

Hierarchical Location and Topic based Query Expansion

Shu Huang¹ Qiankun Zhao^{2*} Prasenjit Mitra¹ C. Lee Giles¹

Information Sciences and Technology¹
Pennsylvania State University
University Park, PA, USA
{shuang, pmitra, giles}@ist.psu.edu

AOL Research Lab²
AOL China
qiankun.zhao@corp.aol.com

Abstract

In this paper, we propose a novel approach to expand queries by exploring both location information and topic information of the queries. Users at different locations tend to have different vocabularies, while the different expressions coming from different vocabularies may relate to the same topics. Thus these expressions are identified as location sensitive and can be used for query expansion. We propose a hierarchical query expansion model, which employs a two-level SVM classification model to classify queries as location sensitive or location non-sensitive, where the former are further classified into same location sensitive and different location sensitive. For the location sensitive queries, we propose an LDA based topic-level query similarity measure to rank the list of similar queries. Experiments with 2G raw log data from CiteSeer and Excite¹ show that our hierarchical classification model predicts the query location sensitivity with more than 80% precision and that the final search result is significantly better than existing query expansion methods.

Introduction

The objective of information retrieval is to provide relevant information to different users out of the overwhelming amount of data according to their searching keywords/queries. Based on the observation that users often issue very short queries, query expansion techniques have been proposed to close the gap between brief expressions and retrieval objectives. Within short queries, the same keywords may be interpreted into different topics for different users, which may also be true for the same users. The basic idea behind query expansion is to add extra keywords to the short queries so that the retrieval objective can be expressed more specifically and accurately. In this way, more accurate retrieval results can be obtained (Billerbeck *et al.* 2003).

Different query expansion techniques proposed in past years (Cai, van Rijsbergen, & Jose 2001)(Park & Ramamohanarao 2004) can be generally grouped into three categories: document-based query expansion, term-based query expansion, and concept-based query expansion. In the document-based approach, top keywords are selected from

relevant documents (in the search results or in the click-through data) to expand the corresponding query keyword. In the term-based approach, similar keywords are selected based on a dictionary, text corpus, or click-through data. In the concept based approach, for each query, all the possible concepts are suggested and users can interactively select the relevant concepts.

We observed that the existing query expansion approaches ignored two important issues. First, users from different locations may have different vocabularies and hence they may refer to the same objects with different query terms, which we identify as *query location sensitivity*. For example, in British English, the term *lorry* represents what *truck* refers to in American English. Also, the term *paddock* has different meanings in Australian English and British English. As a result, the documents of the same topic created by users from different locations differ in their lists of keywords, while the same thing also happens on the queries issued by users from different locations. Therefore, the distribution of keywords in queries will reflect the term usage distribution. Second, the same search keywords may refer to different topics by different users under different context.

Based on the above observations, we propose to combine the information of both location and topics into an efficient methodology for query expansion. First, we show by experiments that some of the queries are location sensitive and others not. Second, for location sensitive queries, we propose two types of expansion strategies: same-location based query expansion and different-location based query expansion. We proposed a hierarchical classification model to classify a new query into different types at two levels (location sensitive versus location non-sensitive, then same location sensitive versus different location sensitive).

For experiments, we use the search log data of Excite, a general search engine; and CiteSeer(Giles, Bollacker, & Lawrence 1998), a search engine focusing on computer science academic documents. IP addresses are used to locate users, while derived topics of each document are used to represent the query objectives. Rather than deterministic document classification, we used the Latent Dirichlet Allocation (LDA) model (Blei, Ng, & Jordan 2003) in which each document is represented as a vector of topics. The similarity of two documents is calculated by the corresponding vectors. Furthermore, the similarity of two queries is determined by

*Work performed while the author was working in PSU
Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹www.excite.com

comparing of the two sets of clicked documents therein. In addition, keywords in top similar queries are added to the initial query as expansion.

