The Use of Relevance Feedback on the Web: Implications for Web IR System Design

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
School of Library and Information Sciences, University of North Texas
P.O. Box 311068, Denton TX 75203 USA
spink@lis.admin.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. & Saracevic, T. 1999. The use of relevance feedback on the web: Implications for web IR system design. 1999 World Conference on the WWW and Internet, Honolulu, Hawaii.

See Other Publications

Abstract: We conducted a transaction log analysis of 51,473 queries from 18,113 users of Excite, a major Web search engine. Approximately 2,500 (approximately 5%) of these queries were from the use of relevance feedback. Given the high level of research activity and historical success of relevance feedback in assisting users in locating relevant information, this is a surprising small percentage of usage. In order to investigate this phenomenon on the Web, we analyzed users sessions that contained relevance feedback queries. We identified states and patterns in these sessions. We also attempt to classify the sessions as successfully or not. This analysis provides insight on the current use of relevance feedback on the Web, its success or failure, and why it is so seldom utilized. We relate our conclusions to system design of information retrieval systems on the Web.

Introduction

Transaction log analysis is a proven analytical technique in information science that can provide excellent data on user searching characteristics (Peters, 1993). In order to gain insight into Web users and their use of advanced searching techniques, we conducted a transaction log analysis of 51,473 queries from 18,113 users of Excite, a major Web search engine. The analysis focused on two levels of investigation, the session level and the query level.

From our analysis, we were able to identify the queries that were a result of a relevance feedback option and isolate the sessions (i.e., sequence of queries by a user over time) that contained relevance feedback queries. Of the over 50,000 queries only about 5% were from Excite's relevance feedback option. This is a surprisingly small percentage of the queries compared to traditional information retrieval (IR) system usage.

Relevance feedback is a classic information retrieval (IR) technique that reformulates a query based on documents identified by the user as relevant (Salton, 1983). Relevance feedback has been and still is a major and active IR research area. Relevance feedback is widely used and reported to be extremely successful in many traditional information retrieval systems. However, why is it not widely used on Web search engines? Is it too difficult for users?

When using the Excite search engine (http://www.excite.com), if one finds a documents that is relevant, the user need only "click" on a hyperlink that implements the relevant feedback option. It 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. There are more complicated IR techniques that are used more frequently, such as Boolean operators and term weighting. We found it surprising that this highly touted and widely researched IR feature implemented in straight forward fashion was so seldom utilized.

We analyzed the sessions that contained the approximately 2,500 relevance feedback queries to isolate the user characteristics. We identified patterns in these sessions. These patterns are composed of states and transitions from and to the same or other states. From these characteristics, we hope to gain insight into the possible causes of this the low use of relevance feedback and, possibly, methods to increase its use among Web users. These methods could be applied to design of IR systems on the Web. This paper extends finding finds from (Jansen, Spink, Bateman, & Saracevic, 1998).

Review of Literature

Relevance feedback is a well-known IR technique (Salton, 1983) to improve the performance of IR systems. It has been widely researched (Salton & Buckley, 1990), (Harman, 1992), (Koenemann, 1996), and (Dunlop,1997). It has been reported to successful improve retrieval performance for at least a small number of iterations (Witten, Moffat, Bell, 1994). Although previous studies have focused on a variety of IR systems, we could locate no study that analyzed the use of relevance feedback on a major Web search engine such as Excite.

Background on Excite

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. The search capabilities of Excite are briefly 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 a ranked relevance order. A number of advanced search features are available. 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. 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 two levels of analysis, sessions and queries, we defined our variables in the following way.

1. **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.

2. **Query**: A query consists of one or more search terms, and possibly includes logical operators and modifiers.

Overall Statistics

Given the way that the transaction log recorded user actions, relevance feedback option was recorded as a empty query. However, a user entering a empty query would also be recorded the same. Using a purely quantitative analysis, we isolated 2,543 null queries, which represents the maximum number of relevance feedback queries. For this study, we had to remove the relevance feedback queries from the mistakes. Therefore, we reviewed the data and removed all queries that were obviously not the result of relevance feedback. If a determination could not be made, the query remained in the study. The results are summarized in Table 1:

Classification	Number of Queries	Percentage
Relevance Feedback	1597	63%
Mistakes	946	37%
Total	2543	100%

Table 1: Percentage of Relevance Feedback Queries.

