session lengths. The chi-square goodness of fit test is used to identify the distribution of the session lengths. To be able to apply the chi-square test, the sample is divided into 13 (k = 13) exclusive and exhaustive outcomes, shown in Table VII.

In our sample, with 191 observations (n = 191), for each outcome, group frequency and the expected frequency are calculated, and used to calculate Q_{k-1}, namely,

$$Q_{k-1} = \sum_{i=1}^{k} = \frac{(Y_i - np_{i0})^2}{np_{i0}}$$

The expected frequency of an observation is the sample size times the probability of that observation, given that the sample comes from the distribution against which the sample is tested. In Table VII expected frequencies are calculated using exponential distribution with mean of 6.667 queries. The chi-square goodness of fit test first calculates the normalized differences between the actual (np_i) and the test (np_{i0}) frequency of each exclusive and exhaustive outcome, then checks significance of the sum of the normalized differences. It is argued that the sum of normalized differences (Q_{k-1}) resembles a chi-square distribution with $k-1$ degrees of freedom. Using Q_{k-1} following hypothesis can be tested.

H_0: $p_i = p_{i0} i = 1, 2, \ldots, k.$

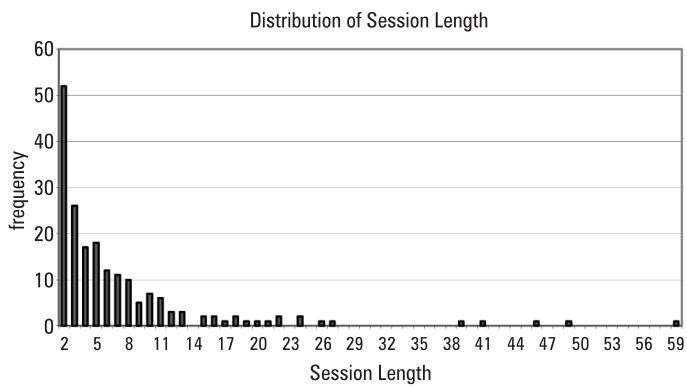
$H1$: $p_i \neq p_{i0}$ for some $j = 1, 2, , k.$

For this chi-square goodness of fit with 5 per cent error, Q_{12} = 6.8566 < 21.83=X^2 0.05(13) (see Table VII). Therefore, $H1$ is strongly rejected as session length is distributed exponentially with a mean of 6.6667 queries. Using this exponential distribution, we can answer questions such as: What is the probability that a user will use ten queries in a session? If we assumed that the duration of each query type is equal, the question given before is equivalent to the question: What is the probability that a user will have ten queries in their session?

Table VII Exponential session length distribution

Observation	Frequency (Y_i)	Expected ($E(Y_i) = np_{i0}$)	$(Yi - np_{i0})^2/np_{i0}$
2	52	49.5	0.126813
3	26	19.71	2.004710
4	17	16.96	0.000072
5	18	14.60	0.787636
6	12	12.57	0.026074
7	11	10.82	0.003160
8	10	9.31	0.052375
9	5	8.01	1.129904
10	7	6.90	0.001270
11	6	5.93	0.000629
12	3	5.11	0.871960
13	3	4.40	0.445091
14 and up	21	27.17	1.406944
Total	191	191	6.856638

Figure 1 Distribution of user session lengths



Markov analysis

The purpose of this section is to analyze user sessions. Questions such as: What is the

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

probability of using a certain query type? For example, if the user modified the query, will the user then use relevance feedback? The probability is 5.3 per cent. The transition matrix (see Table VIII) is obtained by counting the query transitions in the 191 sessions.

The transition matrix in Table VIII provides conditional probabilities for stepwise movements in the duration of sessions. In this analysis we included Previous page (PP). Moreover, using the transition matrix we can obtain limiting probabilities for each query type. However, query type E (end of search) is an absorbing state, therefore, in the long run, limiting probabilities for all of the query types except state E will be zero. Limiting probabilities are useful for examining the ratio of each query type on average in the duration of a session. To obtain the limiting probabilities we used the transition matrix given in Table VIII where query type E is removed from the analysis in Table IX.

