ANALELE ŞTIINŢIFICE ALE UNIVERSITĂŢII „ALEXANDRU IOAN CUZA" DIN IAŞI
Număr special **Ştiinţe Economice** **2010**

PATTERNS IN QUERY REFORMULATION IN ONLINE SEARCHING BEHAVIOR

Sinziana SPIRIDON
Alexandru Ioan Cuza University of Iasi, Faculty of Economics and Business Administration
Iasi, Romania
sinziana.spiridon@gmail.com

Abstract

The present study investigates the characteristics of online searching behavior with emphasis on query reformulation. The topic has been highly studied starting with Bates (1979), Jansen and Spink (1999), Rieh and Xie (2001), Vakkari (2003), Spink, Zhang and Jansen (2007). Earlier studies focused on analyzing patterns with focus on form and topic changes (generalization, specialization, parallel movement etc.), some of the latest studies began focusing on semantic aspects in defining the types of reformulations.

We analyze the frequency of appearance for the different types of query reformulation based on the semantic approach used by Spink, Zhang and Jansen (2007). We therefore focus our analysis on query reformulations by adding or subtracting nouns, adjectives, verbs; conjunctions etc. before, after and between the terms of a prior query. We also analyze reformulations based on the use of synonyms or related terms and topic change.

The analyses regarding patterns of query reformulation focuses on sequences of two and three reformulations.

We examined 50.000 interactions collected from the computer logs of the Faculty of Economics and Business Administration library registering students' online activity during a normal day of school. The results show students use frequent changes in topic and noun and adjective addition in their online searches.

This research is a step on identifying patterns in online searching particularly realized by students with implications in understanding how they use search engines in their educational activities.

Keywords: Query modification, query reformulation, search engine, Web search, patterns.
JEL classification: D83, L86

1. INTRODUCTION

Information search is a continuous learning process characterized by the permanent change of peoples' conceptions about their searching task and need of information as stated by Ingwersen (1992), Belkin, Cool, Stein and Thiel (1995). Information searching on the web implies a complex set of actions including formulating and reformulating interrogation

queries, browsing through the web pages, evaluating content, downloading files, watching movies, listening music etc. All these are expressions of the searcher's state of knowledge and in the same time elements that contribute to its continuous development.

The searching of information in the online environment is influenced by two major aspects: the familiarity with the subject of search and familiarity with internet, search engines and online searching [4], [7], [17]. In this paper we focus on the queries, term, search sessions and query reformulation tactics with the purpose of understanding the patterns of query reformulation in searching sessions with focus on particularities of students in economics.

2. RELEVANT STUDIES AND RESEARCH QUESTIONS

Studies on query reformulation cover various directions of research. Starting with Bates (1979) which realized a detailed classification of reformulation tactics encountered in the searching process and beginning a more complex approach from the linguistic point of view with Spink, Zhang and Jansen (2006). Despite of the accelerated evolution of the internet the classification of reformulation tactics realized by Bates in 1979 remains a fundamental point of departure for understanding and studying online searching behavior. Bates categorizes searching tactics in 4 main types: *monitoring tactics, file structure tactics, search formulation tactics* and *term tactics.*

Fidel (1991) and Rie, Xieh (2001) define the search tactics using as main criteria the effect on the conceptual content of the query. Fidel operates with operational versus conceptual reformulations while Rie and Xieh use 3 types of tactics: query form modification tactics, query content modification tactics and resources oriented tactics.

With Lau, Horvitz (1999), Vakkari (2001), Wildemuth (2004), Rie, Xieh (2005) queries are analyzed further from the point of view of content and form effect of reformulation tactics. The types of search tactics defined and analyzed focus on the same main concepts: decrease of query range, increase of query range, reformulations with no effect on the query range and reformulations that have as effect the change of topic. Still, the important differences regarding the definition of each query reformulation tactics as concepts used in analyses makes it very difficult to compare the results and findings of the various studies on the subject.

The analyses of online searching behavior has multiple applications, most of them oriented on understanding searching behavior, enhancing online searching systems, developing successful web page content and design. The present study has as objective the following research question: *How do students in economics formulate queries while searching online?* It is our purpose to analyze the searching behavior from the linguistic point of view using the classification of Spink, Zhang and Jansen as a model for identifying query reformulation tactics and reformulation patterns. Spink, Zhang, Jansen (2006) define categories of query reformulation tactics from a linguistic point of view analyzing the way searchers use nouns, adjectives, verbs, conjunctions etc. in their searching process.