In the SVM-based query classification model, we use several query features including location-sensitivity score, length of the query, and the number of retrieved documents before and after query expansion. In addition, the label of each query is selected based on the precision of different types of query expansion approaches. We compared the location-based query expansion with three existing approaches: the default CiteSeer approach, the WordNet based approach, and the corpus based approach, and then evaluated the three sub-strategies of our method against the Excite default approach. Experiments show that: (1) the location based query expansion improves the search results significantly for the location sensitive queries; (2) the precision of the query classification model is more than 80%; and (3) our location and topic based approach is significantly better than other query expansion approaches, especially on general webpage search.

The contributions of this paper are as follows:

- We proposed a novel approach of investigating the location sensitivity of search engine queries and applied this observation in the query expansion process.
- By using features extracted from the search results and the search engine log data, a two-level SVM-based query classification model was constructed to group queries into different classes: the location sensitive classes versus the location non-sensitive class, and the same location sensitive class versus the different location sensitive class.
- For location sensitive queries, we proposed two types of location sensitive query expansion approaches: the different location based query expansion approach and the same location based query expansion approach.
- Extensive experiments with a very large dataset showed the usefulness of the location-based query expansion and the effectiveness of the query classification compared with existing approaches.

Our experiments with CiteSeer indicate that the location sensitivity of queries is not so strong in academic area. With the general search log from Excite, we can see the location-based query expansion strategy performs much better. For general web search, we propose to use all the keywords in the webpage file as the content of the webpage, which works in the same way as the CiteSeer documents in the query expansion model. In addition, when ranking the clicked documents for a given query in the user log, we refer to the default documents ranking by CiteSeer and Excite respectively. This leads to a positive enhancement of the default CiteSeer and Excite performance in the evaluation experiments.

The rest of this paper is organized as follows. Section 2 shows the related work. The problem statement with some preliminaries is presented in Section 3. In Section 4 we explain the details of the query classification and location based expansion strategies. Finally, experimental results and

conclusion are illustrated in Section 5 and Section 6 respectively.

Related Work

Query expansion methodology have shown significant improvements in the effectiveness of information retrieval systems. Existing methods can be categorized into document based methods, term based methods and concept based methods.

Many earlier algorithms with conventional probabilistic retrieval approach are document based (Arasu *et al.* 2001). With this approach, an initial query is executed and a set of documents are returned. Then a set of terms are obtained from the top relevant documents, which are combined with the initial query to generate and return a more relevant set of documents. Cai, & et. al. propose a method based on the divergence of the query, which calculates the relevance of queries according to their distribution in documents (Cai, van Rijsbergen, & Jose 2001). Also probabilistic models, such as Markov Chains, are applied to improve the performance by combining different methods at successive stages (Collins-Thompson & Callan 2005).

In the category of term based methods, term relationship has been widely used. Synonyms, co-occurrence, and WordNet are integrated into one language model to explore the relevance of terms (Cao, Nie, & Bai 2005). Similarly, term relationship and information flow are explored to supplement single terms with term sets (Bai *et al.* 2005). Other methods (Cui *et al.* 2002) (Xue *et al.* 2004) (Billerbeck *et al.* 2003) though propose to obtain relevance of queries by mining click-through data, fail to notice other features of the query.

The concept based methods pay more attention to the user interaction (Fonseca, Golgher, & Pössas 2005). For a short query, the algorithm returns a list of concepts to be selected by the user and then added back to the initial query.

In this paper we propose the first two-level location based query expansion model. The location-based query expansion is superior to other query expansion approaches on location sensitive queries. We broke through the document and the term levels, and explored the semantics embedded in the queries at different granularities. The experiments show that our model is significantly better than other query expansion methods.

Preliminaries and Problem Statement

Search engine log data, also called click-through data, keeps the records of interactions between web users and the searching engine. Merging these user sessions, we can construct a triple graph as Figure 1 shows. By looking at the user sessions in real click-through data from CiteSeer and Excite, we have the following observations:

- Users from different countries issue different queries to represent the same documents.
- Users from different countries use the same queries but with different interpretation.