As one can see, fully 37% of the possible null queries were judged not to be relevance feedback queries but instead some sort of mistake. This result in itself is very interesting and noteworthy for Web IR system designers. The high level of failures implies that something with the interface or the system is causing users to enter null queries just under 40% of the time. From observational evidence, some novice users "click" on the search button thinking that it takes them to the screen for searching. Additionally, Peters (1993) states that users many times enter null queries. Regardless of the reason for the mistakes, the maximum possible relevance feedback queries was 1,597. These queries resulted from 823 user sessions, implying an average of 1.99 relevance feedback queries per user session.

We then wanted to isolate patterns, if any, in the user sessions. Working with the 823 user sessions, we classified each query in the session as belonging to one of the following states:

- **Initial Query** was the first string of terms that a user entered for a session.
- **Modified Query** was the second or subsequent entry (i.e., query) that was related to the query before it. Related being defined as processing one or more of the same terms or obviously related to the same topic as the preceding query.
- **Next Page** was a request by the user to view the next page of 10 results.
- **New Query** was a second or subsequent entry by a user that was unrelated to the previous query.
- **Relevance Feedback** was the utilization by the user of the relevance feedback option, "More Like This."
- **Previous Query** was the second or subsequent entry by a user that was exactly like the previous entry.

We first analyzed the number of occurrences of each state.

State Analysis

The number of occurrences of each state is listed in Table 2. There were 2148 unique states in the 804 user sessions. As to be expected, relevance feedback occurred by far (872). Ignoring initial query, the next most

common state was next page (542). This indicates that that there was a number of viewing of subsequent results by users. There were also a large number of modified queries, indicating the addition, removal, or change of query terms.

State	Number of Occurrences
Relevance Feedback	872
Initial Query	804
Next Page	542
Modified Query	467
Previous Query	151
New Query	116
Total	2952

Table 2: Occurrences of Non-Repeating States.

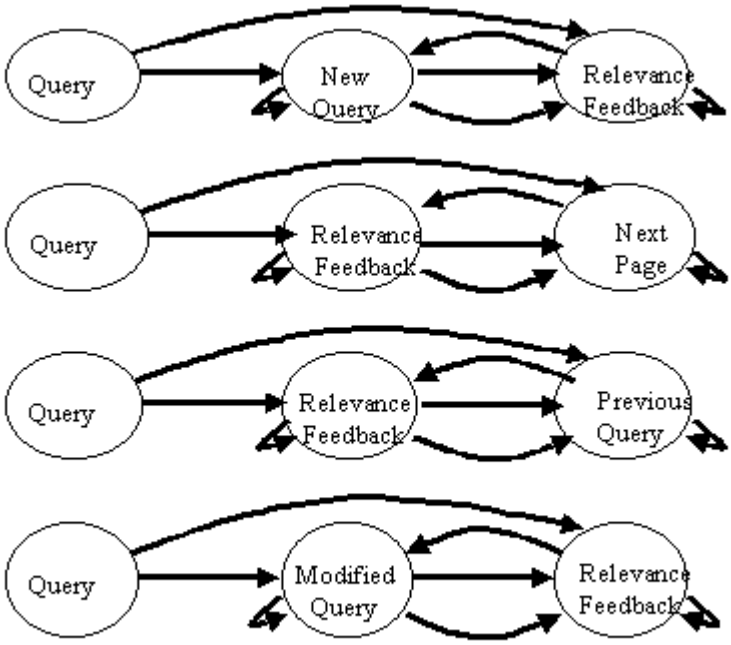
We then examined where each state occurred in the session. The shortest session was two queries. The longest session was seventeen queries. These results are displayed in Table 3.

