Query Topic Detection for Reformulation

Xuefeng He[1], Jun Yan[2], Jinwen Ma[1], Ning Liu[2], Zheng Chen[2]

[1] School of Mathematical Science
Peking University
Beijing 100871, P.R.China
{xfhe, jwma}@math.pku.edu.cn

[2]Microsoft Research Asia
5F, Sigma Center, 49 Zhichun Road
Beijing 100080, P.R.China
{junyan, ningl, zhengc}@microsoft.com

ABSTRACT
In this paper, we show that most multiple term queries include more than one topic and users usually reformulate their queries by topics instead of terms. In order to provide empirical evidence on user's reformulation behavior and to help search engines better handle the query reformulation problem, we focus on detecting internal topics in the original query and analyzing users' reformulation to those topics. Particularly, we utilize the Interaction Information (II) to measure the degree of one sub-query being a topic based on the local search results. The experimental results on query log show that: most users reformulate query at the topical level; and our proposed II-based algorithm is a good method to detect topics from original queries.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Query formulation

General Terms
Algorithms, Experimentation

Keywords
Query reformulation, Topic, Interaction Information

1. INTRODUCTION
Query, which consists of a group of keywords, is playing a key role in the search procedure. Most of previous automatic algorithms were designed by analyzing the terms in queries. According to our observation in the query log, most users reformulate queries following a pattern: the sub-query they choose to delete, or replace, or preserve can be considered as a meaningful topic. Taking query "music video hip hop" as an example, it includes two obvious topics "music video" and "hip hop". Most users change "hip hop" to other queries such as "r&b" or "folk" etc. And some users change the original query to "music video" or "music video hip hop download". Few people change "hip hop" to "hip r&b" in the next step. A topic is usually included as a sub-query in a user's query and it has strong impacts on the quality of search results. The intuition behind this is that search engines will retrieve more relevant results to meaningful topics than those to un-meaningful ones. As for most multiple term queries, they usually include more than one topic. Thus how to detect the topics of a user query and possible refine the query at the topic level will be very important for a search engine's query reformulation success.

2. PROBLEM FORMULATION
Our motivation is to detect the possible topics contained in the original query and analyze the relation between user reformulating behaviors and those topics. To formulate our problem, some basic mathematical definitions are as following:

1. Given a query $q = t_1 t_2 \cdots t_n$ of n terms, where $t_i \ (1 \le i \le n)$ is the ith term of q;
2. Any subsets of $\{t_1, t_2, \ldots, t_n\}$, i.e. any combinations of $t_1 t_2 \cdots t_n$ are defined as sub-queries. The set of sub-queries is defined as $SQ = \{sq_k\}, 1 \le k \le 2^{n+1} - 2$;

Our goal is to detect several sub-queries which could be possible topics from SQ. Then we analyze the user reformulating behaviors to those topics.

3. QUERY TOPIC DETECTION
3.1 Local Search Result
A query typically contains only a few terms, which provide limited information. One straightforward method is to submit a query to a search engine to get the top ranked search pages. Those retrieved results provide some richer information about the query [1]. In other words, we call the retrieved results of query as the local information of this query. Meanwhile, a query has its global information, based on the whole corpus, to provide more information. However, the global based approach can cause high computational complexity and it was shown in [2] that a local based approach outperforms the global based approach. So in this paper, the top ranked search results are utilized to enrich the query.

Given the search results for a query, we need to decide what features should be extracted from the search engine to construct the enrichment. Generally, three kinds of features are considered: the title of a page, the snippet generated by the search engine, and the full plain text of a page [3]. In this paper, we define the top N ranked snippet retrieved by search engine as the local search result of query.

3.2 Interaction Information (II)
For m events $x_{1i}, x_{2j}, \ldots, x_{ml}$, we can get the interaction information (II) between them $I(x_{1i}; x_{2j}; \ldots; x_{ml})$ by:

$$I(x_{1i}; x_{2j}; \ldots; x_{ml}) = \sum_{i,j,\ldots,l} p(x_{1i}, x_{2j}, \ldots, x_{ml}) \left(\sum_{|\tau'| \subseteq |\nu'|} (-1)^{|\nu'|-|\tau'|} \log(p(\tau')) \right),$$

where $\nu' = (x_{1i}, x_{2j}, \ldots, x_{ml})$, and τ' is any subsets of ν'. $|\nu'|$ means the number of elements in ν', and the same with $|\tau'|$. In this way, the interaction information between $I(x_{1i}; x_{2j}; \ldots; x_{ml})$ is defined like this:

$$I(x_{1i}; x_{2j}; \ldots; x_{ml}) = \sum_{|\tau'| \subseteq |\nu'|} (-1)^{|\nu'|-|\tau'|} \log(p(\tau')).$$

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

3.3 Topic Detection

As for a sub-query generated by the original query, the Interaction Information between its terms can be used to measure the information bound up in those terms. The more information bound up one sub-query has, the higher possibility to be a topic it has. Before that, we should build a probability space. And in this space, each term has probability and joint probability with other terms. Because one sub-query is the joint of terms, the joint probability of several terms can be equal to the probability of the sub-query. In our approach, the probability space of one query is built based on the local results of all sub-queries. Given a query $q = t_1\ t_2 \cdots t_n$, where $t_i\ (1 \le i \le n)$ is the ith term of q, the main steps of our approach are as following:

Step1. Get the set of all sub-queries, $SQ = \{sq_k\}, 1 \le k \le 2^n - 2$, where $1 \le k \le 2^n - 2$ is the number of all sub-queries and $2^n - 2 = C_n^1 + C_n^2 + \cdots + C_n^{n-1}$.