The limiting probabilities obtained from the transition matrix given in Table IX are given in Table X. From Table X, we can conclude that almost 55 per cent of user queries are looking at next pages. We know from the previous section that the average session lasts 6.667 queries. Using the limiting probabilities, we can divide the total average queries into average queries during a session (e.g. on average users create 6.667*0.5482 = 3.655 queries looking at the next pages). The reformulation analysis shows that about one in five users reformulated their initial query. However, those users who did reformulate queries created nearly seven queries, many being Next page requests. Relevance feedback represented a small fraction of queries

Table VIII Transition matrix

	From					
To	**M**	**U**	**P**	**PP**	**R**	**E**
M	0.355	0.094	0.302	0.023	0.053	0.174
N	0.252	0.165	0.430	0.003	0.052	0.097
G	0.136	0.032	0.637	0.008	0.035	0.152
P	0.129	0.032	0.387	0.097	0.161	0.194
R	0.120	0.104	0.008	0.144	0.512	0.112
E	0.000	0.000	0.000	0.000	0.000	1.000

Table IX Transition matrix (with query type E removed)

	From				
To	**M**	**U**	**P**	**P**	**R**
M	0.429	0.114	0.365	0.027	0.064
U	0.280	0.183	0.477	0.004	0.057
P	0.160	0.038	0.751	0.009	0.042
PP	0.160	0.040	0.480	0.120	0.200
R	0.135	0.117	0.009	0.162	0.577

Table X The limiting probabilities obtained from the transition matrix in Table IX

Query type	Limiting probabilities (per cent)
π_M **(Modified query – M)**	22.76
π_N **(Unique query – U)**	7.52
π_G **(Next page – P)**	54.82
π_P **(Previous page – PP)**	3.44
π_R **(Relevance feedback – R)**	11.45

Relevance feedback analysis

When a user enters a command for relevance feedback (More Like This), the Excite transaction log counts shows that as a query, but a query with zero terms. From the 51,473 query transaction log, a maximum of 2,543 (5 per cent) queries could have been relevance feedback. In comparison, a study involving IR searches conducted by professional searchers as they interact with users found that some 11 per cent of search terms came from relevance feedback (Spink and 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. More complicated IR techniques, such as Boolean operators and term weighting, are used more frequently by Web users (Jansen *et al.*, 2000). We found it surprising that relevance feedback was so seldom utilized. Relevance feedback warrants further study.

Relevance feedback queries

Given the way that the transaction log recorded user actions, the relevance feedback option was recorded as a null query (i.e. empty and with no terms). However, if a user entered an empty query, the empty query would also be recorded in the same way (i.e. empty and with no terms)

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

– we counted these as mistakes. From the previous analysis, we had found that there were 2,543 queries representing the maximum possible amount of relevance feedback queries. For more detailed analysis we separated the relevance feedback queries from null queries and reviewed the transaction log data to remove all queries that were not relevance feedback. We found the vast majority of null queries were the first query in a session and were obviously not the result of relevance feedback. However, if a determination could not be made, the query was considered as a result of relevance feedback. The results are summarized in Table XI.

From Table XI, one sees that approximately one in three of the possible relevance feedback queries were judged not to be relevance feedback queries, but instead a blank or null first query. This result in itself is very interesting and noteworthy. It implies that users enter null first queries just under 40 per cent of the time. From observational evidence, some novice users "click" on the search button prior to entering terms in the search box, possibly thinking that the button takes them to a screen for searching. Additionally, Peters (1993) shows that many IR systems users enter null first queries.

Classification of relevance feedback sessions

Regardless of the reason for the mistakes, the maximum possible relevance feedback queries were 1,597. These queries resulted from an analysis of all 804 user sessions including relevance feedback, with an average of 1.99 relevance feedback queries per user session. Working with these user sessions, we classified each query in the session as belonging to one of the types listed previously in Table V. We first looked at the number of occurrences of each query type and classified each query within each session. The number of occurrences of each query for all 804 relevance feedback sessions, is shown in Table XII.

