Query Modifications Patterns During Web Searching

Bernard J. Jansen
The Pennsylvania State University
jjansen@ist.psu.edu

Amanda Spink
Queensland University of Technology
ah.spink@qut.edu.au

Bhuva Narayan
Queensland University of Technology
bhuva_narayan@yahoo.com

Abstract

We examine 2,465,145 interactions from 534,507 users of Dogpile.com submitted 6 May 2005. We compare query reformulation patterns. We investigate the type of query modifications and query modification transitions within sessions. Searchers most often modified their query by changing query terms (nearly 23% of all query modifications). Searchers' queries undergoing modification typically transition from Web to Image collections in content shifts (37% of all content transitions), and searchers typically implement assistance at the start of a session or when switching content collections (21% of all assistance usage). This research sheds light on the more complex aspects of Web searching involving query modifications.

Keywords: Web searching, search engines, Web queries, query reformulation, Web sessions

1. Introduction

The goal for the searcher during a Web session is to locate information to satisfy their information need. For evaluation, one can view success or failure at the session level as the critical determinant in the user's perception of the Web search engine's performance. Therefore, the session level is a key paradigm for measuring the performance of Web search engines. Attempts at designing personalized Web systems relying on session-level data have taken a variety of approaches. CiteSeer [1] utilizes an agent paradigm to recommend computer science and computer science-related articles based on a user profile.

Jansen and Pooch [2] designed a client side application for Web search engines that provided targeted searching assistance based on the user interactions during a session. The researchers noted that there are predictable patterns of when searchers seek and implement assistance from the system [3]. These patterns may indicate when the searcher is open to assistance from the system, thereby avoiding task interruptions.

Anick [4] examined the interactive query reformulation support of the AltaVista search engine for searchers using transactions logs. The researcher used a baseline group of AltaVista searchers given no query feedback and a feedback group offered twelve refinement terms along with the search results. There was no significant difference in searching performance between the two groups. However, Belkin, et. al., [5] reported that query expansion may be helpful and improve searching performance.

The purpose of the present study is to expand our knowledge in predicting the future actions of searchers on searching systems into order to provide targeted searching assistance. Specifically, we aim to determine the query modification patterns which users search for information on Web searching systems. We refer to each query modification event during a session as a search state.

This line of research is important because if a Web searching system can predict the future state of searchers, the system can provide targeted searching assistance to aid searchers in their information seeking task. If we can determine an appropriate order of the search process (i.e., number of predictive states), this indicates an upper bound for prediction, which will provide us the most predictive power at the least computational complexity.

2. Related Studies

On the Web, the difficulty of how to define a search session is due to the stateless nature of the client-server relationship. Most Web search engines servers have used the IP address of the client machine to identify unique visitors. With referral sites, Internet service providers (ISP), dynamic Internet Protocol (IP) addressing, and common user terminals it is not always easy to identify a single user session on a Web search engine. Therefore, a single IP address does not always correspond to a single user.

Contained within a single Web session from any of these definitions, the searcher may be engaged in multitasking searching tasks [6] or successive sessions over time that are related to the same topic [7]. There has been some research into using the query context to define the session. He, Göker and Harper [8] used contextual information from a Reuters transaction log and a version of the Dempster–Shafer theory in an attempt identify search engine session boundaries.

Özmutlu and Cavdur [9] attempted to duplicate the findings of [8], but the researchers reported that there were issues relating to implementation, algorithm parameters,

and fitness function. Özmutlu and Cavdur [9, 10] investigated the use of neural networks to automatically identify topic changes in sessions, reporting high percentages (72% - 97%) of correct identifications of topic shifts and topic continuations. Rieh and Xie also investigated query reformulations [11].

This study examines three methods of session identification representing the major approaches taken to identify Web searching sessions. We compare the results among these three methods of session identification.

3. Research Questions

Our research question is: *What are the query modification patterns of searchers during Web sessions?* We investigated the manner of query modification during sessions. We develop a classification method for queries based on prior research in Web search [12, 13]. We provide aggregate results for each query classification category. We then extend this first level classification by analyzing intra-session query transactions (i.e., movement from one type of query to the next).

4. Research Design

4.1 Web Data

Dogpile.com (http://www.Dogpile.com/) is a meta-search engine, owned by Infospace, Inc. When a searcher submits a query, Dogpile.com simultaneously submits the query to multiple other Web search engines, collecting the results from each, removing duplicates results, and aggregating the remaining results into a combined ranked listing using a proprietary algorithm. Dogpile.com integrates the results of the four leading Web search indices (i.e., Ask Jeeves, Google, MSN, and Yahoo!) along with other search engines into its search results listing. So, Dogpile.com provides one of the most complete content collections on the Web to respond to Web searchers' queries. Meta-search engines provide a unique service by presenting the alternate results provided by the various search engines, which have a low rate of overlap [14].

