Query Dependent Pseudo-Relevance Feedback based on Wikipedia

Yang Xu
Institute of Computing Technology
Chinese Academy of Sciences
Beijing,100190,P.R.China
xuyang@ict.ac.cn

Gareth J. F. Jones
Dublin City University
Glasnevin, Dublin 9
Ireland
gjones@computing.dcu.ie

Bin Wang
Institute of Computing Technology
Chinese Academy of Sciences
Beijing,100190,P.R.China
wangbin@ict.ac.cn

ABSTRACT

Pseudo-relevance feedback (PRF) via query-expansion has been proven to be effective in many information retrieval (IR) tasks. In most existing work, the top-ranked documents from an initial search are assumed to be relevant and used for PRF. One problem with this approach is that one or more of the top retrieved documents may be non-relevant, which can introduce noise into the feedback process. Besides, existing methods generally do not take into account the significantly different types of queries that are often entered into an IR system. Intuitively, Wikipedia can be seen as a large, manually edited document collection which could be exploited to improve document retrieval effectiveness within PRF. It is not obvious how we might best utilize information from Wikipedia in PRF, and to date, the potential of Wikipedia for this task has been largely unexplored. In our work, we present a systematic exploration of the utilization of Wikipedia in PRF for query dependent expansion. Specifically, we classify TREC topics into three categories based on Wikipedia: 1) entity queries, 2) ambiguous queries, and 3) broader queries. We propose and study the effectiveness of three methods for expansion term selection, each modeling the Wikipedia based pseudo-relevance information from a different perspective. We incorporate the expansion terms into the original query and use language modeling IR to evaluate these methods. Experiments on four TREC test collections, including the large web collection GOV2, show that retrieval performance of each type of query can be improved. In addition, we demonstrate that the proposed method outperforms the baseline relevance model in terms of precision and robustness.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search and Retrieval - search process, query formulation

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Information Retrieval, Entity, Query Expansion, Pseudo-relevance feedback, Wikipedia

1. INTRODUCTION

One of the fundamental problems of information retrieval (IR) is to search for documents that satisfy a user's *information need*. Often a query is too short to describe the specific information need clearly. For a long time query expansion has been a focus for researchers because it has the potential to enhance IR effectiveness by adding useful terms that can help discriminate relevant documents from irrelevant ones. For all query expansion methods, pseudo relevance feedback (PRF) is attractive because it requires no user input [21][2][31][3]. PRF assumes that the top-ranked documents in the initial retrieval are relevant. However, this assumption is often invalid [3] which can result in a negative impact on PRF performance. Meanwhile, as the volume of data on the web becomes much larger, other resources have emerged which can potentially supplement an initial search better in PRF, e.g. Wikipedia.

Wikipedia is a free online encyclopedia edited collaboratively by large numbers of volunteers (web users). The exponential growth and the reliability of Wikipedia make it a potentially valuable resource for IR [33]. The aim of this study is to explore the possible utility of Wikipedia as a resource improving for IR in PRF. The basic entry in Wikipedia is an entity page, which is an article that contains information focusing on one single entity. Wikipedia covers a great many topics, and it might reasonably be assumed to reflect the diverse interests and information needs of users of many search engines [28]. With the help of enriched text, we can expect to bridge the gap between the large volume of information on the web and the simple queries issued by users. However, few studies have directly examined whether Wikipedia, especially the internal structures of Wikipedia articles, can indeed help in IR systems.

As far as we are aware, there is little work done investigating the impact of Wikipedia on different types of queries. In this paper, we propose a query-dependent method for using PRF for query expansion, on the basis of Wikepdia. Given a query, we categorize it into one of three types: 1) query about a specific entity (**EQ**); 2) ambiguous query (**AQ**); 3)

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

broader query (**BQ**). Pseudo relevant documents are generated in two ways according to the query type: 1) using top ranked articles from Wikipedia retrieved in response to the query, and 2) using Wikipedia entity pages corresponding to queries. In selecting expansion terms, term distributions and structures of Wikipedia pages are taken into account. We propose and compare a supervised method and an unsupervised method for this task. Based on these methods, we evaluate the effect of Wikipedia on PRF for IR. Our experiments show different methods impact differently on the three types of queries. Thus a query dependent query expansion is necessary to optimally benefit retrieval performance.

The contributions of this work are as follows: 1) we thoroughly evaluate the potential of Wikipedia for IR as a resource for PRF, 2) we explore the use of Wikipedia as an entity repository as well as its internal structure for retrieval, and based on these two aspects, different methods for selecting expansion terms are proposed and compared, and 3) our method is conducted in a query dependent way, which is more effective and robust than a single method.

The remainder of this paper is organized as follows: Section 2 provides a detailed account of related work, Section 3 introduces query categorization based on Wikipedia, Section 4 describes our proposed methods for using Wikipedia as pseudo relevant documents, experimental results are reported in Section 5, and we conclude in Section 6.

