A new query expansion method based on query logs mining[1]

Zhu Kunpeng, Wang Xiaolong, Liu Yuanchao
School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
Email:{kpzhu, wangxl, ycliu}@insun.hit.edu.cn

Abstract:

Query expansion has long been suggested as an effective way to improve the performance of information retrieval systems by adding additional relevant terms to the original queries. However, most previous research has been limited in extracting new terms from a subset of relevant documents, but has not exploited the information about user interactions. In this paper, we proposed a method for automatic query expansion based on user interactions recorded in query logs. The central idea is to extract correlations among queries by analyzing the common documents the users selected for them, and the expanded terms only come from the associated queries more than the relevant documents. In particular, we argue that queries should be dealt with in different ways according to their ambiguity degrees, which can be calculated from the log information. We verify this method in a large scale query logs collection and the experimental results show that the method makes good use of the knowledge of user interactions, and it can remarkably improve search performance.

Keywords:

Query expansion, log mining, information retrieval, search engine,

1. Introduction

With the rapid growth of information on the World Wide Web, more and more users need search engine technology to help them exploit such an extremely valuable resource. Although many search engine systems have been successfully deployed, the current search systems are still far from optimal because of using simple keywords to search and rank relevant documents. A well-known limitation of current search engine systems is the difficulty of dealing with synonymy (different words for describing the same things) and

[1] Supported by National Natural Science Foundation of China (60435020, 60673037) and The National High Technology Research and Development Program of China (2006AA01Z197, 2007AA01Z172)

polysemy (same word to describe different things). For example, a farmer may use query "苹果" to get relevant information about the fruit, while computer lovers may use the same query to find related results of this brand computer. When such a query is issued, it is difficult for search engine system to choose which information he/she wishes to get. Another problem of search engines is that web users typically submit very short queries to search engines and the average length of web queries is less than two words (Wen J. R. 2001). Short queries do not provide sufficient indications for an effective selection of relevant documents and thus negatively affect the performance of web search in terms of both precision and recall.

To overcome the above problems, researchers have focused on using query expansion techniques to help users formulate a better query. Query expansion is a method for improving the effectiveness of information retrieval through the reformulation of queries by providing additional contextual information to the original queries. It has been shown to perform very well over large data sets, especially with short input queries (Kraft R. 2004, Carmel D. 2002). However, previous query expansion methods have been limited in extracting expansion terms from a subset of documents, but have not exploited the information about user interactions. Anyone who uses search engines has accumulated lots of click through data, from which we can know what queries have been used to retrieve what documents. These query logs provide valuable information to extract relationships between queries and documents, and which can be used in query expansion. Another problem of current query expansion is that most proposed methods are uniformly applied to all queries. In fact, we think that queries should not be handled in the same manner because we find that there is no need for expansion on some queries. This has also been found in (Dou Z. C. 2007). For example on the query "Google", almost all of the users are consistently selecting results to redirect to Google's homepage, and therefore none of the expansion strategies could provide significant benefits to users. In this paper, we suggest a new query expansion method based on the analysis of user logs. By considering if queries should be expanded and mining correlations among user queries from user logs, our query expansion method can achieve significant improvements in retrieval effectiveness compared to current query expansion techniques. The reminder of this paper is structured as follows. Section 2 is a discussion of previous works for query expansion method. Section 3 introduces a whole procedure of our query expansion method step by step. Section 4 shows empirical evidence of the effectiveness of our method and investigates the experimental results more in detail. Finally, Section 5 summarizes our findings.

