Information Searching Tactics of Web Searchers

Mimi Zhang, Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
311B Information Sciences and Technology Building
University Park PA 16802
E-mail: mzhang, jjansen@ist.psu.edu

Amanda Spink
Faculty of Information Technology
Queensland University of Technology
Gardens Point Campus, GPO Box 2434
Brisbane QLD 4000, Australia
Email: ah.spink@qut.edu.au

Abstract

This paper examines patterns and features of query reformulation within a Web searching session. We pursued this study in response to the growing interest in the area of interactions during information searching. In this study, we randomly selected a stratified sample of Web sessions containing 8,030 queries from an AltaVista (www.altavista.com) transaction log. Then, we analyzed these sessions for query reformulation tactics that the searcher employed. Our results show that changing the query topic was the primary means to modify queries; and most of the time the users were inclined to modify nouns or subtract some types of words when changes were made. The searchers appear to know how to increase and decrease the coverage (i.e., number of results retrieved) of queries. We believe our study can benefit researchers in terms of understanding people's behavior when interacting with Web search engines. It also could benefit search engine providers in terms of improving their services.

Introduction

The Internet is many people's primary information resource nowadays. Due to the large amount of information and its wide distribution, people need to use search engines to aid them in locating information. People's retrieval of online information is an exploratory process (Stojanovic, 2005, p. 281) and "an interactive and iterative process" (Rieh & Xie, 2006, p.752). Thus, when people use search engines to seek information people first form their queries to conduct the initial search on search engines. Based on the searching results and their domain and system knowledge, they reformulate the queries until they come up with satisfactory results. The whole process is defined as query reformulation or query modification. It can be viewed as an "interactive do while loop" by computer scientists or the interaction between users and search engines by human-computer interaction (HCI) researchers. Query reformulation is a crucial step in Web searching in fulfilling users' information needs. This paper reports results from examination of query reformulation and

reporst our findings from the analysis of the transaction logs from AltaVista (www.altavista.com).

Related Research

As a growing research area, Web searching behavior has involved studies by many researchers from different aspects. Two earlier relevant studies include Bates' study (1979) on information search tactics and Fidel's research (1985) on change moves. Bates (1979) defined information search tactics as "a move made to further a search" and named twenty-nine information search tactics according to her knowledge, literatures and comments from other people. She classified these tactics into four categories: monitoring tactics, file structure tactics, search formulation tactics, and terms tactics. Monitoring tactics refer to some as a means to guarantee that the search on the right path and effective. Files structure tactics are employed to get "desired file, source, or information" (Bates, 1979, p.207) via "the file structure of the information facility" (Bates, 1979, p.207). Search formulation tactics mean techniques assisting people in reformulating queries. Terms tactics refer to ways of choosing and modifying terms during search formulation process. Fidel (1985) sorted query reformulation into two categories: operational moves and conceptual moves. Operational moves refers to query modifications that employ the same meaning of query components. In contrast, conceptual moves means to change the meaning of query components.

Some recent relevant research include a series of studies on Excite searchers (c.f., Jansen et al, 1998; Jansen, Spink, & Saracevic, 2000; Spink, Jansen & Ozmultu, 2000; Spink et al, 2001). These users tended to use short queries, not to modify queries, to view only a few results, and not to use advanced search functions. When modifying queries, the users are inclined to use the same amount of key words as the previous query or change only one term. The searchers used a small proportion of terms rather than a wide variety and use some special languages to describe their queries. They usually used search engines to find recreation and entertainment information.

As for query reformulation of Excite users, Spink, Jansen and Ozmultu (2000) report that 32.70% of Excite users did not submit more than one query. When they did modify a query; 34.76% modified queries included the same amount of terms as the original ones. 19.03% of users reformulated their queries by adding a term, and 16.33% of them modified their queries by removing a term. Rieh and Xie (2006) studied multiple Web query reformulations. They concentrated their studies on the reformulation sequences. They analyzed 313 search sessions and classified query reformulation fashions into three categories: content, format, and resource, and eight sub-categories: specified, generalized, parallel, building-block, dynamic, multitasking, recurrent, and format. They also developed a Web query reformulation model based on Saracevic's stratified model and their findings.