2. RELATED WORK

Automatic query expansion is a widely used technique in IR. PRF has been shown to be an effective way of improving retrieval accuracy by reformulating an original query using pseudo-relevant documents from the initial retrieval result [27, 18, 34]. Traditional PRF algorithms such as Okapi [27], Lavrenko and Croft's relevance model [18], and Zhai and Lafferty's mixture model [34] are based on the assumption that the top-ranked documents from an initial search are relevant to the query.

Large amounts of research has been reported on attempts to improve traditional PRF, e.g. using latent concepts [21], query-regularized estimation [29], additional features other than term distributions [3], and a clustered-based re-sampling method for generating pseudo-relevant documents [19].

On the other hand, there has been work on selective query expansion which aims to improve query expansion with decision mechanisms based on the characteristics of queries [11, 6, 32, 13]. These methods share a similar idea, that is, to disable query expansion if the query is predicted to perform poorly. Other relevant work is that of He and Ounis who proposed selecting an appropriate collection resource for query expansion [13].

Cui et al. proposed a query expansion approach by mining user logs [12]. They extracted correlations between query terms and document terms by analyzing user logs, and then used these to select expansion terms for new queries.

A concept-based interactive query expansion method was suggested by Fonseca et al. also using query logs [10]. Association rules are applied to identify concepts related to the current query from the query context derived from user logs. These concepts then provide suggestions to refine the vague (short) query.

Kwok and Chan [17] studied the idea of using an external resource for query expansion. They found that query expansion failure can be caused by the lack of relevant documents in the local collection. Therefore, the performance of query expansion can be improved by using a large external collection. Several external collection enrichment approaches have been proposed [12, 13, 10, 5]. Our work follows this strategy of a query expansion approach using an external collection as a source of query expansion terms.

Recently, Wikipedia[30] has emerged as a potentially important resource for IR, a number of studies have been reported which adopt Wikipedia to assist query expansion, many of these have appeared in the context of the TREC Blog track [24, 35, 7, 8, 16].

One interesting example of the use of Wikipedia in this way is the work of Li et al. [20] who proposed using it for query expansion by utilizing the category assignments of Wikipedia articles. The base query is run against a Wikipedia collection and each category is assigned a weight proportional to the number of top-ranked articles assigned to it. Articles are then re-ranked based on the sum of the weights of the categories to which each belongs. The method shows improvement over PRF in measures favoring weak queries. A thesaurus-based query expansion method using Wikipedia was proposed by Milne, Witten and Nichols [23]. The thesaurus is derived with criteria such that topics relevant to the document collection are included. They propose to extract significant topics from a query by checking consecutive sequences of words in the query against the thesaurus. However, query dependent knowledge is not taken into consideration by the thesaurus. Elsas et al. investigated link-based query expansion using Wikipedia in [7]. They focused on anchor text and proposed a phrase scoring function. Our work differs from these methods, in that, expansion terms are not selected directly from the documents obtained by running the base query on the Wikipedia. Instead Wikipedia entity pages are viewed as a set of pseudo-relevant documents tailored to the specific query.

Similar to the query log based expansion methods, our approach can also reflect the preference of the majority of the users. However, our work differs from these methods in that we try to narrow the gap between the large volume of information on the web and the simple queries issued by users, by mapping queries to Wikipedia entity pages which reflect knowledge underlying the query, rather than using a simple bag-of-words query model.

3. QUERY CATEGORIZATION

In this section we briefly summarize the relevant features of Wikipedia for our study, and then examine the different categories of queries that are typically encountered in user queries and how these can be related to the properties of Wikipedia for effective query expansion.

3.1 Wikipedia Data Set

Wikipedia is a free online encyclopedia which has detailed guidelines governing its content quality and format. Unlike traditional encyclopedias, Wikipedia articles are constantly being updated and new entries are created everyday. These characteristics make it an ideal repository of background knowledge for query expansion.

Wikipedia is organized as one article per topic. Each article thus effectively summarizes the most important information of each topic. A topic in Wikipedia has a distinct, separate existence often referring to a specific entity, such as a person, a place, an organization or miscellaneous. In ad-

Field	Description
Title	Unique identifier for the topic
Overview	Lead section, summary of the topic
Content	Grouped by sections
Category	At least one for each topic
Appendix	References, notes, external reading etc.
Links	Cross-article hyperlinks specified by topics

Table 1: Fields of Wikipedia Articles

dition, important information for the topic of a given article may also be found in other Wikipedia articles [9]. An analysis of 200 randomly selected articles describing a single topic (not including redirect pages and disambiguation pages, described below) showed that only 4% failed to adhere to a standard format. Based on our analysis of Wikipedia page structure, in our work we divide Wikipedia articles into six fields as shown in Table 1.

In addition to topic pages, Wikipedia contains "redirect" pages which provide alternative ways of referring to a topic: e.g. the page *Einstein* redirects to the article *Albert Einstein* which contains information about the physicist. Another useful structure of Wikipedia is its so-called "disambiguation pages" which list the referents of ambiguous words and phrases that denote two or more topics in Wikipedia.

