Lexical Analysis for Modeling Web Query Reformulation

Alessandro Bozzon[1], Paul - Alexandru Chirita[2], Claudiu S. Firan[2], Wolfgang Nejdl[2]

[1]Politecnico di Milano
Polo Regionale di Como
Via Anzani 42, 22100, Como, Italy
bozzon@elet.polimi.it

[2]L3S Research Center
Appelstr. 9a
30167 Hannover, Germany
{chirita,firan,nejdl}@l3s.de

ABSTRACT

Modeling Web query reformulation processes is still an unsolved problem. In this paper we argue that lexical analysis is highly beneficial for this purpose. We propose to use the variation in Query Clarity, as well as the Part-Of-Speech pattern transitions as indicators of user's search actions. Experiments with a log of 2.4 million queries showed our techniques to be more flexible than the current approaches, while also providing us with interesting insights into user's Web behavioral patterns.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Design, Measurement

Keywords

Query reformulation model, Lexical log analysis

1. INTRODUCTION

The high importance of Web search engines is no longer a doubt for anybody. Many Web behavior modeling attempts have been made, building upon simple log statistics [2], machine learning [3], etc. Yet even though about 50% of the queries are actually reformulation queries [4], experts still have little understanding of users' search patterns.

This paper proposes two advances towards modeling the Web search behavior: First, we suggest to use the variation in Query Clarity [1] as an indicator of user's actions, i.e., generalizing, specializing, or refining the query. Second, we analyze the Part-Of-Speech transitions within reformulation processes and argue that they would make a valuable input for future Web user modeling approaches, in applications such as automatic query expansion, sponsored search, etc. The following section examines the performance of our approaches onto real-life Web search sessions.

2. QUERY REFORMULATION PATTERNS

Collection. We performed our empirical investigation onto an Excite log of about 2.4 million queries sent over 8 hours to the search engine. There were 319,566 search sessions, already delimited by the engine in its log using various heuristics. We further split each session in tasks. Two consecutive queries were assigned to the same task if they contained at least one common non-stopword stem. We found 3.08 average tasks per session, and 2.04 average queries per task. Let us now inspect how our two approaches performed.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

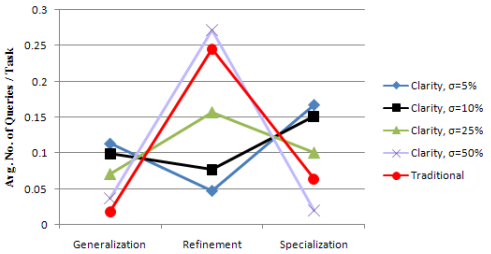

Figure 1: Query reformulation patterns as a function of clarity.

From \ To	New Q.	Navigation (by results pages)	Generaliz.	Specializ.	Refinement	Other (Blank Q., Refresh, ...)	Sum From
New Q.	**30.37**	**6.30**	*2.54*	*4.59*	*2.09*	**7.10**	53.00
Navig.	*2.27*	**16.24**	0.76	*1.40*	0.61	*1.61*	22.91
Gen.	*1.61*	0.80	0.28	0.92	0.33	0.53	4.51
Spec.	*1.83*	*1.39*	*1.28*	0.69	0.55	0.90	6.67
Ref.	*1.05*	0.70	0.38	0.40	0.46	0.42	3.44
Other	*2.09*	*2.27*	0.47	0.78	0.38	*3.42*	9.44
Sum To	39.25	27.73	5.73	8.80	4.45	14.01	100

Table 1: Transitions between query types (%).

Defining Reformulation Types. Traditional approaches model reformulation as a function of the number of keywords per search query: (1) Adding terms is related to specializations, removing terms to generalizations, and substituting them to refinements. We argue that this technique is too shallow, and propose to use *Query Clarity* [1] instead, as an improved indicator of user's actions. We thus build upon the divergence between the language model associated to the query and that associated to the searched collection. In a simplified version, clarity is expressed as follows:

$$Clarity = \sum_{w \in Query} P_{ml}(w|Query) \cdot log \frac{P_{ml}(w|Query)}{P_{coll}(w)} \quad (1)$$

where $P_{ml}(w|Query)$ is the probability of the word w within the query, and $P_{coll}(w)$ is the probability of w within the entire document collection.