2. Query Expansion Based on Relevance feedback

There have been many prior attempts on query expansion. In this paper, we focus on the related work doing query expansion based on relevance feedback (Rocchio J. 1971, Salton G. 1990) information. In this approach, the results returned for the initial query will be marked as relevant or irrelevant according to user's information need, expansion terms can be extracted from the relevant documents. First approaches were explicit (Rocchio J. 1971, Okabe M. 2005) in the sense that the user was the one choosing the relevant results, and then various methods were applied to extract new terms related to the query and the selected documents. Unfortunately, in a real search context, users usually are reluctant to make the extra effort to provide such relevance feedback information (Kelly D. 2003). To overcome the difficulty due to the lack of sufficient relevance judgments, an automatic feedback technique called pseudo-relevance feedback (also known as blind feedback) is

commonly used. This method made a conjecture that, in the absence of any other relevance judgment, the top few documents retrieved on the basis of an initial query are relevant (Attar L. 1977, Croft W.B. 1979). Expansion terms are extracted from the top-ranked documents to formulate a new query for a second cycle retrieval (Lam-Adesina M. 2001, Carpineto C. 2001). However, the method of pseudo-relevance feedback is highly dependent on the quality of the documents retrieved in the initial retrieval. In cases where the top ranked documents retrieved have little relevance to the query, this method will not work well and it may even introduce irrelevant terms into the question and degrade the performance.

Another group of relevance feedback technique is implicit feedback, in which an IR system can make inferences about relevance from searcher interaction, removes the need for the users to explicitly indicate which documents are relevant (Kelly D. 2003, Morita M. 1994). Several previous studies have shown that implicit information may be helpful for inferring user information need and can improve retrieval accuracy through query expansion. Some query expansion methods based on implicit feedback have been proposed in (Cui H. 2003, Lv Y. H. 2006), the implicit information they used is click-through data collected over a long time period in query logs. These query logs provide valuable indications to understand the kinds of documents the users intend to retrieve by formulating a query with a set of particular terms, and expansion terms can be selected from the sets or the results of past queries. One important assumption behind these methods is that the clicked documents are relevant to the query. This presumption is not always right. However, although the clicking information is not as accurate as explicit relevance judgment, the user's choice does suggest a certain degree of relevance. In fact, users usually do not make the choice randomly. Even if some of the document clicks are erroneous, we can expect that most users do click on documents that are relevant. Some previous work on using query logs also strongly supports this assumption (Bar-Yossef 2008, Wen 2002, Billerbeck 2003 and Zhang 2006). Therefore, query logs can be taken as a very valuable resource containing abundant relevance feedback data. In this paper, we present a new query expansion method based on query logs mining, at the same time, in order to avoid the problem of query drift, we utilize clicked results of the present search process as another type of implicit feedback information to deduce users' information need.

Our work differs from the existing ones in two important aspects. First, we introduce a method to evaluate the quality of user queries, which can be measured by the calculation of Kullback-Leibler Distance (Cover T. 1991) among documents in query logs. Query expansion can strongly improve the performance of short queries and ambiguous queries. But this technique can not achieve the same goal on an accurate query; some new added terms will introduce the problem of query drift and degrade the performance. So, we believe that queries should not be dealt with in the same way and measurement of query quality is essential to judge if a query need to be expanded, which has never been researched before. Second, we propose a new query expansion method based on query logs, relevant expansion terms are selected from the past queries with the analysis of relation between queries and documents under the language modeling framework. Comparing to the existing work, the difference is that we extract the terms from the past queries more than the relevant documents, the experiments show that our method gets better performance in some aspects.

3. Query Expansion Method Based on Logs mining

The query expansion method based on logs mining presented in this paper is composed of

two parts: measurement of query quality with ambiguity analysis and terms expansion with query log mining. In this section, the details of these two parts are described.

3.1 Measurement of Query quality with Ambiguity analysis

A good query should be general enough to cover all relevant documents and specific enough to select only relevant ones. But this rule can not be used to evaluate the quality of user queries because the relevant documents are unknown in advance. In fact, many of the queries need to be expanded for their ambiguity, such as the query of "苹果" mentioned above. In this study, we proposed a new method to measure quality of a query based on the calculation of its ambiguity degree, and query logs are adopted as the data resource.

In query logs, the original form of each click-through record is described as:

$$record =< session\_id >< query\_text >< rank >< order >< page\_url >$$

$Session\_id$ is a unique value assigned by the search engine to identify a query task, $rank$ is the document order in all returned results, and $order$ is the order in clicked documents. Our first task is to extract query sessions from the original log data. A query session is formed by the records with the same session id, which can be defined as follows:

$$session =< query\_text > [clicked\ documents]*$$

Each session contains one query and a set of documents which the user clicked on. Because most of queries are repeated, that means one $query\_text$ can correspond to one or more sessions. The central idea of our method is that, for the same query, the clicked documents in the same session should be related with each other and similar in content, but those in the different sessions are not necessarily related. For example, the clicked documents of query "mouse" may be about rodents or computer devices, these two types of documents are not related for the query ambiguity. So the content differences of the clicked documents among the query sessions can be used to measure the ambiguity degree of a query.