In this paper, we continue this line of research by examine a significant number of query reformulation sessions by Web searchers. We focus on the tactical methods of query structure as searchers move on one query to the next within a session, which is named as "search formulation tactics" in Bates' study (1979).

Research Questions

Our goal in this research is to exam how Web searchers reformulate their queries. To achieve this goal, we address two research questions:

- How do people modify their queries during Web sessions?
- What are the results of people's effort of this query reformulation?

This is critical and timely research given the increased focus on Web personalization and the desire to provide more supportive Web systems.

Research Design

To address our research questions, we qualitatively analyzed actual sessions submitted to AltaVista in 2002. The queries examined for this study were submitted to AltaVista over a 24-hour period on Sunday, September 8, 2002. A complete account of the transaction log analysis procedure is outlined in (Jansen, Spink, & Pederson, 2005).

In the log file, each record contains three fields:

1. Time of day: measured in hours, minutes, and seconds from midnight of each day as recorded by the AltaVista server
2. User identification: an anonymous user code assigned by the AltaVista server
3. Query terms: terms exactly as entered by the given user

Using the three fields (time of day, user identification, and query terms), we located the initial query and then recreated the chronological series of actions in a session.

• A *term* is any series of characters separated by white space or other separator.
• A *query* is the entire string of terms submitted by a searcher in a given instance.
• A *session* is the entire series of queries submitted by a user during one interaction with the Web search engine.
• An *initial query* is the first query submitted in a session.
• An *identical query* is a query within a session that is a copy of a previous query within that session.

We qualitatively analyzed two samples of sessions from this transaction log database. We first selected a stratified sample of sessions. That is, the proportion of each session lengths in our sample was equal to the proportion within the entire data set. There were 8,030 queries in this sample from 490 users.

We also randomly selected a 530 query set of just sessions that were two queries from 265 users. Given that the majority of Web sessions are two queries (Jansen, Spink, & Pederson, 2005), it would seem that this segment of the Web population deserves special consideration.

We classified queries according to the codes in Appendix A. The code list is based on the linguistic, information retrieval behavior, information seeking behavior knowledge, and the observation of data.

Results

Query Reformulation Analysis

We first address research question one (*How do people modify their queries during Web sessions?*).

Table 1 presents the results of the pattern classification analysis. A complete legend of codes is presented in Appendix A.

Table 1. Query Classification by Pattern

Pattern	Occurrences	Percentage	Percentage (excluding TC and INT)
TC	2982	37.14%	
INT	1615	20.11%	
SN	557	6.94%	16.22%
CT	462	5.75%	13.46%
SPM	451	5.62%	13.14%
AAN	442	5.50%	12.88%
SP	349	4.35%	10.17%
AAP	247	3.08%	7.19%
SA	194	2.42%	5.65%
ABA	80	1.00%	2.33%
CS	74	0.92%	2.16%
ABN	63	0.78%	1.84%
OTHER	514	6.40%	14.97%
TOTAL (excluding TC and INT)	3,433	42.75%	100.00%
TOTAL	8,030	100.00%	

TC (Topic change) is the most frequent way to modify queries. Except for TC, people often use SN (subtracting noun), CT (change to related term), SPM (spelling change), AAN (noun after term), SP (subtracting phrase), AAP (phrase after term), and SA (subtracting adjective) to reformulate the former queries. Two (SN, AAN) of the seven methods indicate that 29.10% modification fashions are related to changes of noun. Three (SN, SP, SA) of them show that 32.04% modifications are relevant to subtracting something, noun, phrase, or adjective.

Table 2, Table 3, Table 4, and Table 5 present patterns of sessions for query reformulations having different lengths (i.e. number of queries).

