Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1294–1307
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Combining fields for query expansion and adaptive query expansion

Ben He *, Iadh Ounis

Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, United Kingdom

Received 19 June 2006; received in revised form 2 November 2006; accepted 4 November 2006
Available online 26 January 2007

Abstract

In this paper, we aim to improve query expansion for ad-hoc retrieval, by proposing a more fine-grained term reweighting process. This fine-grained process uses statistics from the representation of documents in various fields, such as their titles, the anchor text of their incoming links, and their body content. The contribution of this paper is twofold: First, we propose a novel query expansion mechanism on fields by combining field evidence available in a corpora. Second, we propose an adaptive query expansion mechanism that selects an appropriate collection resource, either the local collection, or a high-quality external resource, for query expansion on a per-query basis. The two proposed query expansion approaches are thoroughly evaluated using two standard Text Retrieval Conference (TREC) Web collections, namely the WT10G collection and the large-scale .GOV2 collection. From the experimental results, we observe a statistically significant improvement compared with the baselines. Moreover, we conclude that the adaptive query expansion mechanism is very effective when the external collection used is much larger than the local collection.

Keywords: Query expansion on fields; Pseudo relevance feedback; Information retrieval; External expansion; Adaptive query expansion; TREC experiments

1. Introduction

In Information Retrieval (IR), *query expansion* usually refers to the technique that uses blind relevance feedback to expand a query with new query terms, and reweigh the query terms, by taking into account a pseudo relevance set (Rocchio, 1971; Amati, 2003). Usually, the pseudo relevance set consists of the top-ranked documents returned by the first-pass retrieval. These top-ranked documents are assumed to be relevant to the topic. Query expansion has proved to be an effective technique for ad-hoc retrieval. For example, in previous Text Retrieval Conference (TREC) ad-hoc tasks, a consistent improvement brought by query expansion has been observed (for example, see Hawking, 2000; Hawking & Craswell, 2001).

* Corresponding author. Tel.: +44 141 330 6292; fax: +44 141 330 4913.
E-mail addresses: ben@dcs.gla.ac.uk (B. He), ounis@dcs.gla.ac.uk (I. Ounis).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.11.002

However, in some cases, query expansion can lead to little improvement in the retrieval performance. For example, in the TREC 2004 Terabyte track, some participants found that query expansion was not helpful for ad-hoc retrieval (Attardi, Esuli, & Patel, 2004; Plachouras, He, & Ounis, 2004). As stressed in Amati, Carpineto, and Romano (2004a), the effectiveness of query expansion is correlated with the quality of the top-ranked documents returned by the first-pass retrieval, on which query expansion is based. The more closely the documents returned by the first-pass retrieval are related to the topic, the better the performance that the query expansion mechanism can achieve. In other words, if the top-ranked documents are poor, this can lead to a failure for query expansion to improve the retrieval performance (Amati et al., 2004a; Yom-Tov, Fine, Carmel, & Darlow, 2005).

Improving the effectiveness and robustness of query expansion has been the focus of some previous work. Two main approaches have been proposed. The first one is selective query expansion, which disables query expansion if query expansion is predicted to be detrimental to retrieval (Amati et al., 2004a; Cronen-Townsend, Zhou, & Croft, 2004; Yom-Tov et al., 2005). The second one is the collection enrichment approach, which performs query expansion using a large high quality external collection, and then retrieves from the local collection using the expanded query (Kwok & Chan, 1998). A high quality collection refers to a collection, in which most documents are highly informative. For example, the TREC newswire collection is a high-quality one, since each of its documents usually contains mostly informative terms and very little noise. In this paper, the notion of *local collection* refers to the collection from which the final retrieved documents are returned, and the notion of *external collection* refers to any other distinct collection. In this paper, we aim to further improve the effectiveness and robustness of query expansion by two different approaches.

First, we propose a novel query expansion mechanism on fields, which appropriately uses field evidence available in a corpora. For example, a Web document can be represented by combining fields, such as its title, the anchor text of its incoming links, and the body of its text. The queries can then be reweighed and expanded using this more refined available information. We suggest that the performance of query expansion is related not only to the quality of the top-ranked documents returned by the first-pass retrieval, but also to the quality of the query term reweighting. We believe that combining evidence from different document fields can refine the statistics of the query terms, and hence improve the query term reweighting. Moreover, the combination of field evidence is applied in the first-pass retrieval in order to refine the quality of the top-ranked documents.

Second, we propose a low-cost adaptive decision mechanism for the application of query expansion. The decision mechanism is based on the pre-retrieval performance prediction technique (He & Ounis, 2005). The proposed mechanism applies both collection enrichment and selective query expansion techniques. In the proposed adaptive query expansion mechanism, the expansion of the query can be either local, using documents in the local collection, or external, using an external resource. The adaptive mechanism predicts the benefit of query expansion using either option, and adopts the most effective one. If both options are predicted to lead to the degradation of the query performance, then query expansion is disabled.