In our method, we assume that the clicked documents in the same session were related, which can be regarded as to be generated by one language model. The calculation of ambiguity degree can be considered as an evaluation of Kullback-Leibler Distance (KLD) among these language models. KLD is often used to measure the divergence of two probability distributions in Information Theory, and it is also can be used to evaluate the irrelevant degree between two language models. Given a query q, we can get a collection of sessions from log data denoted by $S(q) = \{s_1, s_2, \cdots, s_n\}$, each session will be represented by a sequence of the clicked documents, $s_i = \{d_1, d_2, \cdots, d_m\}$. The Inner ambiguity degree of a query is $IA(q)$, then:

$$IA(q) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\forall j \neq i} \frac{KLD(p(s_i) \| p(s_j)) + KLD(p(s_j) \| p(s_i))}{2} \quad (1)$$

That is the average divergence of the sessions. $p(s_i)$ is the probability distribution of the language model which is used to generate the document set s_i, and $KLD(p(s_i) \| p(s_j))$ is defined as:

$$KLD(p(s_i) \| p(s_j)) = \sum_t p(t \mid s_i) \log \frac{p(t \mid s_i)}{p(t \mid s_j)} \quad (2)$$

In order to compute the score of formula (2), we need to be able to estimate the value of $p(t \mid s_i)$, which is the conditional probability of occurrence of word t in s_i. The estimate for $p(t \mid s_i)$ is:

$$p(t \mid s_i) = \alpha \sum_{i=1}^{m} \lambda_i P(t \mid d_i) + (1-\alpha) P(t \mid S) \tag{3}$$

Where α is the interpolation weight determined empirically to smooth the language models, so that non-zero probability can be assigned to terms that do not appear in a given document. $P(t \mid S)$ is the global background collection model. λ_i is a weighting parameter determined by the rank of d_i in the clicked documents, and $P(t \mid d_i)$ is the maximum likelihood estimate of the probability of term t under the term distribution for document d_i. The values of λ_i and $P(t \mid d_i)$ can be calculated by the following formulas:

$$\lambda_i = \frac{0.5}{n} + \frac{i}{n(n-i)} \tag{4}$$

$$P(t \mid d_i) = \frac{tf(t, d_i)}{|d_i|} \tag{5}$$

Here $tf(t, d_i)$ is the raw term frequency of term t in document d_i and $| d_i |$ is the total number of terms in the document. We also give an outer ambiguity which comes from the idea of (Cronen-Townsend S., 2002). They use the concept of "clarity score" to quantify the query's ambiguity, which is the relative entropy between a query language model and the corresponding collection language model. The outer ambiguity of the query can be defined as the reciprocal of "clarity score":

$$OA(q) = \frac{1}{\sum_{t \in V} P(t \mid q) \log_2 \frac{P(t \mid q)}{P_{coll}(t)}} \tag{6}$$

According to the above formulas, we can compute the ambiguity degree for a given query.

$$A(q) = \beta OA(q) + (1-\beta) IA(q) \tag{7}$$

And β is the adjusted parameter. Inner ambiguity degree represents the difference between the related documents of the query. Intuitively, if a query is clear, the clicked documents in its sessions will be focused on the same topic, and the term distributions on these documents should be approximately similar. And outer ambiguity degree represents the difference between the related documents and global documents collection. Therefore, the ambiguity degree of a clear query is smaller than an ambiguous one's. In our test, we set $\beta = 0.4$, because we think inner ambiguity degree is more important for the calculation. We will normalize the value of ambiguity degree from 0 to 1, and give a max length of query expansion, named θ, and use $|A(q) \times \theta|$ to set the number of query expansion terms. The idea is that if a query is more ambiguous, more terms should be added for expansion, and if a query is more clarity, fewer terms should be added in order to avoid importing the irrelevant words.