Table 2. Pattern of Session with One Query Reformulation

Pattern	Occurrences	Percentage
INT-TC	128	48.67%
INT-AAN	24	9.13%
INT-CT	21	7.98%
INT-SPM	16	6.08%
INT-AAP	15	5.70%
INT-SN	13	4.94%
INT-SP	12	4.56%
INT-ABA	5	1.90%
INT-QEQ	5	1.90%
INT-CS	5	1.90%
OTHER	19	7.22%
TOTAL	263	100.00%

Table 3. Pattern of Session with Two Query Reformulations

Pattern	Occurrences	Percentage
INT-TC-TC	254	28.13%
INT-TC-AAN	40	4.43%
INT-SPM-TC	34	3.77%
INT-TC-SN	34	3.77%
INT-TC-SPM	29	3.21%
INT-TC-AAP	26	2.88%
INT-CT-TC	23	2.55%
INT-AAN-TC	21	2.33%
INT-TC-CT	21	2.33%
INT-AAP-TC	18	1.99%
OTHER	403	44.63%
TOTAL	903	100.00%

Table 4. Pattern of Session with Three Query Reformulations

Pattern	Occurrences	Percentage
INT-TC-TC-TC	112	15.66%
INT-TC-TC-AAN	17	2.38%
INT-TC-AAN-TC	15	2.10%
INT-TC-SN-TC	13	1.82%
INT-SPM-TC-TC	12	1.68%
INT-TC-SPM-TC	12	1.68%
INT-TC-TC-SN	10	1.40%
INT-TC-TC-SPM	10	1.40%
OTHER	514	71.89%
TOTAL	715	100.00%

Table 5. Pattern of Session with Four Query Reformulations

Pattern	Occurrences	Percentage
INT-TC-TC-TC-TC	49	8.66%
INT-TC-TC-TC-SN	10	1.77%
INT-TC-SPM-TC-TC	7	1.24%
INT-TC-SN-TC-TC	6	1.06%
INT-AAN-TC-TC-TC	5	0.88%
INT-CT-TC-TC-TC	5	0.88%
INT-TC-AAN-TC-TC	5	0.88%
INT-TC-SN-SN-TC	5	0.88%
INT-TC-TC-AAN-TC	5	0.88%
INT-TC-TC-SPM-TC	5	0.88%
INT-TC-TC-TC-AAN	5	0.88%
OTHER	459	81.10%
TOTAL	566	100.00%

TC (topic change) ranks as the most frequent means to reformulate the queries, as shown in Table 2. In the other three tables it also occurs in all the top rank patterns. All the most popular patterns only include INT and TC. This data depicts that users usually do not stick to one topic, and they possible seem to get their desired information during the first queries on that topic.

We then choose sessions of two queries from the data set to conduct further analysis since most of Web sessions include two queries (Jansen, Spink, & Pederson, 2005). This segment of the Web population seems deserve special consideration. Table 6 presents the classification of the queries of this population.

Table 6. Query Classification by Pattern

Pattern	Occurrences	Percentage	Percentage (excluding INT)
INT	254	50.00%	
TC	206	40.55%	81.10%
AAN	9	1.77%	3.54%
CT	8	1.57%	3.15%
AAP	5	0.98%	1.97%
SPM	5	0.98%	1.97%
OTHER	21	4.13%	8.27%
TOTAL	508	100.00%	100.00%

TC (topic change) is again the dominate position of modifying queries for two queries sessions, which is similar to our previous findings. AAN (noun after term) and CT (change to related term) are the other two reformulation fashions users use often.

Effect of Query Reformulation Analysis

We next address research question two (*What are the results of people's effort of this query reformulation?*). We continue to focus on the sessions with two queries for this research question. We present the effect of the query reformulation (i.e., is the aim to increase or decrease the number of results) and the means to make these reformulations in Table 7, Table 8 and Table 9. We focus on the effect of the query modification in increasing, decreasing, or no change to the number of Web results one would expect the query to retrieve. The legend of codes is presented in Appendix A.