3. RESEARCH DESIGN

3.1. Data collection

The data has been collected from the server of the Faculty of Economics and Business Administration Library in form of a log containing the student – internet interactions from the computers used in the library on 5 of April 2009. The original transaction log contained 53056 records each with 24 fields with data regarding search engines interrogation and also web page browsing. We used a number of 6 relevant fields in our analysis:

- c-ip – computer identification code
- time – moment of each interaction measured in hours, minutes and seconds
- cs-reffered – current web page (including the link for the current web page)
- cs-uri – browser action (including elements of current web page and secondary browser actions)
- status – only records with value 200 were included in the study
- action – only records with value "Allowed" were included in the study

We also used an additional file containing data about the logon and logoff activity on each computer which allowed us to identify the individual sessions of each searcher. We imported the data in Excel software were we manually identified the records containing valid web interrogations.

3.2. Data preparation

For the delimitation of search sessions we used the data regarding searchers logon and logoff the library computers in combination with a threshold of 15 minutes [6] in order to separate distinctive search sessions conducted during the same logon session. We used this delimitation for identifying the first searching query in every session and than recreated the chronological series of actions in each session. The methodology used for collecting the data offers the advantage of not having to deal with queries from non-human submissions, a difficulty encountered in the studies where data are gathered directly from a search engine log.
The analysis uses the following main concepts:

- *Term* - series of characters separated by white space or other separator;
- *Query* - the entire string of terms submitted by a searcher during an interaction with the search engine;
- *Session* - the entire series of queries submitted by a user during the interactions with the Web search engine in the frame time delimited by logon moment and logoff moment;
- *Initial query* - the first query submitted in a session;
- *Identical query* - a query within a session that is a copy of a previous query within that session.

The final data base included a number of 356 query reformulations submitted by 50 users. Data were manually coded using Spink, Zhang and Jansen (2006) list of reformulation tactics and adding one tactic identified in the database and not present in their list, LCG (change to another language). This tactic has been introduced in order to be able to code two reformulations consisting in translation from English to Romanian and respectively from Romanian to English. The original list of tactics is present in Appendix A.

3.3. Data analysis

For the analysis the data has been imported in SPSS where separate variables for encoding the two and respectively three term patterns were created.

4. RESULTS

4.1. Query and session length

The results obtained regarding the number of terms per query match previous results of similar studies. For the present study the queries with 2 terms representing 26.12% of the queries and with 3 terms representing 18.82% are similar with the results obtained by Jansen, Spink, Bateman, Saracevic [8], C. Silverstein, M. Henzinger, H. Marais, M. Moricz [11], Lau, Horvitz [13]. The average number of terms per query obtained is 3.62 significantly larger than the one obtained in the mentioned studies which have values smaller than 3.00. This is determined by the high percentage (57.03%) of queries with more than 3 terms.

The analysis of session length reveals significant differences compared to previous studies in the field given by a higher percentage of 3+ terms sessions. We obtained a percentage of 22.81% of 1 term length sessions, 22.81% of 2 terms length sessions, 54.38% of 3+ terms sessions compared with 31% [15], 14%[14] and 8.9%[11]. The difference might be explained by the characteristics of the population chosen for this study, students in economics familiarized with the internet and the use of search engines and by the characteristics of search tasks mainly informational tasks regarding academic projects.

4.2. Query reformulation

The most frequent query reformulations tactics were TC (topic change) used in 33.66% of reformulations, CT (change to related term) used in 24.60% of reformulations, AAN (add noun after term) in 10.68% of reformulations and SPM (spelling change) in 6.15% of reformulations. Table 1 presents the complete results obtained from the analysis of query reformulation tactics.

Table no. 1 – Query reformulation tactics

	Frequency	Percent
TC	104	33.66
CT	76	24.60
AAN	33	10.68
SPM	19	6.15
SN	15	4.85
ABN	11	3.56
IDC	9	2.91
RO	9	2.91
AAP	8	2.59
AIN	5	1.62
SP	5	1.62

CS	4	1.29
ABP	3	0.97
CG	2	0.65
LCG	2	0.65
AAV	1	0.32
AIA	1	0.32
AIV	1	0.32
SC	1	0.32
Total	309	100.00

The results are similar to those obtained by Spink, Zhang and Jansen [16] regarding the predominant types of reformulation tactics with significant differences regarding the percentage of use.

From the point of view of reformulation effect on query content, the most frequent reformulation tactics had as effect a decrease in the query range (DEC) representing 33.97% of query reformulations, 26.98% of query reformulations increased the range of the query (INC), 25.71% had no effect on the range of the query and 13.33% represented a change of topic.