Our new query expansion techniques are based on the Divergence from Randomness (DFR) framework, which measures the informativeness of a term in a document by the divergence of the term's distribution in the document from a random distribution (Amati, 2003). The idea of combining fields for the first-pass retrieval has been studied in Robertson, Zaragoza, and Taylor (2004) Zaragoza, Craswell, Taylor, Saria, and Robertson (2004) by extending BM25 (Robertson, Walker, Beaulieu, Gatford, & Payne, 1995) to the field retrieval BM25F model, and by extending PL2 (Amati, 2003) to the field retrieval PL2F model (Macdonald, He, Plachouras, & Ounis, 2005). In this paper, similarly to the approach in Macdonald et al. (2005) and Robertson et al. (2004), we extend the parameter-free DLH model (Amati, 2006) to field retrieval.

The remainder of this paper is organised as follows. We introduce the related works in Section 2. In Sections 3 and 4, we propose the query expansion on fields and the adaptive query expansion mechanism, respectively. The two new approaches are extensively evaluated on two standard TREC Web collections. The document weighting model used in the experiments in Sections 3 and 4 is a DFR document weighting model. In Section 5, we present experiments using BM25F to assess the impact of a different document weighting model on the experimental results. In Section 6, we conclude the work and suggest future research directions.

2. Related works

In Section 2.1, we introduce the BM25F model on fields. We introduce a Divergence from Randomness (DFR) document weighting model for the first-pass retrieval in Section 2.2, and introduce a DFR-based term weighting model for query expansion in Section 2.3. These two DFR-based models will be later extended to handle fields. In Section 2.4, we introduce the selective query expansion and collection enrichment techniques, which will be applied in our proposed adaptive query expansion mechanism.

2.1. The BM25F field retrieval document weighting model

In the BM25F model, the relevance score of a document d for a query Q is given by (Robertson et al., 2004):

$$\text{score}(d,Q) = \sum_{t \in Q} w^{(1)} \frac{(k_1+1)tfn}{k_1+tfn} \frac{(k_3+1)qtf}{k_3+qtf} \tag{1}$$

where qtf is the query term frequency and k_1 and k_3 are the parameters. The default setting is $k_1 = 1.2$ and $k_3 = 1000$ (Robertson et al., 1995). $w^{(1)}$ is the *idf* factor, which is given by:

$$w^{(1)} = \log_2 \frac{N - N_t + 0.5}{N_t + 0.5}$$

N is the number of documents in the whole collection and N_t is the document frequency of term t.

The normalised term frequency tfn is given by a so-called per-field normalisation component, where term frequency is normalised in a per-field basis (Zaragoza et al., 2004). This per-field normalisation component applies a linear combination of the normalised term frequencies from different fields as follows:

$$tfn = \sum_f w_f \cdot \frac{tf_f}{(1-b_f) + b_f \cdot \frac{l_f}{avg\_lf}} \tag{2}$$

where w_f is the weight of a field f, tf_f the frequency of the query term in the field f of the document, b_f the term frequency normalisation hyper-parameter of field f, l_f the number of tokens in field f of the document, and avg_lf is the average length of field f in the collection, i.e. the average number of tokens in field f.

Note that the BM25 model's formula (Robertson et al., 1995) is the same as Eq. (1), while its normalised term frequency tfn is given by its term frequency normalisation component:

$$tfn = \frac{tf}{(1-b) + b \cdot \frac{l}{avg\_l}} \tag{3}$$

where tf is the term frequency in the documents, b the term frequency normalisation hyper-parameter, l the document length, and avg_l is the average document length in the collection.

2.2. The DLH document weighting model

The DLH model is a generalisation of the parameter-free hypergeometric DFR model in a binomial case (Amati, 2006). Previous probabilistic weighting models, e.g. BM25 (Robertson et al., 1995), consider the occurrences of a query term in a document to be samples from the document. Unlike BM25, the hypergeometric model assumes that the occurrences of a query term in a document are samples from the whole collection, instead of from the document. The DLH model does not have a term frequency normalisation component, and does not have any parameters that require relevance tuning. Therefore, it has no need for expensive training with relevance judgement, which is not always available in a practical setting. In other words, all the variables of the document weighting formula are automatically set from the collection statistics. Therefore, the hypergeometric model does not have the need for tuning its parameters, in order to achieve an optimised retrieval performance. In the DLH model, the relevance score of a document d for a query Q is given by:

$$\text{score}(d,Q) = \sum_{t \in Q} qtw \cdot \frac{1}{tf + 0.5} \cdot \left(\log_2 \left(\frac{tf \cdot avg\_l}{l} \cdot \frac{N}{F} \right) + 0.5 \log_2 \left(2\pi tf \left(1 - \frac{tf}{l} \right) \right) \right) \quad (4)$$

where tf is the frequency of the term in document d, avg_l the average document length in the collection, l the document length, and qtw is the query term weight. This is given by $qtf/qtf_{\max}$, where qtf is the query term frequency and $qtf_{\max}$ is the maximum qtf among all the query terms. F is the frequency of the term in the collection and N is the number of documents in the collection. Note that the DLH model in Eq. (4) does not have a tf normalisation component, as this is assumed to be inherent to the model.