Table 7. Classification of Effect of Successive Query

Effect	Count	Percentage
TC	207	81.50%
DEC	31	12.20%
INC	15	5.91%
NOC	1	0.39%
TOTAL	254	100.00%

Table 8. Patterns of Query Used to Decrease Range of Query

Pattern	Occurrences	Percentage
AAN	10	32.26%
AAP	6	19.35%
CT	6	19.35%
QEQ	3	9.68%
ABN	2	6.45%
ABP	2	6.45%
CU	2	6.45%
SL	2	6.45%
AAA	1	3.23%
ABD	1	3.23%
AP	1	3.23%
CP	1	3.23%
PBT	1	3.23%
SN	1	3.23%
SF	1	3.23%
DEC	31	

Table 9. Patterns of Query Used to Increase Range of Query

Pattern	Occurrences	Percentage
SPM	5	33.33%
SL	3	20.00%
CT	2	13.33%
SA	2	13.33%
SF	2	13.33%

Pattern	Occurrences	Percentage
SN	2	13.33%
ABA	1	6.67%
CG	1	6.67%
CLT	1	6.67%
PBT	1	6.67%
SM	1	6.67%
SP	1	6.67%
UQEQ	1	6.67%
INC	15	

Web searchers appear to prefer “decrease” to “increase” (refer to Table 7) the number of results. In the age of information explosion, people worry more about how to locate their desired information. Thus, “decrease” is more popular than “increase”. In order to decrease the range or number of query results, AAN (noun after term), AAP (phrase after term), and CT (change to related term) are three most popular methods to realize this target. AAN and AAP both relate to set some restrictions to the previous queries so as to decrease the number of results. For CT, people might change terms to some more specific ones to reduce the number.

As for increasing the range of queries, SPM (spelling change) is the most popular means. The use of SPM supports the idea that people often fail to get their desired information as a result of spelling mistakes.

Using a software application, we automatically submitted the two queries from the two-query sessions to Google separately, retrieving the number of results reported by the search engines in order to gauge the effectiveness of the searchers’ efforts. The results of Google are used as a benchmark here due to the popularity of Google. In Table 10, we present the results this analysis.

Table 10. Result of Query Reformulation

Effect	Occurrences	Success Occurrences	Failure Occurrences	Percentage of Failure
DEC	31	30	1	3.23%
INC	15	13	2	13.33%
TOTAL	46	43	3	6.52%

From Table 10, 6.52% of users fail to realize their targets. 13.33% of users fail to increase the number of the results. Only 3.23% of them fail to decrease the number. It seems people are good at decreasing the range of queries.

Discussion

Based on the transaction log analysis above focusing on modifying queries, people often changed their topics or modified nouns so as to make some changes to their queries. Topic

change is the same concept as Fidel's conceptual moves. It seems people's information need about one topic is easy to satisfy and usually is fulfilled after the initial searching. Researchers found some basic level in human categorization that has a special significance on human's view of world. (Rosch et al, 1976) This means based on knowledge of these basic categories, people could generate some other meaningful information on other levels, and they use this basic level quite often. This could explain why people change topics so frequently. They might use the basic categories most of the time and generate the needed information immediately after the first query.

The data also shows that people are more comfortable in modifying nouns. Modifying nouns is a type of operational move. It is reported that "children learn nouns before predicate terms" (Gentner, 1982, p.327) and "in early-production vocabularies, nouns greatly out number verbs" (Gentner, 1982, p.327). Gentner (1982, p.327) also pointed out "nouns predominate over predicate terms". Nouns are "object-reference terms" (Gentner, 1982, p.328) and have "a particularly transparent semantic mapping to the perceptual-conceptual world" (Gentner, 1982, p.328). Predicate terms have "fewer psychological constraints on their possible conflationary patterns" (Gentner, 1982, p.328) than nouns.

Then, people's searching process could be viewed as a learning process. People have begun to learn as soon as they are born to the world. As children, they are inclined to learn nouns than predicate terms. (Gentner, 1982, p.327) Nouns connect us to the "new" information and "new" knowledge. This learning method is innate and accompanies us all our lives. Thus every time when we search "new" information, we use nouns to approach the retrieval processes. Researchers in cognitive science also pointed out in many languages "nouns and verbs have prototypical functions in discourses" (Lakoff, 1987, p.64). Our findings also support this to a certain extent.