As an open source project, Wikipedia is obtainable via download from http://download.wikipedia.org, which is in the form of database dumps that are released periodically. We downloaded the English Wikipedia dump of January, 2008. Wikipedia articles are stored in their own markup language called *Wikitext* which preserves features, e.g. categories, hyperlinks, subsections, tables, pictures. We index all the pages using the Indri search engine [14]. Preprocessing includes removing Wikipedia pages that are used for management purpose, e.g. those with *Wikipedia* in the title. Wikipedia articles are stemmed using Porter stemming.

3.2 Query Categorization

Queries in web search may vary largely in semantics and the user's intensions they present, in numbers of relevant documents they have in the document repository, and in numbers of entities they involve. In this work, we define three types of queries according to their relationship with Wikipedia topics : 1) queries about a specific entity (EQ), 2) ambiguous queries (AQ), and 3) broader queries (BQ). By EQ, we mean queries that have a specific meaning and cover a narrow topic, e.g. "Scalable Vector Machine". By AQ, we mean queries with terms having more than one potential meaning, e.g. "Apple". Finally, we denote the rest of the queries to be BQ, because these queries are neither ambiguous nor focused on a specific entity. For the first two types of queries, a corresponding Wikipedia entity page can be determined for the query. For EQ, the corresponding entity page is the page with the same title field as the query. For AQ, a disambiguation process is needed to determine its sense. For all three types of queries, a set of top ranked retrieved documents is obtained from an initial search.

Our method automatically categorizes queries, with the help of titles of Wikipedia articles. Of particular interest are the titles of entity pages, redirect pages and disambiguation pages. Queries exactly matching one title of an entity page or a redirect page will be classified as EQ. Thus EQ can be mapped directly to the entity page with the same title. Note that queries with both entity page and disambiguation pages will be counted as EQ, because the existing entity page indicates that there is consensus on a dominant sense for the word or phrase, e.g. "Piracy‘" and "Poaching".

To identify AQ, we look for queries with terms in the ambiguous list of terms/phrases (derived from extracting all the titles of disambiguation pages). All other queries are then classified as BQ. Though the categorization process is simple, we show in later experiments that the results are helpful for retrieval performance enhancement.

3.2.1 Strategy for Ambiguous Queries

We now explain our solution to the problem of mapping ambiguous queries to a specific entity page. In previous research on query ambiguity, one assumption is that when an ambiguous query is submitted, the person that initiates the search knows the intended sense of each ambiguous word [28]. Our solution is also based on this assumption.

Given an ambiguous query, we first obtain the top ranked 100 documents from the target collection to be searched using the query likelihood retrieval model. Then K-means clustering is applied [15] to cluster these documents. Each document is represented by $tf * idf$ weighting and cosine normalization. Cosine similarity is used to compute the similarities among the top documents.

After clustering, we rank these clusters by a cluster-based language model, as proposed by Lee et al. [19], given as follows:

$$P(Q|Clu) = \prod_{i=1}^{m} P(q_i|Clu)$$

where $P(w|Clu)$ is specified by the cluster based language model

$$P(w|Clu) = \frac{|Clu|}{|Clu|+\lambda} P_{ML}(w|Clu) + \frac{\lambda}{|Clu|+\lambda} P_{ML}(w|Coll)$$

where $P_{ML}(w|Clu)$ is the maximum likelihood estimate of word w in the cluster and $P_{ML}(w|Coll)$ is the maximum likelihood estimate of word w in the collection. The smoothing parameter λ is learned using training topics on each collection in experiments. $P_{ML}(w|Clu)$ and $P_{ML}(w|Coll)$ are specified as follows:

$$P_{ML}(w|Clu) = \frac{freq(w, Clu)}{|Clu|}, P_{ML}(w|Coll) = \frac{freq(w, Coll)}{|Coll|}$$

where $freq(w, Clu)$ is the sum of $freq(w, D)$ for the document D which belongs to the cluster Clu, $freq(w, D)$ denotes the frequency of w in D, and $freq(w, Coll)$ is the number of times w occurs in the collection.

The documents in the top-ranked clusters are used to represent the characteristics of the test collection, in terms of pseudo relevant documents in response to a specific query. The top ranked cluster is then compared to all the referents (entity pages) extracted from the disambiguation page associated with the query. The assumption is that the dominant sense for the query should have a much greater degree of matching to the top ranked cluster from the test collection than other senses. Each document is represented by $tf * idf$ weighting and the cosine is used to measure the similarity between one cluster and an entity page. The top matching entity page is then chosen for the query.

Method	**AP**	**Robust**	**WT10G**	**GOV2**
EQ	21	58	31	41
AQ	108	159	54	98
BQ	21	33	15	11
All	150	250	100	150

Table 2: Numbers of each type of query in the TREC topic sets by automatic query categorization.