Type	Query 1	Query 2	Query 3	Query 4	Query 5	Query 6	Query 7	Query 8	Query 9	Total
Initial Query	804	0	0	0	0	0	0	0	0	804
Relevance Feedback	0	371	284	93	63	26	15	8	7	
Next Page	0	282	63	66	56	26	19	14	5	
Modified Query	0	132	133	82	38	35	21	12	6	
New Query	0	19	31	25	14	14	4	3	4	
Previous Query	0	0	54	48	23	11	8	4	1	
Total	804									
Type	**Query 10**	**Query 11**	**Query 12**	**Query 13**	**Query 14**	**Query 15**	**Query 16**	**Query 17**		**Total**
Initial Query	0	0	0	0	0	0	0	0		
Relevance Feedback	2	2	1	0	0	0	0	0		5
Next Page	3	3	0	3	0	1	0	1		11
Modified Query	3	1	3	0	0	0	1	0		8

New Query	2	0	0	0	0	0	0	0		2
Previous Query	1	0	0	0	1	0	0	0		2
Total	11	6	4	3	1	1	1	1		

Table 3: Frequency of State Occurrence at each Session Level.

Given that there were no one query session in this sample (i.e., the shortest session was Query -> Relevance Feedback, a two query session), we see from Table 3, that there were 239 two query sessions, the largest group. However, there were 251 three query session, 120 four query session, 82 five query sessions, followed by a fair number of six and seven query sessions.



We see that the distribution of state occurrences shift as the length of the session increase. For the sessions of two and three queries, the relevance feedback state is the dominant state. As the length of the queries increase, the occurrences of relevance feedback as a percentage of all states decrease. Beginning with session of five queries or more, relevance is no longer the state with the most occurrences. This would seem to indicate that relevance feedback was not successful for these users, and they resorted to other means to find relevant information. This is evidence by the predominance of the modified query state in these lengthier sessions.

State – Transition Analysis

Based on this analysis, we could examine the transitions between states in each session. We isolated four patterns that classified all relevance feedback session. These patterns are displayed in Figure 1. Noted that we account for the returning to the same state. With the identification of these four states, it is clear that the

IR system interface should be tailored to support these patterns of occurrence, namely in the transitions from one state to a different state.

Session Analysis

Given the low occurrences of relevance feedback queries, we attempted to determine if the session containing relevance feedback was successful or not. Without access to the users, this was difficult and required some assumptions. If the user utilized relevance feedback and quit, we gave relevance feedback the benefit of the doubt and counted it as a success (i.e., the user found something of relevance). Probably, many times these were not successfully, so our count of relevance feedback successes if probably on the high side.

If the user utilized relevance feedback and returned to the exact previous query, it is safe to assume that nothing of value was found. There were some sessions where the user used relevance feedback and returned to a similar but not exact query.

Since once could say that the relevance feedback query could have provide some terms suggestions, we classified these session as less than successful.

Some sessions, could also be classified as browsing, namely the user uses relevance feedback and then returns to the session with a totally new query.

The results are summarized in Table 4.

Classification	Number of Occurrences	Percentage
Success	509	63%
Failure	126	16%
Less Than Successful	135	17%
Browsing	34	4%
Total		

Table 4: Classifcation of Relevance Feedback Sessions.

As one can see in Table 4, giving relevance feedback the benefit of the doubt, fully 63% of the relevance session could be construed as being successful. If the less than successful are included, then almost 80% of the relevance feedback session provide some measure of success.

The question then becomes, why is relevance feedback used more on the Web search engine? In order to hopefully gain insight to this, we wanted to see if the population that used relevance feedback different from the population at large.

Comparison to Population at Large

We first examined the query construction of relevance feedback users to the query construction of the general population. This is shown in Table 5. The total percentage for each percentage column does not sum to 100% because the relevance feedback queries are not included. There appears to be little different between the relevance feedback users and the population in general. Assuming that lengthier queries are a

sign of a more sophisticated user, it appears that the relevance feedback population does not difference significantly from the general population of Web users.

Terms Per Query	Number in Relevance Feedback Population	Percent of Relevance Feedback Queries	Percent in General Population
1	972	19.80%	31
2	1045	21.29%	31
3	635	12.94%	18
4	310	6.32%	7
5	195	3.97%	4
6	70	1.43%	1
7	36	0.73%	0.94
8	23	0.47%	0.44
9	3	0.06%	0.24
> 10	22	0.45%	0.36
Total	4908	%	

Table 5: Terms Per Query.

Next, we examined the number of queries per user. This data is displayed in Table 6. In 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 approximately 2 and for the general population it was 1. There were also a significant number of relevance feedback users that had session of 3, 4, 5, and even 6 queries. In the general population, there is a steep drop-off at 2 queries per user. This may indicate that relevance feedback were more persistent in satisfying their information need and therefore more willing to invest the time to use not only relevance feedback but more larger sessions in general.