3.2 Query expansion with Logs mining

There are two steps in our approach to expand an ambiguous query. The first step is to get the candidate terms from the associated queries, and the second step is to determine which candidate words should be added to the new query. In this section, the detail of these two steps will be described.

In the first step, we will use the information of clicked documents to create the correlations of the queries. Generally, we assume that a query is relevant with the documents that the user clicked on, and each record of log data suggests such a relationship. If two queries are related with the same clicked documents, we believe these two queries are associated with each other in some way, and the terms in the associated queries can be used as the candidate terms for query expansion. Here, we used the conditional probability $P(q_j \mid q_i)$ to calculate the correlation between q_i and q_j.

$$P(q_j \mid q_i) = \frac{P(q_j, q_i)}{P(q_i)} = \frac{\sum_{\forall d_k \in D} P(q_j, q_i, d_k)}{P(q_i)} = \frac{\sum_{\forall d_k \in D} P(q_j \mid q_i, d_k) \times P(q_i, d_k)}{P(q_i)} \tag{8}$$

Here we support that $P(q_j \mid q_i, d_k) = P(q_j \mid d_k)$, because the relation of queries is created by the document, so d_k separates q_i from q_j, and we get following formula:

$$P(q_j \mid q_i) = \frac{\sum_{\forall d_k \in D} P(q_j \mid d_k) \times P(d_k \mid q_i) \times P(q_i)}{P(q_i)} = \sum_{\forall d_k \in D} P(q_j \mid d_k) \times P(d_k \mid q_i) \tag{9}$$

In formula (9), $P(d_k \mid q_i)$ is the conditional probability when query is q_i and the clicked document is d_k; $P(q_j \mid d_k)$ is the conditional probability when the clicked document is d_k and the query is q_j. The two conditional probability can be estimated by following:：

$$P(d_k \mid q_i) = \frac{f(q_i, d_k)}{f(q_i)} \tag{10}$$

$$P(q_j \mid d_k) = \frac{f(q_j, d_k)}{f(d_k)} \tag{11}$$

$f(q_i, d_k)$ is used to describe the co-occurrence frequency of query q_i and document d_k in log data. $f(q_i)$ is the frequency of q_i in log data. $f(q_j, d_k)$ is used to describe the co-occurrence frequency of query q_j and document d_k in log data. $f(d_k)$ is the frequency of d_k in log data.

By the calculation of the frequency, we can get the collection of related queries of q_i, and the terms in the queries can be used for query expansion. The weights of terms can be calculated by following formula:

$$P(t \mid q) = \sum_{\forall q_i \ s.t.\ t \in q_i} P(q_i \mid q) \tag{12}$$

In the second step, we will sort the expanded terms by their weights and the number of the terms will be set $\lfloor A(q)\times\theta \rfloor$. We set $\theta = 40$ based on experience. The top $\lfloor A(q)\times\theta \rfloor$ terms will be used for query expansion.

4. Evaluations and Analysis

4.1 Experimental Data and Methodology

Due to the characteristics of our query expansion method, we can not conduct experiments on standard test collection such as the TREC data since they do not contain user logs that we need. We test our method on a dataset collected from the query logs of Sogou(搜狗) (www.sogou.com) search engine. It covers one mouth log data and about 80% of the queries in it contain Chinese words. Approximately 24 million query records and 3 million distinct queries are identified.

We select two hundred test input queries "randomly" according to the overall frequency distributions and extract about one million query sessions from the log data. With respect to documents set, we collect about ten thousand pages from the Internet according to the records in query logs to form the test corpus. In this data set, each document has been retrieved and viewed by users with a certain query, and we can get sufficient click-through information to expand a query with our method.

In order to demonstrate the effectiveness of our method, three experiments were carried out. The first is to investigate the correlation between the query lengths and the ambiguity degrees. In the second, we extract ten queries from the queries set and the performance of query expansion on these queries will be illustrated. At last, the experimental results of our query expansion method will be compared with other systems.