3.2.2 Evaluation

In order to evaluate the accuracy of our query categorization process, we used the four sets of TREC topics used in the retrieval experiments reported in Section 5, with five human subjects. These are taken from four different search tasks and comprise a total of 650 queries. Each participant was asked to judge whether a query is ambiguous or not. If it was, the participant was required to determine which referent from the disambiguation page is most likely to be mapped to the query, according to the description of the TREC topic. If it was not, the participant was required to manually search with the query in Wikipedia to identify whether or not it is an entity defined by Wikipedia (EQ). The user study results indicate that for query categorization, participants are in general agreement, i.e. 87% in judging whether a query is ambiguous or not. However, when determining which referent should a query be mapped to, there is only 54% agreement.

Table 2 shows the result of automatic query categorization process. It can be seen from Table 2 that most queries from TREC topic sets are AQ. That is to say, most of the queries contain ambiguity to some degree. Thus, it is necessary to handle this properly according to query type in query expansion.

To test the effectiveness of the cluster-based disambiguation process, we define that for each query, if there are at least two participants who indicate a referent as the most likely mapping target, this target will be used as an answer. If a query has no answer, it will not be counted by the evaluation process. Experimental results show that our disambiguation process leads to an accuracy of 57% for *AQ*.

4. QUERY EXPANSION METHODS

In this section, we describe our query expansion methods. Using these methods, we investigate the relationship between the query type and expansion methods. Moreover, we look into how to combine evidence from different fields of Wikipedia articles for query expansion. Essentially these methods differ in their criteria of selecting expansion terms. In each of these methods, query specific relevance information is considered.

4.1 Relevance model

A relevance model is a query expansion approach based on the language modeling framework [18]. The relevance model is a multinomial distribution which estimates the likelihood of word w given a query Q. In the model, the query words $q_1, ..., q_m$ and the word w in relevant documents are sampled identically and independently from a distribution R. Thus the probability of a word in the distribution R is estimated as follows :

$$P(w|R) = \sum_{D \in F} P(D)P(w|D)P(Q|D)$$

where F is the set of documents that are pseudo-relevant to the query Q. We assume that $P(D)$ is uniform over the set.

Based on this estimation, the most likely expansion term e from $P(w|D)$ is chosen for the original query. The final expanded query is combined with the original query using linear interpolation, weighted by a parameter λ.

$$P(w|Q^{'}) = (1 - \lambda)P(w|Q) + \lambda P(w|R)$$

The original relevance model and traditional PRF methods use the top retrieved documents from an initial search as pseudo-relevant documents. The problem is that the top retrieved documents frequently contain non-relevant documents or content, which can introduce noise into the expanded query, resulting in query drift. Our approach introduces Wikipedia articles as pseudo-relevant documents. This may still find non-relevant documents, but we will show that it can enhance the quality of pseudo-relevant documents for query expansion. This method forms the baseline in our experiments.

4.2 Strategy for Entity/Ambiguous Queries

One of the issues for web queries is they are often too short to clearly express a user's information need. With the help of Wikipedia, this issue is expected to be reduced to some degree. Entity queries are those "good" queries whose sense is given clearly. On the contrary, it is harder to find relevant documents for ambiguous queries. Both EQ and AQ can be associated with a specific Wikipedia entity page. In this strategy, instead of considering the top-ranked documents from the test collection, only the corresponding entity page from Wikipedia will be used as pseudo-relevant information. We briefly introduced entity pages in the first part of Section 3.1. An entity page contains the most representative information for the entity, which most Wikipedia users have an agreed consensus on.

Our strategy firstly ranks all the terms in the entity page, then the top K terms are chosen for expansion. The measure to score each term is defined as: $score(t) = tf * idf$, where tf is the term frequency in the entity page. idf is computed as $log(N/df)$, where N is the number of documents in the Wikipedia collection, and df is the number of documents that contain term t. The measure is simple, yet we will show in later experiments that it is very effective.

The performance of existing PRF methods is often affected significantly by parameters, e.g. the number of feedback documents used. The proposed method eases the problem by utilizing the fact that one article exists for each entity in Wikipedia which focuses on details of the subject.

4.3 Field Evidence for Query Expansion

Currently, structured documents are becoming more and more common, in consequence several studies have been conducted on exploiting field combination for improved IR [13, 26]. Although just semi-structured, our observations show that the evidence of different fields of Wikipedia can be used for improving retrieval, e.g. the importance of a term appearing in the *overview* may be different than its appearance in an *appendix*.

We examine two methods for utilizing evidence from different fields. The first is similar to that proposed by Robertson et al. for the BM25 model in [26], we replace the term frequency in a pseudo relevance document from original relevance model with a linearly combined weighted term frequencies. The second method is a supervised learning approach which classifies "good" expansion terms from "bad" ones. Features derived from fields are used by the classifier. Note that the field based expansion methods are applicable to all the queries. For EQ and AQ, the pseudo relevant documents can be either a query specific entity page, or just the same as BQ, i.e. the top ranked entity pages from the initial search.

