Predicting Query Reformulation During Web Searching

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
jjansen@acm.org

Danielle Booth
College of Information Sciences and Technology
The Pennsylvania State University
nephari@gmail.com

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

Abstract

This paper reports results from a study in which we automatically classified the query reformulation patterns for 964,780 Web searching sessions (composed of 1,523,072 queries) in order to predict what the next query reformulation would be. We employed an n-gram modeling approach to describe the probability of searchers transitioning from one query reformulation state to another and predict their next state. We developed first, second, third, and fourth order models and evaluated each model for accuracy of prediction. Findings show that *Reformulation* and *Assistance* account for approximately 45 percent of all query reformulations. Searchers seem to seek system searching assistant early in the session or after a content change. The results of our evaluations show that the first and second order models provided the best predictability, between 28 and 40 percent overall, and higher than 70 percent for some patterns. Implications are that the n-gram approach can be used for improving searching systems and searching assistance in real time.

Keywords

Stochastic process, n-grams, query reformulation, Web queries, Web sessions

ACM Classification Keywords

H.3.3 [1] Information Search and Retrieval – *Search process*

Introduction

Query reformulation is the process of altering a given query in order to improve search or retrieval performance. Belkin and fellow researchers [3] reported that query reformulation assistance may be helpful and improve searching performance. When implemented in real systems, however, searchers seldom utilize this system support, resulting in ineffective and inefficient searches [2]. Some researchers have attempted contextual help to assist in query reformulation [7]; however, searchers can

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-247-4/09/04.

become frustrated with information that is pushed to them by the system [1]. One issue hindering the use of these contextual help systems may be a lack of understanding about when searchers desire system intervention. The cognitive load of information searching in complex contextual situations is high [4]. The retrieval or interjection of assistance into the search process may be too much of a cognitive load, requiring a task switch from focusing on the search process to mentally processing the intervention. Therefore, the searcher may simply ignore any assistance offered.

What if information searching systems could more accurately predict what type of query reformulation the searcher was most likely to implement? What if the system could tell when the searcher was most open to system intervention? What if the system could then offer targeted query reformulation assistance at the most receptive point in the search process? These questions motivate our research.

Prior Research

Query reformulation has been an active area of research in the information retrieval, given that the query is the primary expression of searchers information needs. Rieh and Xie [9] conducted a qualitative analysis of query reformulation using 313 sessions from a Web search engine log. The researchers report three facets to query reformulation (content, format, and resource), with multiple sub-facets of each of these given areas. He, Göker and Harper [5] in automatically classifying queries reformulation, used contextual information from a Reuters transaction log for analysis of Web sessions. Employing a version of the Dempster–Shafer theory in an attempt to identify search engine session boundaries, the researchers also identified a series of query states in order to detect the session start and end states. However, the researchers' focus was on determining the average Web search session duration rather than mapping query reformulation states. He, Göker, and Harper [5] automatically tagged queries, but they did not investigate the prediction of moving from one state to another within a session.

Research Question

The following research question is addressed in this study: *What order of state transition provides the best predictability for query reformulation during Web searching?*

Using a transaction log from a Web search engine, we developed heuristics to classify each query into one of six unique query reformulation states, implemented these heuristics in a program, and executed this program against the entire transaction log. With these results, we could then show the distribution of query reformulation states of the entire dataset. We then developed a probability transition matrix, which provides the percentage of transitions among each of the six query reformulation states.

Research Design

For this research study, we collected the data from the meta-search engine, Dogpile (http://www.dogpile.com/). On 6 May 2005, we collected records of Web searcher – system interactions in a transaction log that represents a portion of the

We first conducted an overall analysis of the dataset, shown in Table 1.

Sessions	964,780	
Queries	1,523,793	
Terms		
Unique	298,796	7.03%
Total	4,250,656	
Terms / query	2.79 sd = 1,54	
Session size		
1 query	691,672	71.64%
2 queries	153,056	15.85%
3+ queries	120,052	12.51%
	964,780	100.0%

table 1. Aggregate statistics from the Dogpile search log.

There were 2,465,145 interactions during the data collection period. Of these interactions, there were 1,523,793 queries submitted by 534,507 users (identified by unique IP address and cookie) containing 4,250,656 total terms. Our classification algorithm identified 964,780 unique sessions. There were 298,796 unique terms in the 1,523,793 queries. These statistics for query length, session length, and term usage are in line with those reported in prior works [c.f., 8, 11].

searches executed on Dogpile[1]. The terminology and procedure that we used in this research is similar to that used in other Web transaction log studies [c.f., 6, 8]. The original transaction log contained 4,056,374 records, with each record containing seven fields:

- *User Identification*: a code to identify a particular computer based on the computer's Internet Protocol address.
- *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.
- *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.