Table XI Percentage of relevance feedback queries

Classification	Number of queries	Percentage of queries
Relevance feedback	1,597	63
Null first queries	946	37
Total	2,543	100

Table XII Occurrences of query types

Query type	Number of queries	Percentage of queries
Relevance feedback (R)	1,597	42.3
Next page (P)	693	18.4
Modified query (M)	467	12.3
Unique query (U)	1,020	20.27
Total	3,777	100

There were 3,777 queries in the 804 user sessions including relevance feedback. Relevance feedback was used in by far the most occurrences followed by unique query. There were also a number of next page and modified queries, indicating the addition, removal, or change of query terms.

Query transitions

We examined the occurrence of each query type within each relevance feedback session as opposed to the overall totals (Table IX):

- The shortest session was two queries and the maximum session length was 17 shifts in query type.
- If a query type occurred in succession, we counted it as only occurring once. For example, in a session of Query →Relevance feedback → Relevance feedback, the Relevance feedback query would be counted as only occurring one time within that session. We did this to simplify the pattern and isolate the state transitions from one state to another. However, using the above example, if the Relevance feedback occurred last in the session, following another query type, it would be counted again. For example, Query → Relevance feedback → New query → Relevance feedback. In this example, Relevance feedback would be counted twice.
- Given that there was no one-query session in this sample (i.e. the shortest session was Query → Relevance Feedback, a two query session), there were 239 two-query sessions, the largest group. However, there were 251 three-query sessions, 120 four-query sessions, 82 five query sessions, followed by a fair number of six- and seven-query sessions. After that, there is a sharp drop-off in session length.

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

- For the sessions of two and three queries, relevance feedback is the dominant query type.
- As the length of the sessions increased, the occurrences of relevance feedback as a percentage of all query types decreased.
- Beginning with sessions of five queries or more, relevance feedback is no longer the query type with the most occurrences. The decreased use of relevance feedback requires further investigation.

Success of relevance feedback

Given the low occurrences of relevance feedback queries, we attempted to determine if the session containing relevance feedback was successful or not. However, we can make some generalizations about the use of relevance feedback. If relevance feedback was used and then the user quit searching, we counted it as a success (i.e. assumed the user found something of relevance). This pattern seems consistent with what one may expect of a successful search session. Many times these sessions may not be successful, so this count is on the high end. If the user utilized relevance feedback and returned to the exact previous query, we assumed that nothing of value was found (i.e. an unsuccessful session). This pattern seems consistent with what one may expect of an unsuccessful search session (i.e. assumed the user found nothing of relevance).

There were some sessions when relevance feedback was used and then the user returned to a similar, but not exact, query. Since the relevance feedback query could have provided some terms suggestions, we classified these sessions as partially successful, again giving relevance feedback the benefit of the doubt. The results of this analysis are summarized in Table XIII.

As one can see, giving relevance feedback the benefit of the doubt, fully 63 per cent of the relevance session may be construed as being successful. If the partially successful are included, then over 80 per cent of the relevance feedback sessions provide some measure of success. This is a fairly high percentage; although as mentioned, we are presenting probably the maximum number of successful sessions. The question then becomes, why is relevance feedback not used more on the Web search engine? In order to gain greater insight to this behavior, we compared the population of RF users to the larger population in our data set, to see if there was some difference that set this sub-set apart from the larger population.

Table XIII Classification of relevance feedback sessions

Classification	Number of occurrences	Percentage
Successful	509	63
Unsuccessful	156	20
Partially successful	139	17
Total	804	100

Comparison of relevance feedback users to larger population of Excite users

We examined the query construction of relevance feedback users to the query construction of the general population. The actual numbers from the larger population are unimportant. The important item of comparison is the percentages. The actual numbers are available (Jansen *et al*. 2000). The comparison is displayed in Table XIV.

There generally appears to be little difference between the relevance feedback users and the population in general in number of query terms used, other than zero term queries (e.g. the relevance feedback queries). The mean number of terms per query was 1.98 for the relevance feedback population and 2.2 for the larger population. Assuming that lengthier queries are

Table XIV Terms per query

Terms per query	Number of queries in RF population	Percent of RF queries	Percent in general Excite population
0	872	21	6
1	972	23	31
2	1,045	25	31
3	635	15	18
4	310	7	7
5	195	5	4
6	70	2	1
7	36	1	0.94
8	23	1	0.44
9	3	0	0.24
>10	22	1	0.36
Total	4,183	100	100

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

a sign of a more sophisticated user, it appears that the relevance feedback population does not differ considerably from the larger population of Excite, and possibly, Web users. The number of queries per user for each group is shown in Table XV.