Unsupervised Method.

Robertson et al. [26] gives a detailed analysis of the disadvantages of linear combination of the scores obtained from scoring every field, and recommends a simple weighted field extension. We use a similar idea in our work to explore the impact of different fields in Wikipedia. By assigning different weights to fields, we modify the $P_{ML}(w|D)$ in the original relevance model to the following form:

$$P_{ML}(w|D) = \frac{\sum_{f=1}^{K} W_f * TF_f(w, D)}{|D|}$$

where K (here $K = 6$) is the number of fields in a document, and $\sum_{f=1}^{K} W_f = 1$. Parameter tuning is needed for each single pair of parameters. We evaluate different weight settings for the fields, shown in next section.

Supervised Method (Support Vector Machines).

An alternative way of utilizing evidence of field information is to transfer it into features for supervised learning. The learning process is to train a classifier which distinguishes "good" expansion terms from "bad" ones. This method is inspired by the work of Cao et al. [3], where they suggest that "good" terms should be identified directly according to their possible impact on retrieval effectiveness.

We use Support Vector Machines (SVMs) [1], which are a popular supervised learner for tasks such as this, as a classifier. A radial-based kernel (RBF) SVM with default settings based on LIBSVM [4] is used. Parameters are estimated with a 5-fold cross-validation to maximize the classification accuracy of the training data.

In our work, we want to select good expansion terms and re-weight terms. It is important for us to know the probability estimated for each term belonging to each class. We set LIBSVM to train an SVM model for probability estimates. We compute posterior probabilities from SVM outputs using the method of [25], $P(+1|x) = \frac{1}{exp(A*S(x)+B)}$, where A and B are the parameters, $S(x)$ is the score calculated by the SVM. Each expansion term is represented by a feature vector $F(e) = [f_1(e), f_2(e), f_3(e), ..., f_n(e)]$.

The first group of features are the term distributions in the PRF documents and collections. Term distributions (TD) have already been used and proven to be useful in traditional approaches. The assumption is that the most useful features for term selection make the largest difference between the feedback documents and the whole collection. The features that we used include: (1) TD in the test collection; (2) TD in PRF (top 10) from the test collection; (3) TD in the Wikipedia collection; and (4) TD in PRF (top 10) from the Wikipedia collection. TD in the PRF documents from Wikipedia is given below, the others can be defined similarly.

$$f_{prfWiki}(e) = log\frac{\sum_{D \in F \bigcap W} tf(e, D)}{\sum_t \sum_{D \in F \bigcap W} tf(t, D)}$$

where F is the set of feedback documents, W is the Wikipedia collection.

The second group of features is based on field evidence. As described in section 3.1, we divide each entity page into six fields. One feature is defined for each field; this is computed as follows :

$$f_{Field_i}(e) = log(tf_i(e) * idf/fieldLength_i)$$

where $tf_i(e)$ is the term frequency in the field i for the entity page, and $fieldLength_i$ is the length of the field i.

	AP	**ROBUST**	**WT10G**	**GOV2**
Accuracy	74.15%	75.30%	72.99%	75.75%

Table 3: Term classification results

Training and test data are generated using a method similar to Cao et al. [3], that is, to label possible expansion terms of each query as good terms or non-good terms to see their impact on retrieval effectiveness. We define *good* expansion terms as those that improve the effectiveness when $\lambda_{fb} = 0.01$ and hurt the effectiveness when $\lambda_{fb} = -0.01$. Terms that do not satisfy these criteria are counted as *bad* terms. We now examine the classification quality. We use four test collections, see Table 4. We divide queries from the same collection into three equal size groups, and then perform a leave-one-out cross validation to evaluate classification accuracy, shown in Table 3.

5. EXPERIMENTS

5.1 Experiment Settings

In our experiments, documents are retrieved for a given query by the query-likelihood language model with Dirichlet smoothing. We set the Dirichlet prior empirically at 1,500 as recommended in [22]. The experiments were carried out using the Indri search engine. Indri provides support for PRF via a modified version of Lavrenko's relevance model. We implemented our new methods on top of Indri.

Experiments were conducted using four standard Text Retrieval Conference (TREC) collections : Associated Press is a small homogeneous collection; Robust2004, is the test collection for the TREC Robust Track started in 2003 to focus on poor performing queries; and two Web collections: the WT10G collection and the large-scale .GOV2 collection. Further details of these collections are given in Table 4.

Corpus	Size	# of Doc	Topics
AP88-90	0.73GB	242,918	51-200
Robust2004	1.9GB	528,155	301-450&601-700
WT10g	11GB	1,692,096	451-550
GOV2	427GB	25,205,179	701-850

Table 4: Overview of TREC collections and topics.