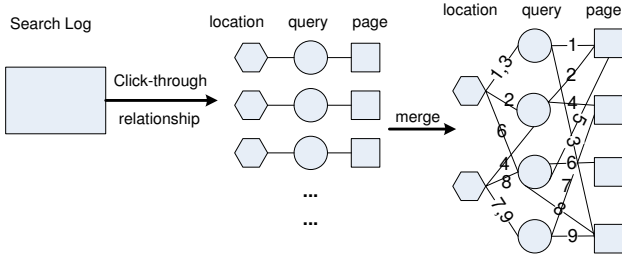


Figure 1: Representation of search log data

Table 1: Different queries, different location, same DocID in Excite log data

Query	Country	Excite DocID
wintv	DE	70007
hauppauge	LU	70007
bruxelles airport	US	84510
zaventem luchthaven	BE	84510
searchitall	CA	5856
search engine	US	5856

The above two observations obtained from the log data motivate our research on investigating the location factor in information retrieval and especially in query expansion. Table 1 shows some examples from the log data.

Given a query q , the default search results returned by the search engine is represented as L_f , which is a list of ranked documents. Suppose that each document is represented as a vector of topics, in which each entry represents the weight of the corresponding topic. Based on the click-through data the similarity between any two queries can be calculated as the cosine similarity of the corresponding summing vector for each query. For example, query q is connected to d_1 , d_2 , and d_3 , query q' is connected to d_2 , d_4 , and d_5 , then the similarity between q and q' is the cosine value between V and V' , where

$$\vec{V} = \vec{d}_1 + \vec{d}_2 + \vec{d}_3 \text{ and } \vec{V}' = \vec{d}_2 + \vec{d}_4 + \vec{d}_5$$

Suppose q comes from location L_1 , q' is from location L_2 , and we use q' to expand q . If $L_1 = L_2$, the expansion is called *same location based query expansion*, else it is called *different location based query expansion*. The *ground-truth* is the list of ranked relevant document extracted from the search engine log data.

To formalize the observations, we define a new concept *location sensitive query* as follow:

Definition 1: Given a query q , suppose the default search result list is L_f , the result list after same location based query expansion is L_s , the result list after different location based query expansion is L_d , and the ground-truth of the results list is L . Query q is defined as a location sensitive query if: $Q(L_s, L) \geq Q(L_f, L)$ or $Q(L_d, L) \geq Q(L_f, L)$, where $Q(L_s, L)$ is the quality of the returned results L_s compared against the ground-truth result L respectively.

Furthermore, if $Q(L_s, L) > Q(L_d, L)$, then q is defined as a *same location sensitive query*; if $Q(L_d, L) > Q(L_s, L)$

then q is defined as a *different location sensitive query*.

The location difference in our experiments is identified at the country level. The quality of the returned results can be measured by different metrics such as Precision, MAP, NDCG, and Tau (Kendall 1995)(Agichtein, Brill, & Dumais 2006).

To identify the location sensitive queries, we define another concept of *location sensitivity score* as follow:

Definition 2: Given a query q , and a list of relevant documents $\{d_1, d_2, d_3, \dots, d_n\}$ as the ground-truth of q . Suppose q is issued from m countries and the set of documents clicked by users from country i is represented as D_i , then the location sensitivity score for query q is defined as:

$$LSS(q) = \sum_{i=1, i \neq j}^m Sim(\sum_{ds \in D_i} \vec{d}_s, \sum_{dt \in D_j} \vec{d}_t)$$

$$0 < s, t \leq n \text{ and } 0 < i, j \leq m$$

The location sensitivity score describes the topic distribution of one query across different countries. A query is not location sensitive or the location sensitivity score is 1, if users across different countries access the relevant documents with the same pattern. Here access pattern refers to the number of times a document was accessed in log data. The location sensitivity score is between 0 and 1. The larger the location sensitivity score is the less location sensitive the corresponding query is.

Figure 2 shows the process of identifying the location sensitive queries and classifying them into different groups for different location-based query expansion strategies. From the search log, we first derive a list of queries to be expanded. Then the query expansion strategies of same location based and different location based are applied. The result quality of different types of query expansion is evaluated and applied to train a two-level SVM classification model. Once the model is constructed, when a new queries comes, we first extract its corresponding features and then predict its location sensitivity using the two-level classification model.