The relevance feedback population had longer sessions than the population at large. The median number of queries per user for the relevance feedback population was two, and for the larger population it was one. There were also a number of relevance feedback users with sessions of three, four, five and, even six queries. In the larger population, there is a steep drop-off at two queries per user. This may indicate that relevance feedback users were more persistent in satisfying their information need and therefore more willing to invest the time and effort to use not only relevance feedback but also longer sessions in general.

Discussion

Our study revealed some interesting findings about Web users' query reformulation and use of relevance feedback options. Results show limited use of query reformulation and relevance feedback by Excite users – only one in five users reformulated queries. Those users who entered more than one query, entered an average of 6.666 queries. Most relevance feedback sessions were successful. The most common pattern of searching was a single query followed by viewing a list of the first ten Web sites. Of the over 51,473 queries in the data set, less than 5 per cent were from Excite's relevance feedback option. This is a small percentage of the queries. We also noted the high number of null queries entered by users. We identified four possible query types: unique query, relevance feedback, modified query, and next page. Of these query types, not counting the unique query state, relevance feedback was the most common, occurring 872 times. There was an average of 1.99 relevance feedback queries per user session.

In the general Excite population, most users entered only a single query. A third of users went beyond the single query, with a smaller group using either query modification or relevance feedback, or viewing more than the first page of results. The shortest user session was two queries and the distribution of query type shifts as the length of the user session increases. For the user sessions of two and three queries, the relevance feedback query is dominant. As the length of the sessions increases, the occurrences of relevance feedback as a percentage of all query types decreases. Given the low occurrences of relevance feedback queries, some 63 per cent of the relevance feedback sessions could be construed as being successful. If the partially successful user sessions are included, then more than 80 per cent of the relevance feedback sessions provide some measure of success.

There appears to be little difference between the relevance feedback users and the population in general. Both populations had mean query lengths of about two terms. Next, we examined the number of queries per user. The relevance feedback population had significantly longer queries than the population at large. The median number of queries per user for the relevance feedback population was two, and for the general population it was one.

IR studies suggest that relevance feedback is useful for Web users, although in the data set we examined only a small percentage of Web users take advantage of this feature. On the other hand, although it is successful 63 per cent of the time, this implies a 37 per cent failure rate or at least a not totally successful rate of 37 per cent. This may be one reason why relevance

Table XV Queries per session

Query per user	Number of users	Percentage of RF users	Percentage of general population
1	3	0.36	67
2	375	45	19
3	223	27	7
4	97	12	3
5	64	8	2
6	34	4	0.80
7	11	1	0.44
8	4	0.48	0.18
9	8	0.97	0.20
10	6	0.72	0.09
11	1	0.12	0.04
>12	1	0.12	0.04
	827	100	100

Internet Research: Electronic Networking Applications and Policy
Volume 10 · Number 4 · 2000 · 317–328

feedback is so seldom utilized. Its success rate on the Web is just too low. It points to the need for an extremely high success rate before Web users consider it beneficial. As for characteristics of the relevance feedback population, they do not to differ in terms of query construction, but they exhibit more persistence in attempting to locate relevant information. This is manifested in longer sessions. This could be for several reasons; perhaps the subjects they are searching for are more intellectually demanding. Unfortunately, a cursory analysis of the query subject matter and terms does not support this conclusion. It probably points to some detail of sophistication in searching techniques. So, relevance feedback options appear to attract a more sophisticated Web user.

If these results can be generalized to other Web search engines other than Excite, they point to the need to tailor the interface if the goal is to increase the use of relevance feedback. Also, the precision of the relevance feedback option must be increased.

Limitations

The methodology of this study has both strengths and weaknesses. Strengths of the study include the analysis of a large data set from users of a commercial Web search engine. These results provide large-scale findings often difficult to obtain under laboratory conditions. Weaknesses of the study include the limitation to one Web search engine without comparison to users of other Web search engines or meta-search tool users.