4.2 Data Collection

We collected the records of searcher – system interactions in a transaction log that represents a portion of the searches executed on Dogpile.com on 6 May 2005. The original general transaction log contained 4,056,374 records, each containing seven fields:

- *User Identification*: a code to identify a particular computer

- *Cookie*: an anonymous cookie automatically assigned by the Dogpile.com server to identify unique users on a particular computer.
- *Time of Day*: measured in hours, minutes, and seconds as recorded by the Dogpile.com server on the date of the interaction.
- *Query Terms*: the terms exactly as entered by the given user.
- *Location*: a code representing the geographic location of the user's computer as denoted by the computer's IP address.
- *Source*: the content collection that the user selects to search (e.g., *Web, Images, Audio, News,* or *Video*), with *Web* being the default.
- *Feedback*: a binary code denoting whether or not the query was generated by the *Are You Looking for?* query reformulation assistance provided by Dogpile.com.

We imported the original flat ASCII transaction log file of 4,056,374 records into a relational database. We generated a unique identifier for each record. We used four fields (*Time of Day*, *User Identification*, *Cookie*, and *Query*) to locate the initial query and then recreate the sequential series of actions from a particular user, determined by *User Identification* and *Cookie*. An analysis of the dataset shows that the interactions of Dogpile.com searchers was generally similar to Web searching on other Web search engines [15].

4.3 Data Preparation

The terminology that we use in this research is similar to that used in other Web transaction log studies [c.f., 2]. For this research, we are interested in queries submitted by humans, and the transaction log contained queries from both human users and agents. There is no acknowledged methodology for precisely identifying human from non-human submissions in a transaction log. Therefore, researchers normally use a temporal or interaction cut-off [16].

We selected the interaction cut-off approach by removing all sessions with 100 or more queries. This cut-off is substantially greater than the reported mean number of queries [17] for human Web searchers. This cutoff most likely introduced some agent sessions; however, we were reasonable certain that we had included most of the queries submitted primarily by human searchers.

4.4 Data Analysis

We used the following algorithm to classify content changes within sessions.

4.4.1 *Method 1: IP, Cookie, and Content Change:*

We used a contextual method to identify sessions. We again used the searcher's IP address and the browser cookie to determine the initial query and subsequent queries. Instead of using a temporal cut-off, we used changes in the content of the user queries.

For this method, we assigned each query into a mutually exclusive group based on an IP address, cookie, query content, use of the feedback feature, and query length. The classifications are:

- *Assistance:* the current query was generated by the searcher's selection of an *Are You Looking For?* query (see http://www.dogpile.com).
- *Content Change*: the current query is identical but executed on another content collection.
- *Generalization:* the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more general information.
- *New:* the query is on a new topic.
- *Reformulation:* the current query is on the same topic as the searcher's previous query and both queries contain common terms.
- *Specialization*: the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more specific information.

The *initial query* (Q_i) from a unique IP address and cookie always identified a new session. In addition, if a *subsequent query* (Q_{i+1}) by a searcher contained no terms in common with the previous query (Q_i), we also deemed this the start of a new session. Naturally, from an information need perspective, these sessions may be related at some level of abstraction. However, with no terms in common, one can also make the case that the information state of the of the user changed, either based on the results from the Web search engine or from other sources [18]. Also, from a system perspective, two queries with no terms in common represent different executions to the inverted file index and content collection. We classified each query using an application that evaluated each record in the database. We built our algorithm from that presented by He, Göker, and Harper [8].

5. Results

5.1 Query Reformulation during Sessions

For our research question (*What are the query modification patterns of searchers during Web sessions?*), we used the algorithm discussed in section 4.4.1, which classified each query into one of our mutually exclusive groupings. From this classification, we were able to analyze the occurrences of type of query modifications and the transactions from one type of query to another within a session. We compare results to using just the IP address and cookie and to IP address, cookie, and a 30 minute time limit.

We see from Table 1 that more than 8% of the query modifications were for *Reformulation*, with another approximately 8% of query modifications resulting from system *Assistance*. If we exclude the *New* queries, *Reformulation* and *Assistance* account for nearly 45% of all query modifications. This finding would seem to indicate that a substantial portion of searchers go through a process of defining their information need by exploring various terms and system feedback to modify the query as an expression of their information need. Another 16% of query modifications are *Specialization*, supporting prior reports that precision is a primary concern for Web searchers [19].

Table 1: Query reformulation.