Table no. 2 – Query reformulation tactics used for decreasing the range of the query

	Frequency	Percent
AAN	33	40.74
CT	14	17.28
ABN	11	13.58
AAP	8	9.88
AIN	5	6.17
IDC	3	3.70
ABP	3	3.70
AAV	1	1.23
RO	1	1.23
AIV	1	1.23
AIA	1	1.23
Total	81	100.00

The most frequent reformulation tactics used for decreasing the range of the query are AAN (add noun after term) used in 40.74% of reformulations, CT (change to related term) in 17.28% of reformulations and ABN (add noun before term) in 13.58% of reformulations.

Table no. 3 - Query reformulation tactics used for increasing the range of the query

	Frequency	Percent
CT	16	38.10
SN	15	35.71
SP	5	11.90

IDC	4	9.52
CS	1	2.38
RO	1	2.38
Total	42	100

The reformulation tactics used most frequently to increase the range of the query are CT (change to related term) used in 38.10% of reformulations, SN (subtracting noun) in 33.71% of reformulations and SP (subtracting phrase) in 11.90% of reformulations.

Table no. 4 - Query reformulation tactics with no effect upon the range of the query

	Frequency	Percent
CT	45	56.96
SPM	19	24.05
CS	3	3.80
RO	3	3.80
CG	2	2.53
IDC	2	2.53
LCG	2	2.53
TC	2	2.53
SC	1	1.27
Total	79	100

Query reformulation tactics used frequently in reformulations with no effect on the range of the query were CT (change with related term) – 56.96% and SPM (spelling change) – 24.05%.

4.3. Query reformulation patterns

The most frequent reformulation pattern of two sequences used to initiate a search is topic change (TC) used in 48.65% of the search sessions with at least two queries.

Table no. 5 – Query patterns of two queries for search session initiation

	Frequency	Percent
INT-TC	18	48.65
INT-CT	7	18.92
INT-AAN	4	10.81
INT-AAP	3	8.11
INT-ABN	2	5.41
INT-SPM	2	5.41
INT-RO	1	2.70
Total	37	100

The results of analyzing all the sequences of two queries from the sessions with two or more than two queries show as most frequent patterns the succession of two topic changes (TC-TC), the succession of an initial term followed by a topic change (INT-TC), the succession of two changes to related term (CT-CT) and the succession of a change to related term followed by a topic change (CT-TC) as shown in table 4 presented below.

Table no. 6 - Query patterns of two queries in sessions with two or more queries

	Freq.	Percent
TC-TC	42	11.80
INT-TC	18	5.06
CT-CT	17	4.78
CT-TC	16	4.49
TC-CT	14	3.93
TC-AAN	13	3.65
AAN-CT	11	3.09
CT-SPM	8	2.25
AAN-TC	7	1.97
CT-AAN	7	1.97
INT-CT	7	1.97
SPM-TC	6	1.69
TC-ABN	6	1.69
RO-TC	5	1.40
SPM-CT	5	1.40
TC-SPM	5	1.40
ABN-CT	4	1.12
ABN-TC	4	1.12
CT-SN	4	1.12
INT-AAN	4	1.12
SN-AAN	4	1.12
AAN-SN	3	0.84
AAP-CT	3	0.84
AIN-AIN	3	0.84
CT-RO	3	0.84
INT-AAP	3	0.84
SN-ABP	3	0.84
TC-RO	3	0.84

Table no. 7 – Query patterns of three queries from search sessions of three or more queries (selection of most relevant patterns)

	Freq.	Percent

TC-TC-TC	14	3.93
INT-TC-TC	10	2.81
CT-TC-TC	7	1.97
TC-CT-TC	6	1.69
TC-TC-AAN	6	1.69
AAN-CT-CT	4	1.12
CT-CT-CT	4	1.12
CT-CT-TC	4	1.12
TC-SPM-TC	4	1.12
CT-CT-AAN	3	0.84
CT-TC-CT	3	0.84
INT-TC-CT	3	0.84
SPM-TC-TC	3	0.84
TC-AAN-CT	3	0.84
TC-AAN-TC	3	0.84
TC-TC-CT	3	0.84
TC-TC-SPM	3	0.84

The results presented in table 5 mention as most frequent three queries patterns the succession of three consecutive topic changes (TC-TC-TC), the sequences beginning with an initial query followed by two consecutive topic changes (INT-TC-TC) and the pattern containing a change to related term, followed by a topic change and than another change to related term (CT-TC-CT).

5. CONCLUSION AND FURTHER STUDIES

The main differences between the results obtained in the present study and the studies used as theoretical and empirical background are the significantly higher percentage of queries with 3+ terms and session with 3+ queries. The differences might be determined by the particularities of the population used for realizing the analysis and will make de subject of further studies on a larger database.