Step2. Enrich each sub-query by submitting it into search engine and get the top N ranked snippets. In this way, we can get $(2^n - 2) \cdot N^*$ snippets for query q. The snippet set is defined as $C(q) = \{sn_i\}, 1 \le i \le (2^n - 2) \cdot N^*$, where sn_i is the ith snippet. $N^* = \min(N, N')$, where N' is the actually number of retrieved snippets by search engine for each query.

Step3. The probability of sub-query in $C(q)$ is defined as: $p(sq_k) = \dfrac{\sum_{i=1}^{(2^n-2)\cdot N^*} (sq_k \text{ can be found in } sn_i)}{(2^n - 2) \cdot N^*}$, where sq_k is the probability of occurrence of sq_k in collection.

Step4. As for sq_k, we get its sub-query set, which is defined as: $SQ_k = \{sq_{k1}, sq_{k2}, \ldots, sq_{k2^k-2}\}$. According to the theory of Interaction Information we mentioned in Section 4.1.2, the II value of sq_k is:

$$I(sq_k) = \sum_{|sq_{kj}| \subseteq |sq_k|} (-1)^{|sq_k| - |sq_{kj}|} \log\left(p(sq_{kj})\right).$$

Step5. Order all sub-queries in a descendant value of Interaction Information and split it into two parts by the threshold of zero. In other words, the first part includes the sub-queries with positive II and the last part includes the sub-queries with negative II.

Step6. Finally, the list of topic list we detected from SQ is defined as: $topic\ list = \{T_1, T_2, \ldots T_i, \ldots, T_M\}$, where T_i is one sub-query which has positive Interaction Information and M is the total number, and $I(T_i) >= 0, I(T_i) >= I(T_{i-1}),\ \ 1 \le i \le M$.

4. EXPERIMENTS

Our whole data set, including 9,621,160 query pairs, comes from the query log of a common used commercial search engine. Each query pair contains two queries, where the first one is the original query and the other one is the query reformulated by a user in a session. Corresponding to three categories of reformulations, three data sets are extracted, where Set A contains query term deletion, Set B contains query term substitution, and Set C contains query term expansion. According to the statistical data, the proportions of three categories (A, B and C) are **22%, 31%** and **47%**. As for each query pair (q, p) in data Set A and Set B, q is used to detect topics and p is used to analyze the relation between topics and user behaviors. As for each query pair (q, p) in data Set C, p is used to detect topics and q is used to analyze the relation between topics and user behaviors.

After detecting topics, three precisions related to the query reformulation are defined as following:

$$P(A) = \frac{\#\ user\ delete\ topic\ or\ preserve\ topic\ in\ the\ original\ query}{|A|},$$

where $P(A)$ measures the precision of Set A. Similarly, we can measure the precision of Set B, C, and overall. The results of the three Sets and the overall precision are shown in Table 1.

Table 1. Results.

	Set A	Set B	Set C	Overall
Precision	0.78	0.82	0.89	0.84

A short summary about the analysis of topics are as following:

1. 78% users choose to preserve topic or delete topic when implement query term deletion. Such as "msn messenger 7.5" to "msn messenger", where "msn messenger" is detected as a topic by our approach;
2. 82% users choose to preserve topic or replace topic when implement query term substitution. Such as "britney spears baby picture" to "cute baby picture", where "britney spears" and "baby picture" are detected as two topics by our approach;
3. During 89% users' reformulated queries, the original queries are still topics. In other word, users still focus on the original query though some other information added;
4. Averagely 84% users reformulate queries at the topic level.

5. CONCLUSIONS

In this paper, we study the query topic detection problem and analyze user query reformulation behaviors based on query log. For topic detection, a probability space is built based on the local search results of its all sub-queries from each original query, and the sub-query's degree of being a topic is measured by utilizing Interaction Information (II). Two contributions in this paper are: (1) Interaction Information is utilized to measure the degree of a sub-query being a topic and a topic detection algorithm is proposed and validated based on the local search results; (2) we analyze user reformulation behavior at the topic level. Our experimental results show that averagely 84% users reformulate queries at the topic level, where the topic is detected by our approach. Meanwhile, it also indicates that our II-based approach is a good method to detect topics from the original query.

6. REFERENCES

[1] M. Ljosland. Evaluation of Web Search Engines and the Search for Better Ranking Algorithms. Presented at SIGIR 99 Workshop on Evaluation of Web Retrieval, University of California, Berkeley, 1999.

[2] http://www.iprospect.com

[3] D. Shen, J. T. Sun, Q. Yang and Z. Chen. Building bridges for web query classification. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 131-138, New York, NY, USA. ACM Press, 2006.