Location-based Query Expansion

In this section, we present the details about constructing the two-level classification model and the location-based query expansion strategies.

The Two-Level Query Classification Model

To construct the query classification model, there are three subtasks: feature extraction, label generation, and model training.

Feature Extraction By mining the log data and using LDA, we extract eight critical features for each query: LSS, NO, NA, NA', NDO, NDA, NDA' and DLD. LSS is location sensitivity score defined above. NO is the number of terms in the original query. NA is the number of terms added to the original query after different location-based query expansion, while NA' is the number of terms added to the original query after same location-based query expansion. NDO is the number of retrieved documents before query expansion. NDA is the number of retrieved documents after different

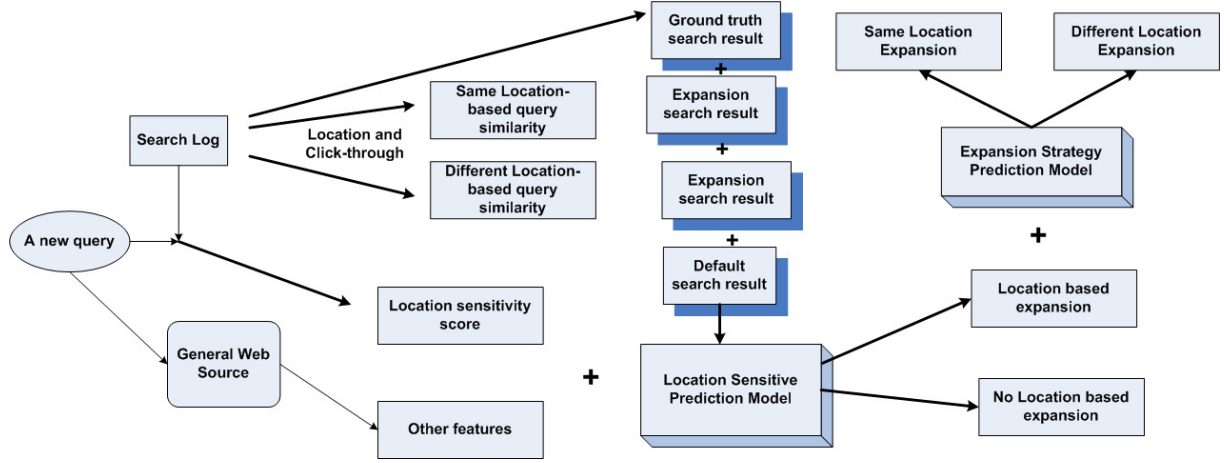


Figure 2: Identify location sensitive queries for different expansion strategies

location-based query expansion, and NDA' is the number of retrieved documents after same location-based query expansion. DLD describes the location diversity of documents related to a query. It is defined as the proportion of the number of documents clicked by users issuing query q to the number of countries where users issuing the query q .

$$DLD|q = \frac{|clicked\ documents\ of\ q|}{|countries\ where\ q\ is\ issued|}$$

Label Generation Two labels of queries are used in SVM modeling: sensitivity label SL and type label TL . SL shows whether the query is location sensitive, and TL shows whether the query is different location sensitive or same location sensitive. To generate the labels, we make three query expansion trials: same location based approach, different location based approach, and ignoring location based approach. In the ignoring location based approach, we only consider the query similarity from topic distribution vectors but ignore the location information.

The labels are decided by comparing the $NDCG$ (Agichtein, Brill, & Dumais 2006) of the three query expansion approaches. Given a query q , SL is set "+1" if a higher $NDCG$ is obtained after different location or same location based approach, and "-1" if the ignoring-location based strategy gets a higher $NDCG$. TL is set "+1" if the $NDCG$ of the different location based approach is higher than the same location based approach, and "-1" otherwise.