Retrieval effectiveness is measured in terms of Mean Average Precision (MAP). Given an initial query Q_{orig}, the

relevance model first retrieves a set of N documents and forms a relevance model from them. It then forms an expanded query Q_{RM} by adding the top K most likely non stopword terms from the relevance model. The expanded query is formed with the following structure:

$$\#weight(\quad 1.0 \quad Q_orig \qquad \lambda_{fb} \quad Q_{RM} \quad)$$

5.2 Baselines

The baseline of our experiments is the query-likelihood language model (**QL**) and the relevance model (**RMC**). Besides, we also consider the relevance model based on Wikipedia (**RMW**). Note that our RMW method retrieves the top ranked N Wikipedia articles that are not "redirect pages". For RMC and RMW we fixed the parameters for the rest experiments: $N = 10$, $K = 100$ and $\lambda = 0.6$, for a fair comparison.

Method	**AP**	**Robust**	**WT10G**	**GOV2**
QL	0.1428	0.2530	0.1831	0.2967
RMC	0.1707	0.2823	0.1969	0.3141
RMW	0.1622	0.2904*	0.2094*	0.3392*

Table 5: Performance comparisons using MAP for all the test topics on test collections. * indicate statistically significant improvements over RMC. We use the paired t-test with significance at p<0.05.

As can be seen from Table 5, Wikipedia is a good resource for relevance feedback on large collections. Using Wikipedia for generating PRF, RMW brings comparable performance with that produced by RMC for Robust2004, WT10G and GOV2. For AP, although RMW does not work as well as RMC, RMW still improves performance over QL. This serves as the baseline for our following experiments. We also believe that the test collections and Wikipedia have their own advantages. For test collections, the initial search emphasizes characteristics of the collection (e.g. GOV2 consists of web pages from government web sites), while Wikipedia appears to be more general in terms of topics.

5.3 Using Entity Pages for Relevance Feedback

We now turn our attention to our proposed method utilizing only the entity page corresponding to the query for PRF (**RE**), (see section 4.2). The results are presented in Table 6. Note that in our proposed method, not all the queries can be mapped to a specific Wikipedia entity page, thus the method is only applicable to EQ and AQ. Results for EQ and AQ are shown in Table 6 and Table 7 respectively. Note that results in Table 7 are based on automatic disambiguation. QL, RMC and RMW in the two tables are the averages for the same groups of queries.

As can be seen in Table 6 and Table 7, RE outperforms RMC and RMW on all collections. This indicates that what entity pages provide is relevant information that is closely related to the query, but might be missing in the test collection. Thus exploiting an entity page as the sole relevance feedback source works for both small and large collections. We also notice that the improvement of RE over RMW is greater for entity queries than for ambiguous queries. This is because the accuracy of the automatic disambiguation process is rather low (see section 3.2.2). If AQ is not associated with the most relevant entity page, its performance suffers from the RE method.

Method	**AP**	**Robust**	**WT10G**	**GOV2**
QL	0.2208	0.3156	0.2749	0.3022
RMC	0.2484	0.3401	0.2458	0.3168
RMW	0.2335	0.3295	0.2821*	0.3453*
RE	0.2494**	0.3580**	0.2897*	0.3889**

Table 6: Performance comparisons using MAP for *entity queries* on test collections. * and ** indicate statistically significant improvements over RMC and RMW, respectively. We use the paired t-test with significance at p<0.05.

Comparing Tables 5, 6 and 7, we can see that entity queries have better performance than the average, while ambiguous queries have lower performance than the average. For ambiguous queries, using the top retrieved documents as pseudo relevant documents usually includes more non-relevant documents than for entity queries. The disambiguation process helps to find the relevant entity page for ambiguous queries, thus both types of queries can benefit from the query expansion. Based on these results, refinement of the disambiguation process for AQ could be expected to further improve performance.

Method	**AP**	**Robust**	**WT10G**	**GOV2**
QL	0.1391	0.2258	0.1801	0.2892
RMC	0.1520	0.2485	0.1881	0.3101
RMW	0.1588	0.2619*	0.1868	0.3186*
RE	0.1692**	0.2728**	0.2037**	0.3329**

Table 7: Performance comparisons using MAP for *ambiguous queries* on test collections. * and ** indicate statistically significant improvements over RMC and RMW, respectively. We use the paired t-test with significance at p<0.05.

5.4 Field Based Expansion

Inspired by the result that an entity page can indeed help to improve retrieval performance, we next go on to investigate utilizing field evidence from entity pages. We will first see the results of two different term selection methods based on field evidence, then an analysis between them is given. Note that in our experiments on field based expansion, the top retrieved documents are considered pseudo relevant documents for EQ and AQ.

For the supervised method, we compare two ways of incorporating expansion terms for retrieval. The first is to add the top ranked 100 *good* terms (**SL**). The second is to add the top ranked 10 *good* terms, each given the classification probability as weight (**SLW**). The relevance model with weighted TFs is denoted as (**RMWTF**). From Table 8, we can see that both methods enhance retrieval performance. Among the three methods, RMWTF and SLW generate similar results. However, the SLW is subject to the accuracy of term classification, thus we choose RMWTF for the query dependent method introduced in the next section. Although we do not give results for BQ for space reasons, our experiments show that BQ is improved by the field based expansion. In addition, characteristics of BQ could be investigated so that pseudo relevant documents could be tailored for BQ, and the field based expansion still be applied.