The most frequent patterns of query reformulation include the tactics of topic change and change with a related term for both two reformulations sessions and three reformulation sessions. This might indicate a lack in defining searching strategies. The analysis defined two categories of search patterns: a category with a higher degree of coherence including changes in terms and the use of nouns, verbs. On the other hand the data base included sessions with less coherence including mostly changes of topics or changes to related terms. This results presents interest for a future study including the analysis of factors influencing searching behavior such as subject familiarity and the experience of using search engines.

References

[1] Bates Marcia J., *Information Search Tactics*, Journal of the American Society for Information Science, vol. 30, nr. 4, p.205, 1979

[2] Belkin Nicholas J., Cool Colleen, Stein Adelheit, Thiel Ulrich, *Cases, Scripts, and Information-Seeking Strategies: On the Design of Interactive Information Retrieval Systems*, Expert Systems with Applications, Vol. 9, 1995
[3] Efthimiadis Efthimis N., *Query Expansion*, Annual Review of Information Systems and Technology (ARIST), vol. 31, pp. 121-187, 1996.
[4] Fenichel, Carol-Hensen, *Online Information retrieval - Identification of measures that discriminate among user with different levels and types of experience*, Drexel University, 1979
[5] Fidel Raya, *Searchers' Selection of Search Keys: I. The Selection Routine*, Journal Of The American Society For Information Science, vol. 42(7), pp. 490-500, 1991
[6] He Daqing, Goker Ayse, Harper David J., *Combining evidence for automatic Web session identification*, Information Processing and Management, Vol. 38, pp. 727–742, 2002
[7] Hsieh-yee Ingrid, *Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers*, Journal of the American Society for Information Science, Vol. 44(3), pp. 161-74, 1993
[8] Jansen Bernard J., Spink Amanda, Bateman Judy, Saracevic Tefko, *Real life information retrieval: a study of user queries on the Web*, ACM SIGIR Forum, Vol. 32(1), pp. 5 – 17, 1998
[9] Rieh Soo Young, Xie Hong (Iris), *Patterns and Sequences of Multiple Query Reformulations in Web Searching: A Preliminary Study*, Proceedings of the 64th ASIST Annual Meeting, Vol. 38, pp. 246-255, 2001
[10] Rieh Soo Young, Xie Hong (Iris), *Analysis of multiple query reformulations on the web: The interactive information retrieval context*, Information Processing & Management, vol. 42(3), pp. 751-768, 2006
[11] Silverstein Craig, Henzinger Monika, Marais Hannes, Moricz Michael, *Analysis of a Very Large Web Search Engine Query Log*, ACM SIGIR Forum, vol. 33(1), pp. 6 – 12, 1999
[12] Spink Amanda, Bateman J., Jansen Bernard. J., *Searching heterogeneous collections on the web: Behavior of Excite users*, Information Research: An Electronic Journal, vol. 5(2), 1998
[13] Lau Tessa, Horvitz, Eric, *Patterns of Search: Analyzing and Modeling Web Query Refinement*, User Modeling: Proceedings of the Seventh International Conference, UM99, Springer, Viena, pp.119-28, 1999
[14] Spink Amanda, Chang Carol, Goz Agnes, *Users' Interactions With The Excite Web Search Engine: A Query Reformulation And Relevance Feedback Analysis*, Internet Research Journal, Vol. 9(2), pp. 117 – 128, 1999
[15] Spink Amanda, Wolfram Dietmar, Jansen Bernard J., Saracevic Tefko - Searching the Web: The Public and Their Queries, Journal of The American Society For Information Science And Technology, vol. 52(3), pp. 226–234, 2001
[16] Spink Amanda, Chang Carol, Jansen Bernard J., *Patterns and Transitions of Query Reformulation during Web Searching*, International Journal of Web Information Systems, vol. 3(4), pp. 328 – 340, 2007
[17] akkari Pertti, *Changes in Search Tactics and Relevance Judgments when Preparing a Research Proposal A Summary of the Findings of a Longitudinal Study*, Information Retrieval, Vol. 4, pp. 295–310, 2001, Kluwer Academic Publishers
[18] Wildemuth, B., *The Effects of Domain Knowledge on Search Tactic Formulation,* Journal of the American Society for Information Science and Technology, Vol. 55 (3), pp. 246 – 258, 2004
[19] Wolfram Dietmar, *Term co-occurrence in Internet Search Engine Queries: An Analysis of the Excite Data Set*, Canadian Journal of Information and Library Science, Vol. 24, pp. 12-33, 1999

Table no. 1 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)