Data Analysis

The transaction log covered a complete 24-hour period with the possibility that certain searchers made several visits to the search engine. Therefore, we had to define a 'session' within the transaction log. Although some researchers have used no boundary or an arbitrary temporal cut-off, we believe that this is inconsistent with reported studies of Web searching sessions [5].

Instead, we used a contextual method to define a session using the searcher's IP address and the browser cookie to determine the *initial query* and *subsequent queries*. We then identified content changes in the sequence of queries for each searcher in the dataset. To implement 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 groups are generally consistent with prior work in query classification [5]. The classifications of query reformulation that we used are defined as:

- *New:* the query is the first query from a unique *User Identification – Cookie* or the query is on a new topic from this searcher. We considered the query on a new topic if there were no terms in common with the previous query from a particular searcher. Although this approach for defining a new topic is not foolproof from a systems viewpoint, a new query is a new execution again the inverted file index. *New* is the first classification applied to the dataset.
- *Assistance:* query generated by the searcher's selection of an *Are You Looking For?* feature. The *Assistance* field was the second classification checked for after New. The *Assistance* field was helpful in illustrating when the searcher sought out assistance from the system.
- *Content Change*: the searcher executed a query on another content collection. The available content collections were *Web, Images, Audio, News,* and *Video*.

[1] We expect to make this Web search engine transaction log available to the research community once the current non-disclosure agreement expires and upon successful negotiation with Infospace.

This was the third condition checked for during data analysis. Although it is possible to change the query and the content collection simultaneous, we could locate no occurrences of it in the data set.

- *Generalization:* the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more general information. We determined a query reformulation to be *Generalization* if the query contained fewer terms than and similar terms to the previous query by a particular user.
- *Reformulation:* the current query is on the same topic as the searcher's previous query, and both queries contain common terms. We determined a query reformulation to be *Reformulation* if the query contained the same number of terms as the previous query by a particular searcher with at least one term being in common to both queries.
- *Specialization*: the current query is on the same topic as the searcher's previous query, but the searcher is now seeking more specific information. We determined a query reformulation to be *Specialization* if the query contained more terms than and similar terms to the previous query by a particular searcher.

Using an automated program that we developed for this research, we classified each query in the database using an approach similar to that used by He, Göker, and Harper [5] to identify temporal sessions in Web searching.

Once we had developed our probability transition matrix, we could then begin calculating the state – transition pattern for each session (i.e., the number of order of states and transitions for a given session) and a hash table for the entire data set (i.e., all session patterns). We used n-grams to develop the probabilistic patterns.

N-grams are a probabilistic modeling approach used for predicting the next item in a sequence and are (n-1) order Markov models, where n is the n of the gram (i.e., sub-sequence or pattern) from the complete sequence or pattern. An n-gram model predicts state x_i using states x_{i-1}, x_{i-2}, x_{i-3}, …x_{i-n}. The probabilistic model is then presented as: $P(x_i \mid x_{i-1}, x_{i-2}, x_{i-3}, \ldots x_{i-n})$, with the assumption the next state only depends on the last $n - 1$ states, which is, again, a (n-1) order Markov model.

Results

We now return to our research question and presentation of the results of our analysis. The state occurrences (zero order model) are shown in Table 2.

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

table 2. Occurrences of query reformulation.

Results concerning the query reformulation states that are most likely to follow one another in Web searching

From Table 3, there appears to be a connection between the searcher shifting content collections and the use of system assistance with a majority (58 percent) of assistance usage occurring just before a content change or just after a content change (41 percent). The use of assistance during these transitions accounted for 25 percent of all assistance usage. Users appear to be somewhat receptive to *Assistance* at the start of the session (21 percent), just after their initial query. As stated previously, we see high occurrences of *Reformulation* after *New* (21 percent) and after *Specialization* (32 percent), with a variety of modification variations on this base pattern. This would indicate that searchers use 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 percent), representing a standard building block methodology of searching. *Specialization* also appears to be a These are probably the searchers who have a well-defined expression of their information need.

are shown in the transition probability matrix (i.e., first order analysis) in Table 3.

	New	**Content Change**	**Reformulation**	**Generalization**	**Generalization w/ reformulation**	**Specialization**	**Specialization w/ reformulation**	**Assistance**	**Total**
New	0%	13%	**21%**	7%	7%	**22%**	9%	**21%**	100%
Content Change	0%	0%	17%	11%	8%	16%	7%	**41%**	100%
Reformulation	0%	11%	0%	14%	18%	19%	**23%**	15%	100%
Generalization	0%	10%	18%	0%	5%	**37%**	12%	18%	100%
Generalization w/ reformulation	0%	6%	**32%**	6%	0%	18%	**27%**	11%	100%
Specialization	0%	9%	**32%**	16%	**22%**	0%	12%	9%	100%
Specialization w/ reformulation	0%	6%	**28%**	14%	**36%**	8%	0%	7%	100%
Assistance	0%	**58%**	12%	7%	11%	5%	8%	0%	100%

table 3. Transition Probability Matrix.