From \ To	N.V.Aj.Av	N.V.Aj	N.V.Av	N.Aj.Av	V.Aj.Av	N.V	N.Aj	N.Av	V.Aj	V.Av	Av.Aj	N	V	Aj	Av	U	Sum From
N.V.Aj.Av	*4.45*	0.42	*2.40*	0.67	0.12	0.35	0.14	0.36	0.03	0.06	0.00	0.13	0.03	0.00	0.00	0.02	9.18
N.V.Aj	0.43	0.96	0.31	0.11	0.02	0.95	0.72	0.12	0.13	0.02	0.00	0.65	0.12	0.04	0.00	0.04	4.62
N.V.Av	*2.43*	0.32	*1.47*	0.37	0.07	0.39	0.15	0.27	0.03	0.05	0.00	0.22	0.05	0.00	0.01	0.02	5.85
N.Aj.Av	0.67	0.11	0.37	0.15	0.02	0.09	0.13	0.09	0.01	0.01	0.00	0.09	0.01	0.01	0.00	0.01	1.77
V.Aj.Av	0.12	0.02	0.07	0.02	0.01	0.02	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.31
N.V	0.36	*1.22*	0.45	0.09	0.02	**6.08**	*1.17*	0.23	0.27	0.04	0.00	*4.72*	0.89	0.06	0.01	0.20	15.81
N.Aj	0.15	0.94	0.17	0.16	0.01	*1.12*	*2.97*	0.30	0.26	0.02	0.01	*2.29*	0.15	0.23	0.01	0.11	8.92
N.Av	0.36	0.14	0.29	0.10	0.01	0.22	0.30	0.29	0.02	0.02	0.00	0.40	0.03	0.02	0.01	0.03	2.26
V.Aj	0.03	0.22	0.04	0.02	0.01	0.29	0.29	0.03	0.16	0.01	0.00	0.23	0.10	0.05	0.00	0.03	1.50
V.Av	0.06	0.02	0.06	0.01	0.00	0.05	0.02	0.03	0.01	0.02	0.00	0.03	0.02	0.00	0.00	0.01	0.35
Av.Aj	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.07
N (Noun)	0.14	0.90	0.28	0.11	0.01	**6.20**	*3.11*	0.53	0.29	0.04	0.01	**19.45**	*1.14*	0.29	0.02	*1.41*	33.92
V (Verb)	0.03	0.18	0.06	0.01	0.00	*1.48*	0.23	0.04	0.16	0.03	0.00	*1.35*	0.93	0.06	0.01	0.30	4.88
Aj (Adjective)	0.01	0.06	0.01	0.01	0.00	0.11	0.44	0.02	0.09	0.01	0.01	0.33	0.06	0.32	0.01	0.13	1.62
Av (Adverb)	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.02	0.00	0.01	0.00	0.03	0.01	0.02	0.02	0.01	0.17
U (Unknown)	0.02	0.04	0.02	0.01	0.00	0.32	0.18	0.05	0.04	0.01	0.00	*2.91*	0.61	0.25	0.04	4.29	8.78
Sum To	9.26	5.56	6.00	1.85	0.31	17.67	9.91	2.40	1.53	0.35	0.07	32.85	4.14	1.34	0.16	6.61	100

Table 2: Part-Of-Speech transitions between various lexical compound patterns (in percentages).