Method	AP	Robust	WT10G	GOV2
SL	0.1640	0.2902	0.2107	0.3237
SLW	0.1702	0.2921*	0.2145*	0.3298*
RMWTF	0.1768*	0.2934*	0.2153*	0.3274*

Table 8: Supervised Learning vs Linear Combination. * indicates statistically significant improvements over RMC. We use the paired t-test with significance at p<0.05.

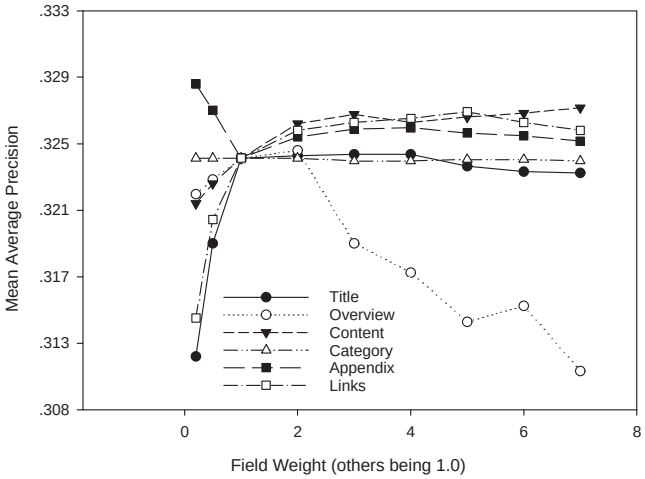

Figure 1: Performance on different field weights on the GOV2 collection.

Figure 1 shows the results of assigning different weight to fields on GOV2. As can be seen in Figure 1, performance improves as weights for *Links*, *Content* increase. On the other hand, the increase of weight to *Overview* leads to deterioration of the performance. This shows that the positions where a term appears have different impacts on the indication of term relevancy.

5.5 Query Dependent Expansion

We now explore a query dependent expansion method. We assign different methods for queries according to their types as follows: RE for EQ and AQ, RMWTF for BQ. RE is chosen because EQ and AQ benefit more from RE than from RMWTF. For AQ, a disambiguation process is conducted to determine the corresponding entity page. We denote the query dependent method as (**QD**). As can be seen in Table 9, the improvement of QD over RMC is significant across test collections.

		AP	Robust	WT10G	GOV2
RMC	MAP	0.1707	0.2823	0.1969	0.3141
	IMP	119	174	47	88
QD	MAP	0.1777	0.3002*	0.2194*	0.3348*
	IMP	116	191	67	96

Table 9: Query Dependent vs traditional Relevance Model. IMP is the number of queries improved by the method over QL. * indicates statistically significant improvements over RMC. We use the paired t-test with significant at p<0.05.

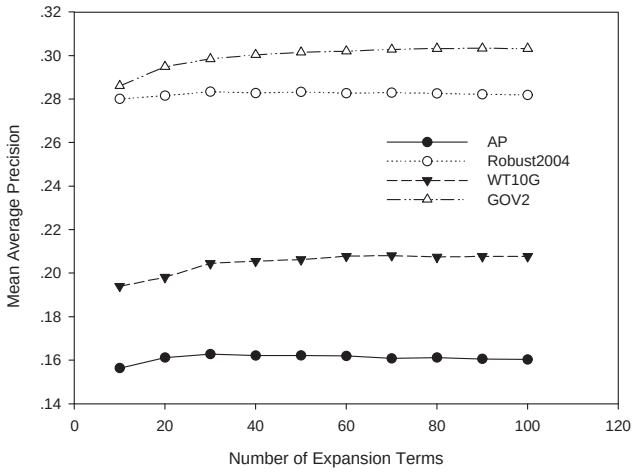

Figure 2: Parameter (K) sensitivity of expansion terms over different data sets, $N = 10$,$\lambda_{fb} = 0.6$.

Table 9 also presents an analysis of the robustness by giving the numbers of queries improved by RMC and QD over QL respectively. QD shows improvement for robustness on Robust2004, WT10G and GOV2 collections.

5.6 Parameter Selection

Both the RMC and RMW methods have parameters N, k and λ_{fb}. We tested these two methods with 10 different values of N, the number of feedback documents: 10, 20,..., 100. For λ_{fb}, we tested with 5 different values: 0.4, 0.5, 0.6, 0.7 and 0.8. Due to space limitations, we present only the final results. Results show that setting $N = 10$, $K = 50$ and $\lambda_{fb} = 0.6$ work best for the values tested. Figure 2 shows the sensitivity of RMW to K. The results for other methods are similar. Figure 2 shows that retrieval performance varies little as the number of expansion terms increases.