Implications for Web design

Our findings emphasize the need to approach design of Web IR systems, search engines, and even Web site design in a significantly different way than the design of IR systems as practiced to date. They also point to the need for further and in-depth study of Web users. For instance:

- The low use of 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 a 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 deserves 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 feature in IR research. If users use it so little, what is the impetus for testing relevance feedback in the present form? Users are voting with their fingers, but is research going the other way?
- Given that relevance feedback may be useful for Web users, methods could be investigated to automate the implementation of relevance feedback. Instead of offering the relevance feedback mechanism to the user, the system could automatically implement relevance feedback and offer the resulting documents to the user.

Conclusion and further research

Our analysis points to the need for more detailed analysis of the query reformulation and relevance feedback in the context of Web searching. We are currently conducting an analysis of a 2.5 million query set supplied by Excite, Inc. The results of the new analysis will be compared the findings from our previous studies.

References

Chu, H. and Rosenthal, M. (1996), "Search engines for the World Wide Web: a comparative study and evaluation methodology", *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, Baltimore, MD, October pp. 127-35.

Ding, W. and Marchionini, G. (1996), "A comparative study of Web search service performance", *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, Baltimore, MD, October, pp. 136-42.

Efthimiadis, E. (1996), "Query expansion", *Annual Review of Information Science and Technology*, Vol. 31, pp. 121-88.

Gordon, M. and Pathak, P. (1999), "Finding information on the World Wide Web: The retrieval effectievness of search engines", *Information Processing and Management,* Vol. 35, pp. 141-80.

Harman, D.K. (1992), "Relevance feedback revisited", in Belkin, N.J., Ingwersen, P. and Pejtersen, A.M. (Eds), *SIGIR 92: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 15th Annual International Conference of Research and Development in Information Retrieval*, 21-24 June, pp. 1-10.

Huberman, B.A., Pirolli, P., Pitkow, J.E. and Lukose, R.M. (1998), "Strong regularities in World Wide Web surfing", *Science*, Vol. 280 No. 5360, pp. 95-7.

Jansen, B.J., Spink, A. and Saracevic, T. (1999), "The use of relevance feedback on the Web", *Proceedings of WebNet 99*, Hawaii, October.

Lawrence, and Giles, C.L. (1998), "Searching the World Wide Web", *Science*, Vol. 280 No. 5360, pp. 98-100.

Jansen, B.J., Spink, A. and Saracevic, T. (2000), "Real life and real users: a study and analysis of user queries on the Web", *Information Processing and Management*, Vol. 36 No. 2, pp. 207-27.

Peters, T.A. (1993), "The history and development of transaction log analysis", *Library Hi Tech*, Vol. 4 No. 2, pp. 41-66.

Spink, A., Bateman, J. and Jansen, B.J. (1999), "Searching the Web: survey of Excite users", *Internet Research: Electronic Networking Applications and Policy*, Vol. 9 No. 2, pp. 117-28.

Spink, A. and Losee, R.M. (1996), "Feedback in information retrieval", in Williams, M. (Ed.), *Annual Review of Information Science and Technology, Vol. 31*, pp. 33-78.

Spink, A. and Saracevic, T. (1997), "Interaction in information retrieval: selection and effectiveness of search terms", *Journal of the American Society for Information Science*, Vol. 48 No. 8, pp. 728-40.

Spink, A., Jansen, B.J., Chang, C. and Goz, A. (1999), "Users' interactions with the Excite Web search engine: a query reformulation and relevance feedback analysis", *Proceedings of the 1999 Canadian Association for Information Science (CAIS) Conference*, Sherbrooke, Canada, June, pp. 342-54.

Tillotson, J., Cherry, J. and Clinton, M. (1995), "Internet use through the University of Toronto: demographics, destinations and users' reactions", *Information Technology and Libraries*, September, pp. 190-8.

Tomaiuolo, N.G. and Packer, J.G. (1996), "An analysis of Internet search engines: assessment of over 200 search queries", *Computers in Libraries*, Vol. 16 No. 6, pp. 58-62.