Search Patterns	Occurrence	%	Occurrence (excluding *New*)	% (excluding *New*)
New	964,780	63.34%	-	-
Reformulation	126,901	8.33%	126,901	22.73%
Assistance	124,195	8.15%	124,195	22.25%
Specialization	90,893	5.97%	90,893	16.28%
Content change	65,949	4.33%	65,949	11.81%
Specialization with reformulation	55,531	3.65%	55,531	9.95%
Generalization with reformulation	54,637	3.59%	54,637	9.78%
Generalization	40,186	2.64%	40,186	7.20%
	1,523,072	100.00%	558,292	100.00%

IEEE
COMPUTER
SOCIETY

Table 2. Transition among content.

Content Transition	Occurrences	%
Web to Images	12,080	37.21%
Web to Audio	2,411	7.43%
Web to Video	1,298	4.00%
Web to News	602	1.85%
Images to Web	6,882	21.20%
Images to Audio	553	1.70%
Images to Video	2,096	6.46%
Images to News	202	0.62%
Audio to Web	1,537	4.73%
Audio to Images	581	1.79%
Audio to Video	1,410	4.34%
Audio to News	53	0.16%
Video to Web	1,036	3.19%
Video to Audio	1,006	3.10%
Video to News	143	0.44%
News to Web	370	1.14%
News to Images	123	0.38%
News to Audio	25	0.08%
News to Video	59	0.18%
	32,467	100.00%

Table 3. Transitions of query modifications within Web sessions.

Query Pattern Shift	Occurrences	% (within sub-category)	% (within entire dataset)
Assistance to content change	18,474	57.84%	4.29%
Content change to assistance	10,688	40.91%	2.48%
Generalization to specialization	6,790	37.17%	1.58%
Generalization with reformulation to reformulation	8,455	31.91%	1.96%
Specialization to reformulation	13,049	32.02%	3.03%
Specialization with reformulation to generalization with reformulation	9,719	36.35%	2.26%
Reformulation to specialization with reformulation	8,826	22.70%	2.05%
New to assistance	58,471	26.42%	13.58%
New to specialization	62,405	28.20%	14.50%

To explore this further, we first investigate the transitions by searchers among the content collections. For example, if a searcher entered a new query using the *Web* content collection, then executed this query on the *Image*

collection; this transition would be labeled as *Web – Image*. Table 2 contains the results of this analysis.

As we see from Table 2, the major content transitions were from *Web* to *Images* (37%) and *Images* to *Web* (21%). The *Web* was the default content collection, and it appears that *Images* is by far the second most popular content collection.

We also investigated the transitions by searchers from one query to the next in terms of query classification. For example, if a searcher entered a new query then reformulated the next query; this transition would be labeled as *New – Reformulation*. We conducted this analysis for the entire data set at the session level. Due to page limitations, we report only the most frequently occurring transitions from each classification. Table 3 contains the results of this analysis.

We see from the results in Table 3, that there appears to be a connection between the searcher shifting content collections and the use of system assistance with near majorities (58%) of assistance usage occurring just before a content change or just after (41%) a content change. These shifts accounted for 25% of all assistance usage.

We see high occurrences of query reformulation after *Generalization* (17%) and *Specialization* (32%), with a variety of reformulation variations. This would indicate that searchers use the interactions with the system, probably the results listings, to explore the information space with new query terms. There also appears to be a tendency to go from *Generalization* to *Specialization* (37%). *Specialization* also appears to be a tendency immediately after the initial query, with 28% of searchers immediately moving to narrow their queries. Searchers also appear to be open to *Assistance* (26%) at the start of the session.

5.2 Accuracy of Classification

We conducted a verification of our classification algorithm by manually classifying 2,000 queries. We arrived at 5 categories of errors, developed a priori:
1) *Misspelling:* a word was misspelled or a previously misspelled word causing a change resulting in a misclassification (causes a false *New* or *Reformulation*).
2) *Cookie:* either cookie not defined or change in cookie but not a change in user (causes a false *New*).
3) *Special character change*: the original query contained special characters (causes a false *New* or *Reformulation*).
4) *Time gap*: time gap between queries was too large to be considered a session, but Q_i and Q_{i-1} were still related (causes a false *New*).
5) *Other:* a miscellaneous collection of other reasons (causes a false *New*).

We see from Table 4 that most of the errors were due to misspellings (i.e., the algorithm counted the word as a new term when in realty the searcher had misspelled a term in the original query and corrected the term in the subsequent query. Most misspellings occurred due to missing spaces in words. However, the sum total of all misclassifications was 4.45%, resulting in a 95.55% accuracy rate for the algorithm.