6. CONCLUSION

In this paper, we have explored utilization of Wikipedia in PRF. In this work TREC topics are categorized into three types based on Wikipedia. We propose and study different methods for term selection using pseudo relevance information from Wikipedia entity pages. We evaluated these methods on four TREC collections. The impact of Wikipedia on retrieval performance for different types of queries has been evaluated and compared. Our experimental results show that the query dependent approach can improve over a baseline relevance model.

This study suggests several interesting research avenues for our future investigations. More investigation is needed to explore the characteristics and possible technique refinements for the broader queries. For ambiguous queries, if the disambiguation process can achieve improved accuracy, the effectiveness of the final retrieval will be improved. For the supervised term selection method, the results obtained are not satisfactory in terms of accuracy. This means that there is still much room for improvement. We are going to explore more features for the learning process. Finally, in this paper, we focused on using Wikipedia as the sole source

of PRF information. However, we believe both the initial result from the test collection and Wikipedia have their own advantages for PRF. By combining them together, one may be able to develop an expansion strategy which is robust to the query being degraded by either of the resources.

7. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China under Grant No. 60603094, the Major State Basic Research Project of China (973 Program) under Grant No. 2007CB311103 and the National High Technology Research and Development Program of China (863 Program) under Grant No. 2006AA010105. The authors are grateful to the anonymous reviewers for their comments, which have helped improve the quality of the paper.

8. REFERENCES

[1] C. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[2] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In *Text REtrieval Conference*, 1994.

[3] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of SIGIR 2008*, pages 243–250.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/c̃jlin/libsvm.

[5] P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the web. In *Proceeding of SIGIR 2007*, pages 7–14.

[6] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A framework for selective query expansion. In *Proceedings of CIKM 2004*, pages 236–237.

[7] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *Proceedings of SIGIR 2008*, pages 347–354.

[8] C. Fautsch and J. Savoy. UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogsphere. In *Proceedings of TREC 2008*.

[9] S. Fissaha Adafre, V. Jijkoun, and M. de Rijke. Fact discovery in Wikipedia. In *Proceedings of Web Intelligence 2007*, pages 177–183.

[10] B. M. Fonseca, P. Golgher, B. Pôssas, B. Ribeiro-Neto, and N. Ziviani. Concept-based interactive query expansion. In *Proceedings of CIKM 2005*, pages 696–703.

[11] G. R. Giambattista Amati, Claudio Carpineto and F. U. Bordoni. Query difficulty, robustness and selective application of query expansion. In *Proceedings of ECIR 2004*, pages 127–137, 2004.

[12] J.-Y. N. Hang Cui, Ji-Rong Wen and W.-Y. Ma. Query expansion by mining user logs. *IEEE Transactions on knowledge and data engineering*, 15(4):829–839, 2003.

[13] B. He and I. Ounis. Combining fields for query expansion and adaptive query expansion. *Information Processing and Management*, 2007.

[14] Indri. *http://www.lemurproject.org/indri/*.

[15] J.A.Hartigan and M. Wong. A k-means clustering algorithm. *Applied Statistics*.

[16] W. W.-J. H. K. Balog, E. Meij and M. de Rijke. The University of Amsterdam at TREC 2008: Blog, Enterprise, and Relevance Feedback. In *Proceedings of TREC 2008*.

[17] K. L. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of SIGIR 1998*, pages 250–256.

[18] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of SIGIR 2001*, pages 120–127.

[19] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of SIGIR 2008*, pages 235–242.

[20] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using Wikipedia as external corpus. In *Proceedings of SIGIR 2007*, pages 797–798.

[21] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *Proceedings of SIGIR 2007*, pages 311–318.

[22] D. Metzler, T. Strohman, H. Turtle, and W. Croft. Indri at trec 2005: Terabyte track. In *Proceedings of TREC 2004*.

[23] D. N. Milne, I. H. Witten, and D. M. Nichols. A knowledge-based search engine powered by Wikipedia. In *Proceedings of CIKM 2007*, pages 445–454.

[24] G. Mishne. *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, Amsterdam, 2007.

[25] J. Platt. Probabilities for SV machines. *Advances in large margin classifiers*, pages 61–74.

[26] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of CIKM 2004*, pages 42–49.

[27] S. E. Robertson, S. Walker, M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *In Proceedings of the 4th Text REtrieval Conference (TREC)*, 1996.

[28] M. Sanderson. Ambiguous queries: test collections need more sense. In *Proceedings of SIGIR 2008*, pages 499–506.

[29] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of SIGIR 2006*, pages 162–169.

[30] Wikipedia. *http://www.wikipedia.org*.

[31] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.

[32] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty. In *Proceedings of SIGIR 2005*, pages 512–519.

[33] M. M. Zesch Torsten, Gurevych Iryna. Analyzing and accessing Wikipedia as a lexical semantic resource. In *Biannual Conference of the Society for Computational Linguistics and Language Technology 2007*, pages 213–221.

[34] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM 2001*, pages 403–410.

[35] W. Zhang and C. Yu. UIC at TREC 2006 Blog Track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2007.