Classification Model Training In the location-based query classification, we employ Support Vector Machines (SVM) (Chang & Lin 2007) to generate a two-stage prediction model. After extracting the eight query features and two labels from the log data, we apply them to SVM and a two-stage model is generated to predict which type of query expansion should be applied to a given query.

This two-stage model can classify the queries at two levels. Given a query, if it is predicted as not location sensitive in the first stage model, no query expansion will be applied. But if it is predicted as location sensitive, the second-stage

stage model will be used. According to the second prediction result, the different-location or same-location based expansion will be applied to the query respectively.

Location-Based Query Expansion

In this part, we will illustrate the topic-based document clustering and the application of location sensitivity in location-based similarity measure.

In topic-based document clustering, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan 2003) is applied to generate the topic distribution. As the input of LDA, the document collection contains all the documents reviewed by users in the click-through data. For each query, there is a list of reviewed documents related to it. By processing all the documents with LDA, a topic distribution on all the documents is produced. Each document is associated with a topic vector which specifies the topic distribution of the document.

Based on the topic vector representation, the location-based similarity measure is presented as follow:

For a given query q , there is a list of relevant documents that have been clicked by end users $\langle d_1, d_2, d_3, \dots, d_n \rangle$, where each document d_i is represented as a vector of topic distribution $\langle t_1, t_2, t_3, \dots, t_m \rangle$ generated by LDA. Here m is the number of topics and i is between 1 and n . Then, the similarity between any two queries can be calculated as the cosine similarity between their topic vectors.

By taking into account of the location, we can propose a location-based query similarity measure as follow:

$$Sim(q_1, q_2 | L_1, L_2) = Cos(\sum \vec{d}_i, \sum \vec{d}_j)$$

(d_i is clicked in response to $q_1 | L_1$,
 d_j is clicked in response to $q_2 | L_2$)

That is, for query q_1 from location set L_1 and q_2 from L_2 , their location-based similarity can be measured as the cosine similarity between the vector representations of documents relevant to them and clicked by users from location

sets $L1$ and $L2$ respectively. Suppose the whole set of locations is C , if $L1 = L2$, then $Sim(q1, q2|L1, L2)$ is the same location based query similarity, if $L1 = C - L2$, then $Sim(q1, q2|L1, L2)$ is the different location based query similarity.

Experiments

Dataset

In experiments, we use 2G raw search log data from CiteSeer (130825 queries from 55947 unique IPs) and 129830 queries in search log data from Excite. In experiments on CiteSeer data, 281,379 documents clicked in the user log are used. In experiments on Excite data, 180,150 webpages are used.

Ground truth

In our experiments, the location of queries is identified at the level of country. Each IP address in the user log is mapped to a country. For the documents relevant to queries, we associate each query with the documents clicked by the same user in a time period of maximum thirty minutes. The same user is identified by the same IP address.

We use the clicked documents in the user logs as the baseline. When ranking the clicked documents for a given query, we refer to the default documents ranking by the search engine. This will lead to a positive enhancement to the default CiteSeer/Excite performance in the evaluation experiments, especially the metric of NDCG and Kendall's Tau.

Evaluation

In our experiments, the query expansion methods are evaluated over four widely accepted evaluation metrics: Normalized Discounted Cumulative Gain (NDCG), Precision at K , Mean Average Precision (MAP), and Kendall's Tau. In evaluation, we use short queries which contain no more than three terms. For CiteSeer, we randomly select 30% from all the short queries, which are 3863 in all. For Excite, we randomly select 2400 short queries as test data.

In the experiments with CiteSeer, we compare our location and topic-based method with the default CiteSeer results, the corpus-based query expansion (Mitra & Wiederhold 2002), and Wordnet-based query expansion approaches (Collins-Thompson & Callan 2005). First, to test the precision of the two-level query classification model, we conduct cross-validation on the 3863 queries. Part of queries is randomly selected as training data, and the left queries are used as test data. A group of experiments with increased number of training queries and decreased number of predicting queries are executed. Figure 3 shows the changes of predicting precision with the increase of the number of training queries. As figure 3 shows, our two-stage SVM prediction model has a good precision in predicting the query location sensitivity. The average precision of the first level prediction and second level prediction are 83.78 and 95.76 respectively. In the first level, the maximal and minimal precision are 88.86 and 80.93 respectively; while for the second level prediction, the maximum and minimum are 96.21 and 94.37 respectively.