4.2 Results

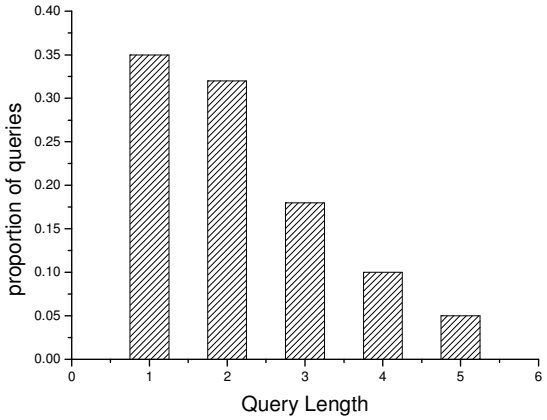

Fig 1. Distribution of query lengths

Figure 1 illustrates the distribution of query lengths according to the number of words. In our experiment, we notice that 35% of the queries contain only one keyword and 32% of the queries contain two keywords. The average length of all queries is 2.18. The result

shows that most people like to use short queries to retrieve information. We do not select the queries contained more than 5 words, because these queries are seldom used and we can not get enough log data for calculation.

Figure 2 illustrates the relation between the query lengths and the statistical analysis values of their ambiguity degree. Let the ambiguity of query q_i is $a_i = A(q_i)$, then the average $\overline{a}$ and the variance σ can be defined as:

$$\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i \tag{13}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(a_i - \overline{a})^2}{n}} \tag{14}$$

We observe that the average values of short queries are higher than the ones of long queries. This verifies that the short queries are more ambiguous than the long queries and the query expansion technique should be applied on short queries more than long queries. The results also approve the effectiveness of our method to measure the ambiguity of queries. But it should be emphasized that not all short queries are bad queries. The variance analysis proves that query length is not a better criterion to measure the quality of queries. The variance is often used to describe the deviation of the data from its mean center. We observe that the variance is larger when query length is 2 and 3, which means the ambiguity values of queries in these two groups make a greater fluctuation around their mean value.

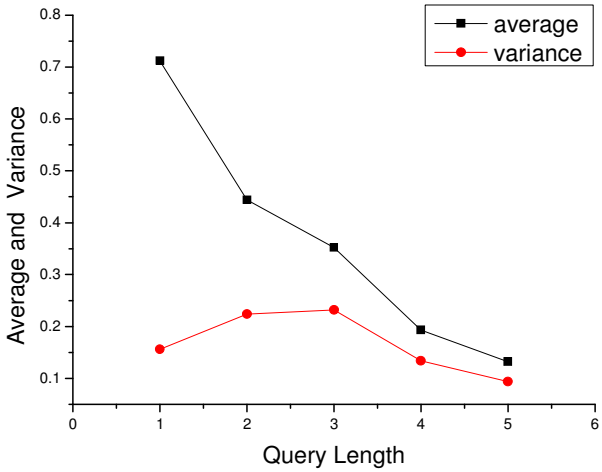

Fig 2. Average and variance analysis of ambiguity degree

In the second experiment, we extract ten queries from queries set which are shown in Table 1 and each query will be divided into both short and long version in order to see how query expansion affects retrieval results on short queries and long queries. In our experiment, the “long queries” come from the queries which length is 4 or 5, and “short queries” only contain one word. After pre-processing documents, including phrasing,

removing stop words and useless characters, we get a thesaurus which contains about sixty thousand words. The results are the precision-recall performance of these queries which will be counted by manual.

ID	Short Queries	Long Queries
1	苹果	苹果 褐斑病 防治
2	成都	成都 旅游 景点
3	足球	足球 过人 技术 视频
4	网易	网易 邮箱 申请
5	比尔盖茨	比尔盖茨 慈善 基金
6	经济	国际 经济 形势
7	DNA	DNA 提取侦破 技术
8	汽车	汽车 保险 计算 方法
9	华为	华为 招聘 信息
10	手机	手机 生产 厂家

Table 1. List of Queries in Both the Long Query Set and the Short Query Set

The retrieval results are shown in Table 2. According to the calculation of ambiguity degree, we believe the queries in Short Queries set are more ambiguous than the queries in Long Queries set, so the average precision of Short Queries set should be lower than the one of Long Queries set. Similar to the retrieval process, query expansion is also affected by the ambiguity of original queries. Compared with an accurate query, the query expansion method can achieve a more improvement on an ambiguous one. The results confirm our expectation just described. Without query expansion, the average precision on Short Queries set is 22.63% which is lower than 28.80% of Long Queries set. The improvement gained with query expansion on Short Queries set is observably higher than that obtained on Long Queries set, and the results show the application of query expansion on Short Queries set is more valuable.