Query Per User	Number of Users	Percentage of RF Users	Percentage of General Population
1	3	0.36%	67.00
2	375	45.29%	19.00
3	223	26.93%	7.00
4	97	11.71%	3.00
5	64	7.73%	1.60
6	34	4.11%	0.80
7	11	1.33%	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

Summary

We conducted a transaction log analysis of 51,473 queries from 18,113 users of Excite, a major Web search engine. Of the over 50,000 queries only about 5% were from Excite's relevance feedback option. This is an extremely small percentage of the queries. In order to gain insight into the possible causes of this phenomena, we analyzed the sessions that contained the approximately 2,500 relevance feedback queries.

Given the way that the transaction log recorded user actions, relevance feedback option was recorded as a empty query. Fully 37% of the possible relevance feedback queries were judged not to be relevance feedback queries but instead some sort of mistake

We isolated states within each session, identifying 6 possible states, query, relevance feedback, modified query, previous query, next page, and new query. Of these state, relevance feedback was the most common, occurring 872 times.

We then examined the occurrence of each state at each query in the session. The shortest session was two queries. We sew that the distribution of state occurrences shifts as the length of the session increase. For the sessions of two and three queries, the relevance feedback state is the dominant state. As the length of the queries increase, the occurrences of relevance feedback as a percentage of all states decrease.

Based on this analysis, we isolated four patterns that classified all relevance feedback session. These patterns are displayed in Figure 1. Noted that we account for the returning to the same state.

Given the low occurrences of relevance feedback queries, we attempted to determine if the session containing relevance feedback was successful or not. As one can see, given relevance feedback the benefit of the doubt, fully 63% of the relevance session could be construed as being successful. If the less than successful are included, then almost 80% of the relevance feedback session provide some measure of success.

We then compared the relevance population to the population at large. We first examined the query construction of relevance feedback users to the query construction of the general population There appears to be little different between the relevance feedback users and the population in general.

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 about 2 and for the general population it was approximately1.

Conclusion

The data and analysis suggest that relevance feedback is successful for Web users, although only a small percentage of Web users take advantage of this feature. On the other hand, although it is successful over 60% of the time, this implies a 40% failure rate or at least a not totally successful rate of 40%. This may be one reason relevance feedback is so seldom utilized. Its success rate on the Web is just too low.

As for user characteristics of the relevance feedback population, they do not appear to differ in terms of sophistication from the other Web users, but they exhibit more doggedness in attempting to locate relevance information. This could be for several reasons. One may suspect that the subjects they are searching for are more intellectually demanding. A cursory analysis of the query subject matter and terms does not support this conclusion.

There does appear to be four distinct sessions patterns of relevance feedback users on the Web. If these can be generalize to other Web search engines other than Excite, remains to be seen. However, at the very least it points to the need to tailor the interface to support these patterns if the goal is to increase the use of relevance feedback. Another option may be to automate the search engine to retrieve relevant documents of any result the user examinees. This approach would be similar to research by Lieberman (1998). The results could then be presented to the user without the user initiating the process.

References

Dunlop, M. (1997). The effect of accessing nonmatching documents on relevance feedback. ACM Transaction on Information Systems, 15(2), 137 – 153.

Harman, D. (1992). Relevance feedback revisited. Paper presented at *14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Chicago, IL.

Jansen, B., Spink, A., Bateman, J. & Saracevic, T. (1998) Real life information retrieval: A study of user queries on the Web. *SIGIR Forum, 32*(1), 5-17.

Koenemann, J. (1996). Supporting interactive information retrieval through relevance feedback. *Proceedings of the CHI'96 Conference*, 49-50.

Lieberman, H. (1998). Integrating user interface agents with conventional applications. Paper presented at *International Conference on Intelligent User Interfaces*, San Francisco, CA.

Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech, 42*(11), 41-66.

Salton, G. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*. *41*(4), 288-297.

Witten, I., Moffat, A., & Bell, T. (1994). *Managing Gigabytes: Compressing and Indexing Documents and Images*. New York: Van Nostrand Reinhold.