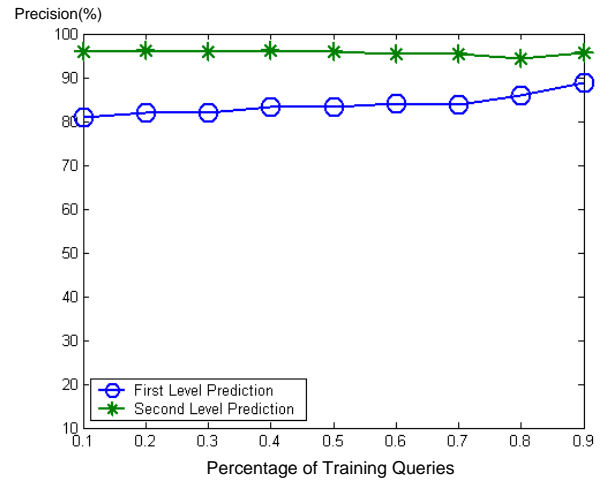


Figure 3: Precision of the two-level query classification model

Table 2: Comparison of four query expansion strategy on location sensitive queries with CiteSeer

Evaluation Strategy	CiteSeer Default Strategy	Topic & Location-based Strategy	Corpus-based Strategy	Wordnet-based Strategy
Precision (%)	19.40	19.79	13.51	14.53
NDCG (%)	32.84	33.34	19.92	23.32
MAP (%)	85.60	92.24	54.72	62.36
Kendall's Tau	-0.923	0.001	-0.935	-0.930

Second, with the ground truth and sensitivity labels, 702 location sensitive queries are picked out from all the 3863 queries. We compare our method with default CiteSeer result, the Wordnet-based query expansion result, and the Corpus-based query expansion result. The strategy of Wordnet-based query expansion method is to expand the initial query with synonyms picked out from the Wordnet (Collins-Thompson & Callan 2005). And the Corpus-based strategy is to calculate the relevance of words based on their context and then expand the initial query with the top relevant words.

Table 2 shows the Precision, NDCG, MAP, and Kendall's Tau of the four strategies on the 702 location sensitive queries. When comparing with default method, the topic and location-based method is 7.7% and 0.923 better in terms of MAP, and Kendall's Tau. The improvements show that, even the ground truth has a positive effect on the default CiteSeer results, the document ranking of our method is much better than default CiteSeer results. Also because the queries in CiteSeer data all aim at academic documents, the effects of location on vocabulary diversity is reduced in some degree.

In experiments with Excite search engine, we compare the default Excite search result with three sub-strategies in our query expansion method. Among the 2400 short queries, 218 are location sensitive, in which 123 queries are different location sensitive and 95 are same location sensitive. In

Table 3: Comparison of sub-strategies of location based method on 118 location sensitive queries with Excite

Evaluation Strategy	Excite Default Strategy	Topic & Different-location Strategy	Topic & Same-location Strategy	Topic & Ignoring-location Strategy
Precision(%)	2.67	3.40	3.40	3.14
NDCG(%)	9.46	20.76	14.25	20.05
MAP(%)	1.13	22.88	12.15	19.49
Kendall's Tau	-0.99	-1.0	-0.973	-1.0

the results, the NDCG values of 1563 queries with the four strategies are all 0. The probable reason is that the index of webpages on Excite search engine changes over time. But with other queries, we can still see the significant improvement caused by our location and topic based method.