We also found that people like to subtract terms to make query changes, such as noun, phrase, or adjective. Subtracting is an operational move. Subtracting words usually will lead to a broader query and it is a typical way to increase the range of queries. People seem familiar with the basic IR techniques.

From the effect of query reformulation analysis, we also found that people usually to realize their targets in terms of restricting or "loosening" the number of results. All the failures happened due to polysemy. After using search engines for a long time and because of frequent use of the Web, people are familiar with search engines and some fundamental search skills. They basically know how to form and reform queries.

Conclusion and limitations

From our research, we find that

- While modifying queries, people often changed their topic, modified nouns or subtracted terms so as to make some changes of their queries; and
- People had some basic skills to modify queries to make desired effective changes.

Since every research has strengths and weaknesses, our project has strengths and weaknesses as well. We conducted an in-depth analysis of query reformulation, pointed out some interesting findings, and also provided reasonable explanations.

However, this project also has some shortcomings. We used data from 2002. Web searching behaviors could have changed. However, nearly 20% of queries include only one term has remained the same from 1998 to 2002. (Jansen, Spink & Pedersen, 2005). We also identified users according to IP and cookies. This is a good method but not a hundred percent precise approach to identifying individual users. Some temporal cur-off may be needed. This would have impacted our percentage of TC query modifications.

Our future research includes the development of an automated process to analysis large data sets for query reformulation. If applications could automatically identify reformulation patterns in real-time, these applications could assist searchers in located relevant information by aiding in the reformulation process.

References

Bates, M. (1979) Information search tactics. *Journal of the American Society for Information Science*, 30(4), 205-214.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In S. Kuczaj (Ed.), *Language Development: Language, cognition, and culture* (pp.301-334). Hillsdale, NJ: Erlbaum.

Fidel, R. (1985). Moves in online searching. *Online Review*, 9(1), 61-74.

Jansen, B. J., 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.

Jansen, B. J., Spink, A, and Pederson, J. (2005) *Trend analysis of AltaVista Web searching*. *Journal of the American Society for Information Science and Technology*. 56(6), 559-570.

Jansen, B. J., 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.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago and London: University of Chicago Press.

Rieh, S. Y. & Xie, H. (2006). Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing and Management*, 42(3), 751-768.

Rosch, E. etc. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382-439.

Spink, A., Jansen, B. J., & Ozmultu, C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research*, 10(4), 317-328.

Spink, A., Wolfram, D., Jansen, B. J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Stojanovic, N. (2005). On the conceptualization of the query refinement task. *Library Management*, 26(4/5), 231-293.

Appendix A:

Table 11. Pattern Description

Pattern	Description
AAA	adjective after term
AAD	definite article after term
AAN	noun after term
AAP	phrase after term
AAS	synonym after term
AAV	verb after term
AAX	suffix after term
ABA	adjective before term
ABD	definite article before term
ABN	noun before term
ABP	phrase before term
ABS	synonym before term
ABV	verb before term
AIA	adjective in between terms
AIB	Boolean in between terms
AIC	conjunction in between terms
AIN	noun in between terms
AIP	phrase in between terms
AIS	synonym in between terms
AIV	verb in between terms
AP	add a phrase
CC	change conjunction
CG	change from plural to singular
CLT	change to last term
CP	change from singular to plural
CS	change to synonym
CT	change to related term
CU	change to url
INT	initial query
PBT	preposition between terms
QEQ	quote entire query
RO	return to original query
SA	subtracting adjective
SC	subtracting conjunction (and, or..)
SD	subtracting definite article (the)
SF	subtracting first term
SL	subtracting last term
SM	subtracting middle term
SN	subtracting noun
SP	subtracting phrase
SPM	spelling change
SV	subtracting verb
TC	topic change
UQEQ	unquote entire query

Table 12. Effect Description

Effect	Description
DEC	decreases the range of the query
INC	increases the range of the query
NOC	no change in the range of the query
TC	change in topic