The value in each cell is the probability of going from the row state to the corresponding column state. The most frequently occurring states for each row (i.e., the start state) are bolded, directly answering our research question. By definition, *New* is always the first state in a session. We bolded transition probability that are significantly higher than other transition probabilities.

For results directly addressing our research question (*What order of state transition provides the best predictability for query reformulation during Web searching?*), we refer to Table 4. We analyzed the entire dataset and divided the log file into five separate sub-sets, analyzing each sub-set individually. Results for the entire dataset and each of the subsets were similar, so we report results only for the entire dataset in this manuscript.

Order of Model	**Precision**	**Sessions**	**Actual Unique Patterns**
0	0	965,439	8
1st	0.28	436,647	48
2nd	0.40	120,711	392
3rd	0.44	62,174	2686
4th	0.47	34,451	12,025

table 4. Occurrences of Query reformulation.

Table 4 presents the results from our analyses of the dataset at five different model orders, zero to four (i.e., one to five states respectively). Although some prior work has excluded smaller order sessions of the collected data from model analysis [c.f., 10], we believe it important to apply all models to the entire dataset to obtain an accurate evaluation of the models' performance, applicability, and effectiveness. Additionally, we present the results for the zero order model. Although providing little predictability, the zero order model provides a base line for comparison.

Table 4 shows that the predictably of the models increase but rapidly begins to level off. The first order model (i.e., one state to the predictive the next) has a precision of 28 percent. A second order model (i.e., two states to predict the third) has a precision of 40 percent. From second to third, the increase in precision is 4 percent. The fourth order model has the highest precision (47 percent), providing the best predictability. However, very few sessions (34, 451) had this many states.

Conclusion

Searchers appeared to execute a great deal of *Reformulation* as they try to express more precisely their information need (see table 2). They typically move to narrow their query at the start of a session, moving to *Reformulation* in the mid and latter portions of the sessions. Implications for designing contextual help include that assistance to narrow the query and alternate query terms would be.

There are low rates of implementing system assistance in conjunction with these states. Instead, the most systems assistance usage occurred immediately after *Content Changes*. System *Assistance* (see table 3) use at this state suggest that searchers are more open to system intervention during these content collection shifts. As to why searchers are less likely to implement *Assistance* immediately after query submission – they may be too cognitively focused on correctly expressing their information need to attend to anything else. System assistance should be most specifically targeted to when a searcher is making a cognitive shift (i.e., using a different content vertical) when it appears searchers are open to system intervention.

Acknowledgements

Our sincere thanks to Infospace for providing the data set used for this research project. The Air Force Office of Scientific Research (AFOSR) and the National Science Foundation (NSF) funded portions of this research.

References

[1] Adamczyk, P. D. and Bailey, B. P. If not now, when?: The effects of interruption at different moments within task execution. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2004)* (Vienna, Austria, 24-29 April, 2004).

[2] Anick, P. Using terminological feedback for Web search refinement - a log-based study. In *Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Toronto, Canada, 28 July - 1 August, 2003).

[3] Belkin, N., Cool, C., Kelly, D., Lee, H.-J., Muresan, G., Tang, M.-C. and Yuan, X.-J. Query length in interactive information retrieval. In *Proceedings of the 26th Annual International ACM Conference on Research and Development in Information Retrieval* (Toronto, Canada, 28 July - 1 August, 2003).

[4] Belkin, N., Oddy, R. and Brooks, H. Ask for information retrieval, parts 1 & 2. *Journal of Documentation*, 38, 2 1982), 61-71, 145-164.

[5] He, D., Göker, A. and Harper, D. J. Combining evidence for automatic Web session identification. *Information Processing & Management*, 38, 5 (September 2002), 727-742.

[6] Jansen, B. J. and Pooch, U. Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*, 52, 3 2001), 235-246.

[7] Meadow, C. T., Hewett, T. T. and Aversa, E. A computer intermediary for interactive database searching ii: Evaluation. *Journal of the American Society for Information Science*, 331982), 357-364.

[8] Park, S., Bae, H. and Lee, J. End user searching: A Web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research*, 27, 2 2005), 203-221.

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

[10] Su, Z., Yang, Q. and Zhang, H.-J. A prediction system for multimedia pre-fetching in Internet. In *Proceedings of the Eighth ACM international conference on Multimedia 2000* (Marina del Rey, California, US, 30 October - 4 November, 2000).

[11] Wang, P., Berry, M. and Yang, Y. Mining longitudinal Web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54, 8 2003), 743-758.