For evaluation, we randomly select a sample of 1000 short queries from Excite log, in which 118 location sensitive queries are detected. Table 3 shows the comparison of the four strategies on the 118 location sensitive queries. Because the precision value is calculated at the base of 20(K=20) and the number of clicked webpage for a query is usually 1, so the average precision values shown in Table 3 are close to 5%. On one hand, the boost of precision value shows that our method has more precision; on the other hand, the significant increase of NDCG and MAP values both reflect the improvement of ranking caused by our method. Table 3 shows that our location and topic based method produces significant improvement on general search engine log. Comparing the improvements on Citeseer data and Excite data, it is observed that the query location sensitivity is much more obvious in general webpages than in academic documents.

Conclusions and Future work

We proposed a statistic model of topic and location-based query expansion based on LDA and SVM methods. By clustering the documents to different topics, we scale down the document relevance to the topic relevance, and use the topic relevance to identify the similarity between queries. In addition we make use of the location information to determine whether the query is location sensitive and which type of query expansion should be applied. Our experiments on the Citeseer and Excite data show that on one hand, our model can effectively select the location sensitive queries; on the other hand, for location sensitive queries, our query expansion methods significantly improve the search results.

Acknowledgement

Authors would like to thank Jim Jansen for providing the Excite search engine log Data. This work was supported in part by the National Science Foundation.

References

Agichtein, E.; Brill, E.; and Dumais, S. 2006. Improving web search ranking by incorporating user behavior information. In *Proc. of the ACM Conference on Research and Development in Information Retrieval (SIGIR'06)*.

Arasu, A.; Cho, J.; Garcia-Molina, H.; Paepcke, A.; and Raghavan, S. 2001. Searching the web. *ACM Transactions on Internet Technology (TOIT)* 1(1):43.

Bai, J.; Song, D.; Bruza, P.; Nie, J.-Y.; and Cao, G. 2005. Query expansion using term relationships in language models for information retrieval. In *Proc. of 14th International Conference on Information and Knowledge Management (CIKM'05)*.

Billerbeck, B.; Scholer, F.; Williams, H. E.; and Zobel, J. 2003. Query expansion using associated queries. In *Proc. of 12th International Conference on Information and Knowledge Management (CIKM'03)*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Machine Learning Research* 3:993–1022.

Cai, D.; van Rijsbergen, C. J.; and Jose, J. M. 2001. Automatic query expansion based on divergence. In *Proc. of 10th International Conference on Information and Knowledge Management (CIKM'01)*.

Cao, G.; Nie, J.-Y.; and Bai, J. 2005. Integrating word relationships into language models. In *Proc. of the ACM Conference on Research and Development in Information Retrieval (SIGIR'05)*.

Chang, C.-C., and Lin, C.-J. 2007. *LIBSVM: a Library for Support Vector Machines*.

Collins-Thompson, K., and Callan, J. 2005. Query expansion using random walk models. In *Proc. of 14th International Conference on Information and Knowledge Management (CIKM'05)*.

Cui, H.; Wen, J.-R.; Nie, J.-Y.; and Ma, W.-Y. 2002. Probabilistic query expansion using query logs. In *Proc. of 11th international Conference on World Wide Web (WWW'02)*.

Fonseca, B. M.; Golgher, P.; and Pössas, B. 2005. Concept-based interactive query expansion. In *Proc. of 14th International Conference on Information and Knowledge Management (CIKM'05)*.

Giles, C. L.; Bollacker, K. D.; and Lawrence, S. 1998. Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*.

Kendall, M. G. 1995. *Rank Correlation Methods, Second Edition*. New York: Hafner Publishing Co.

Mitra, P., and Wiederhold, G. 2002. Resolving terminological heterogeneity in ontologies. In *Proc. of Workshop on Ontologies and Semantic Interoperability at the 15th European Conference on Artificial Intelligence (ECAI'02)*.

Park, L. a. F., and Ramamohanarao, K. 2004. Hybrid pre-query term expansion using latent semantic analysis. In *Proc. of the 4th International Conference on Data Mining (ICDM'04)*.

Xue, G.-R.; Zeng, H.-J.; Chen, Z.; Yu, Y.; Ma, W.-Y.; Xi, W.; and Fan, W. 2004. Optimizing web search using web click-through data. In *Proc. of 13th International Conference on Information and Knowledge Management (CIKM'04)*.