6. Discussion

Results show that the majority of content switching occurs between *Web* and *Image* collections. In other query modifications, searchers appear to execute a great deal of *Reformulation* as they try to express more precisely their information need. They typically move to narrow their query at the start of the session, moving to *Reformulation* in the mid and latter portions of the sessions. Web search engine users seem to be receptive to system searching *Assistance* at the start of the session or when switching among content collections.

Table 4.Query reformulation.

Search Patterns	Occurrence	%	Occurrence (excluding *New*)	% (excluding *New*)
New	964,780	63.34%	-	-
Reformulation	126,901	8.33%	126,901	22.73%
Assistance	124,195	8.15%	124,195	22.25%
Specialization	90,893	5.97%	90,893	16.28%
Content change	65,949	4.33%	65,949	11.81%
Specialization with reformulation	55,531	3.65%	55,531	9.95%
Generalization with reformulation	54,637	3.59%	54,637	9.78%
Generalization	40,186	2.64%	40,186	7.20%
	1,523,072	100.00%	558,292	100.00%

COMPUTER SOCIETY

7. Conclusion and Future Research

For future research, these algorithms may facilitate cross-system investigations. An attempt to standardize query reformation detection would also enhance comparative transaction log analyses.

Acknowledgements

8. References

[1]   S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," *IEEE Computer,* vol. 32, pp. 67-71, 1999.

[2]   B. J. Jansen and U. Pooch, "Web User Studies: A Review and Framework for Future Work," *Journal of the American Society of Information Science and Technology,* vol. 52, pp. 235-246, 2001.

[3]   B. J. Jansen, "Seeking and implementing automated assistance during the search process," *Information Processing & Management,* vol. 41, pp. 909-928, July 2005.

[4]   P. Anick, "Using Terminological Feedback for Web Search Refinement - A Log-Based Study," in *the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 2003, pp. 88-95.

[5]   N. Belkin, C. Cool, D. Kelly, H.-J. Lee, G. Muresan, M.-C. Tang, and X.-J. Yuan, "Query length in interactive information retrieval," in *the 26th Annual International ACM Conference on Research and Development in Information Retrieval*, Toronto, Canada., 2003, pp. 205 - 212.

[6]   A. Spink, M. Park, B. J. Jansen, and J. Pedersen, "Multitasking during web search sessions," *Information Processing & Management,* vol. 42, pp. 264-275, 2005.

[7]   A. Spink, J. Bateman, and B. J. Jansen, "Searching Heterogeneous Collections on the Web: Behavior of Excite Users," *Information Research,* vol. 4, pp. 317-328, 1998.

[8]   D. He, A. Göker, and D. J. Harper, "Combining Evidence for Automatic Web Session Identification," *Information Processing & Management,* vol. 38, pp. 727-742, September 2002.

[9]   H. C. Özmutlu and F. Cavdur, "Application of automatic topic identification on Excite Web search engine data logs," *Information Processing & Management,* vol. 41, pp. 1243-1262, 2005.

[10]   S. Özmutlu and F. Cavdur, "Neural network applications for automatic new topic identification," *Online Information Review,* vol. 29, pp. 34-53, 2005.

[11]   S. Y. Rieh and H. Xie, "Analysis of multiple query reformulations on the web: The interactive information retrieval context," *Information Processing & Management,* vol. 42, pp. 751-768, 2006.

[12]   A. Spink and B. J. Jansen, *Web Search: Public Searching of the Web*. New York: Kluwer, 2004.

[13]   P. Wang, M. Berry, and Y. Yang, "Mining Longitudinal Web Queries: Trends and Patterns," *Journal of the American Society for Information Science and Technology,* vol. 54, pp. 743-758, 2003.

[14]   A. Spink, B. J. Jansen, C. Blakely, and S. Koshman, "A Study of Results Overlap and Uniqueness Among Major Web Search Engines," *Information Processing & Management,* vol. 42, pp. 1379-1391, 2006.

[15]   B. J. Jansen, A. Spink, C. Blakely, and S. Koshman, "Web Searcher Interaction with the Dogpile.com Meta-Search Engine," *Journal of the American Society for Information Science and Technology,* vol. 57, pp. 1875-1887, 2006.

[16]   C. Silverstein, M. Henzinger, H. Marais, and M. Moricz, "Analysis of a Very Large Web Search Engine Query Log," *SIGIR Forum,* vol. 33, pp. 6-12, 1999.

[17]   B. J. Jansen, A. Spink, and T. Saracevic, "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," *Information Processing & Management,* vol. 36, pp. 207-227, 2000.

[18]   N. Belkin, R. Oddy, and H. Brooks, "ASK for Information Retrieval, Parts 1 & 2," *Journal of Documentation,* vol. 38, pp. 61-71, 145-164, 1982.

[19]   B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information Processing & Management,* vol. 42, pp. 248-263, 2005.