In Figure 1 we depict the percentage of queries associated to each reformulation pattern, using both the traditional approach and query clarity with different σ values (we take values residing within the $ClarityOfPreviousQuery \pm \sigma$ as denoting refinement queries, values above that interval marking specialization queries, and values below it indicating generalizations). It is obvious to see that introducing "clarity" allows for more flexibility in determining query reformulation actions. The traditional approach is similar to using clarity with a fairly large σ, about 45% of the clarity value of the previous query. Although the best parameter depends on each application, we believe that such a high σ misses out on identifying the *real* underlying intentions of each user, i.e., generalizing or specializing her query. We thus argue that clarity with σ values around 5-10% would model the Web search behavior much better.

We consider a simplified reformulation model as follows: After the first query was issued, if the output is satisfactory, the user would either navigate through the results or start a new search task. Otherwise, she would attempt to improve her query either by narrowing its focus, or by broadening it. Then, the same steps are taken until a satisfactory output is obtained, or until giving up. In Table 1 we use clarity with $\sigma = 10\%$ to test our assumptions. Most transitions occur either towards a new query, or towards the next page of results for the current search request (i.e., no reformulation). Unlike with traditional approaches (see Figure 1), for clarity the highest amount of reformulations are *specializations*, rather than refinements, which is closer to the intuitive search model outlined above. Also, interestingly, from a specialized query, about 20% of all actions are generalizations, indicating that in some cases users consider to have narrowed their search too much, and thus try to relax it a little bit, again in accordance to the above model.

Part-Of-Speech Pattern Transitions. Applications such as automatic query expansion could be improved by knowing *which* POS are more likely to be added or removed by each user. We thus analyzed the POS transition patterns in Table 2. As expected, most queries are composed only of nouns (about 33%). It seems that a good amount of these are ambiguous terms, fact indicated by the high transition rate towards N-V queries (about 20% of all noun queries). We believe this is not due to a real addition of verbs, but rather of multi-sense nouns (e.g., "play", which can act both as a noun and as a verb). Moreover, there are very few queries composed exclusively of adjectives, adverbs, or both, and they usually do not get reformulated at all. The same is valid when a verb is added to the above mentioned patterns. Finally, there is a significant amount of queries containing all 4 major POS, about 10%. The major transition patterns from these queries involve either adding even more words, or removing the adjective(s). Interestingly, queries with N-V-Adv are usually further extended, whereas for queries with N-V-Adj the tendency is to remove words. This indicates that users generally consider N-V-Adv queries to be broad, and thus in need of specialization, whereas N-V-Adj are seen as too specific, or perhaps too badly formulated, which demands for less terms.

Applications. Besides providing an understanding of users' search behavior, reformulation models could be employed in a variety of applications. For example, when the user reformulates her request, the engine could automatically infer that the initial query was not successful, and use this information in order to improve the new search results (e.g., by putting more bias on the newly added terms, for specializations, or by learning which POS transition patterns are characteristic to the user, and consequently adapting automatic query expansion to favor these patterns, etc.).

3. CONCLUSIONS AND FURTHER WORK

This paper proposed Query Clarity as an indicator of user's search reformulation actions, and showed it to be more flexible than traditional approaches, which build onto naïve instruments such as the number of keywords. Moreover, we also analyzed the Part-Of-Speech transition patterns over a large search engine log. In further work we intend to deploy our two techniques into applications such as personalized Web search, in order to better adapt the algorithm to each user's search patterns.

4. REFERENCES

[1] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proc. of the 25th Intl. ACM SIGIR Conf. on Research and Development in Inf. Retr.*, 2002.

[2] Rosie Jones and Daniel C. Fain. Query word deletion prediction. In *Proc. of the 26th Intl. ACM SIGIR Conf.*, 2003.

[3] Tessa Lau and Eric Horvitz. Patterns of search: Analyzing and modeling web query refinement. In *Proc. of the 7th Intl. Conf. on User Modeling*, 1999.

[4] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, and Tefko Saracevic. Searching the web: the public and their queries. *J. Amer. Soc. Inf. Sci. Technol.*, 52(3):226–234, 2001.