Recall	Short Queries Without QE	Short Queries With QE	Long Queries Without QE	Long Queries With QE
10	45.00	69.33(+54.07)	54.63	67.20(23.01)
20	32.33	52.50(+62.39)	41.72	56.40(35.19)
30	26.50	46.32(+74.79)	35.50	47.54(33.92)
40	23.06	40.67(+76.37)	31.33	42.60(35.97)
50	20.78	38.25(+84.07)	27.75	38.67(+39.35)
60	18.39	35.81(+94.73)	25.17	34.44(+36.83)
70	16.90	32.22(+90.65)	21.30	30.22(+41.88)
80	15.56	28.83(+85.28)	18.52	25.83(+39.47)
90	14.13	26.67(+88.75)	16.67	23.67(+41.99)
100	13.67	25.35(+85.44)	15.40	21.33(+38.51)
Average	22.63	39.60(+74.95)	28.80	38.79(+34.69)

Table 2. Comparison with and without QE on both Long Query Set and Short Query Set

The results in Table 2 also prove that query expansion technique can not achieve the same performance on the accurate queries compared with the ambiguous ones; some new added terms will introduce the problem of query drift and degrade the performance.

In order to evaluate our query expansion method, we will compare its performance not only with that of the original queries, but also with that of local context analysis (LCA) which extracts the expanded terms from the related documents. The results are shown in Fig 3.

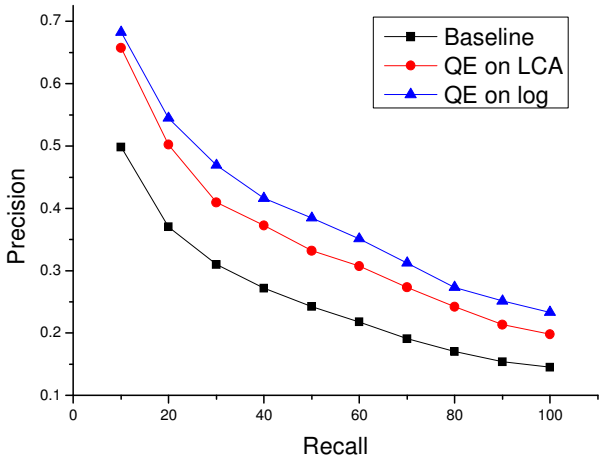

Fig 3. Comparison of query expansion

For local context analysis, we use 30 expansion terms from 100 top-ranked documents for query expansion. The smoothing factor δ in local context analysis is set to 0.1. The experiments showed that query expansion techniques can greatly improve the performance of precision rate and recall rate for information retrieval, especially for the documents collection with a wide range of content. The results also show that the method of query expansion based on query logs gets better performance than other systems. The reason of the poorer performance achieved by QE on LCA is that the initial search results of are unsatisfactory. This situation affects the performance of the expansion algorithm, resulting in irrelevant terms be added to the original query and thus failed to achieve the better results. In our method, the expansion algorithm is based on the mining of a large scale query logs, relevant expansion terms are selected from the past queries with the analysis of relation between queries and documents under the language modeling framework. Our method can availably reduce the situation of expanding irrelevant terms and decrease the bad impact of unsatisfactory initial search results.

5. Conclusions

In this article, we presented a new method for query expansion based on query logs mining. This method aims first to calculate the ambiguity degree of the query by exploiting the user logs. The result can be used to measure the quality of the query and decide the expanded length of the query. And in the next step, we use the information of clicked documents to

create the correlations of the queries, and the high-quality expansion terms are selected from the past queries with the analysis of relation between queries and documents. This is an effective way to avoid the problem of query drift by reducing the irrelevant expansion terms. We tested our method on a data set that is extracted from the real Web environment. A series of experiments conducted on the data set showed that the query expansion method based on query logs mining can achieve substantial improvements in performance. It also outperforms local context analysis, which is one of the most effective query expansion methods in the past. Our experiments also show that query expansion is more effective for ambiguous queries than for clear queries. This also proved that queries should not be dealt with in the same way and measurement of query quality is essential to judge if a query need to be expanded, because some expansion terms can degrade the performance of high-quality queries.

6. References

Wen, J. R., Nie, J. Y., and Zhang, H. J., 2001, Clustering user queries of a search engine. Proceedings of the 10th International World Wide Web Conference, pp. 162-168.

Kraft, R., and Zien, J., 2004, Mining anchor text for query refinement. Proceedings of the. 13th international conference on World Wide Web, pp. 666-674.

Carmel, D., Farchi, E., Petruschka, Y., and Soffer, A., 2002, Automatic query refinement using lexical affinities with maximal information gain. Proceedings of the 25th International ACM SIGIR Conference on research and development in information retrieval, pp. 283-290.

Dou, Z. C., Song, R. H., and Wen, J. R., 2007, A large-scale evaluation and analysis of personalized search strategies. Proceedings of the 16th International World Wide Web Conference, pp. 581-590.

Rocchio, J., 1971, Relevance feedback in information retrieval. In The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313-323.

Salton, G., and Buckley, C., 1990, Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), pp. 288–297.

Okabe, M., Umemura, K., and Yamada, S., 2005, Query expansion with the minimum user feedback by transductive learning. Proceedings Human Language Technology Conference. Empirical Methods in Natural Language Processing, pp. 963-970.

Kelly, D., and Teevan, J., 2003, Implicit feedback for inferring user preference: A Bibliography. ACM SIGIR Forum, 37(2), pp. 18-28.

Attar, L., and Fraenkel, A.S., 1977, Local feedback in full-text retrieval systems. Journal of the Association for Computing Machinery, 24(3), pp. 397-417.

Croft, W.B. and Harper, D.J., 1979, Using probabilistic models of document retrieval without relevance information. Journal of Documentation, 35(4), pp. 285-295.

Lam-Adesina, M., and Jones, G. J. F., 2001, Applying summarization techniques for term selection in relevance feedback. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1-9.

Carpineto, C., De Mori, R., Romano, G., and Bigi, B., 2001, An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems, 19(1), pp. 1-27.

Morita, M., and Shinoda, Y., 1994, Information filtering based on user behavior analysis and best match text retrieval. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 272-281.

Cui, H., Wen, J. R., Nie, J. Y. and Ma, W. Y., 2003, Query expansion by mining user logs. IEEE Transactions on Knowledge and Data Engineering, 15(4), pp. 829-839.

Lv, Y. H., Sun, L., Zhang, J. L., Nie, J. Y., Chen, W., and Zhang, W., 2006, An iterative implicit feedback approach to personalized search. Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 585–592.

Bar-Yossef, Z. and Gurevich, M., 2008, Mining search engine query logs via suggestion sampling. Proceedings of the 34th International Conference on Very Large Data Bases, pp. 54-65.

Wen, J. R., Nie, J. Y., and Zhang, H. J., 2002, Query clustering using user logs. ACM Transactions on Information Systems, 20(1), pp. 59-81.

Billerbeck, B., Scholer, F., Williams, H. E., and Zobel, J., 2003, Query expansion using associated queries. Proceedings of the 12th international conference on Information and knowledge management, pp. 2-9.

Zhang, Z. and Nasraoui, O., 2006, Mining search engine query logs for query recommendation. Proceedings of the. 15th international World Wide Web conference, pp. 1039-1040.

Cover, T. and Thomas, J., 1991, Elements of Information Theory. New York: John Wiley and Sons.

Cronen-Townsend S., Zhou Y., Croft W. B. Quantifying query ambiguity. In Proc. of Human Language Technology, 2002, pp:94--98