2.3. Divergence from randomness query expansion mechanism

The Divergence from Randomness (DFR) framework employs a query expansion mechanism that is a generalisation of Rocchio's method (Rocchio, 1971). The DFR query expansion mechanism has two steps.

First, it applies a DFR term weighting model to measure the informativeness of the terms in the top-ranked documents. The idea of the DFR term weighting model is to infer the informativeness of a term by the divergence of its distribution in the top-ranked documents from a random distribution. The most effective DFR term weighting model is the Bo1 model that uses the Bose–Einstein statistics (Amati, 2003; Macdonald et al., 2005; Plachouras et al., 2004). Using this model, the weight w of a term t in the top-ranked documents is given by:

$$w(t) = tf_x \cdot \log_2 \frac{1 + P_n}{P_n} + \log_2(1 + P_n) \quad (5)$$

where tf_x is the frequency of the query term in the top-ranked documents, P_n is given by $\frac{F}{N}$, F the frequency of the term in the collection, and N is the number of documents in the collection. In this paper, following the default setting of Amati (2003), we extract the 10 most informative terms from the top 3 returned documents. The original query terms may also appear in the 10 extracted terms.

In the second step, the DFR query expansion mechanism expands the query by merging the extracted terms with the original query terms. The query term weight qtw is given by a parameter-free query expansion formula:

$$qtw = \frac{qtf}{qtf_{\max}} + \frac{w(t)}{\lim_{F \to tf_x} w(t)} = F_{\max} \log_2 \frac{1 + P_{n,\max}}{P_{n,\max}} + \log_2(1 + P_{n,\max}) \quad (6)$$

where $\lim_{F \to tf_x} w(t)$ is the upper bound of $w(t)$, $P_{n,\max}$ is given by $F_{\max}/N$, and $F_{\max}$ is the frequency F of the term with the maximum $w(t)$ in the top-ranked documents. If an original query term does not appear in the most informative terms extracted from the top-ranked documents, its query term weight remains equal to the original one. Note again that the above query expansion formula does not have any parameter that requires tuning.

2.4. Selective query expansion and collection enrichment

The basic idea of *selective query expansion* is to disable query expansion if the query is predicted to perform poorly. A selective query expansion mechanism was proposed by Amati, Carpineto, and Romano (2004b) in the context of the DFR framework. It predicts the performance of query expansion by the *query difficulty*, which looks at the divergence of a query term's distribution in the top-ranked documents from this distribution in the whole collection. The larger the divergence is, the better retrieval performance query expansion can result in. An alternate selective query expansion mechanism was proposed by Cronen-Townsend et al. (2004) in the context of language modeling. In the latter approach, the query performance is predicted using the *clarity score* (Cronen-Townsend, Zhou, & Croft, 2002), which is defined as the Kullback–Leibler divergence of the query model from the collection model. Again, the larger the divergence is, the better retrieval performance query expansion can provide.

In Kwok and Chan (1998) studied the idea of using an external resource for query expansion. They suggested that the failure of query expansion is 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, which possibly contains more relevant documents and has better collection statistics for the query term reweighting. This *collection enrichment*[1] approach was later applied in the TREC Robust tracks (Grunfeld, Kwok, Dinstl, & Deng, 2003; Kwok, Grunfeld, Sun, & Deng, 2004). Moreover, Singhal et al. proposed the conservative collection enrichment approach to handle the case where the external collection is not related to the topic (Singhal, Choi, Hindle, & Lewis, 1998). In the conservative collection enrichment approach, the query is expanded on the local collection, and the external collection is only used for query term reweighting.

In this paper, we extend the DLH document weighting model and the Bo1 term weighting model for query expansion to field retrieval. Moreover, we devise a novel adaptive query expansion mechanism that applies both selective query expansion and collection enrichment techniques. This mechanism selects the appropriate resource, either the local collection, or an external resource, for query expansion on a per-query basis on fields in Section 3 and the adaptive query expansion mechanism in Section 4.

3. Combining field evidence for query expansion

In this section, we present a query expansion mechanism on fields that is based on an extension of the Bo1 model for query expansion. We also extend the DLH model to handle fields. In Robertson et al. (2004), Robertson et al. denote the BM25 model on fields as the BM25F model. In this work, we follow their notation by denoting the DLH model on fields as the *DLHF* model, and the Bo1 model on fields as the *Bo1F* model.

We propose the DLHF model in Section 3.1 and the query expansion mechanism on fields in Section 3.2. The related evaluation is presented in Section 3.3.

3.1. The DLHF field retrieval document weighting model

In this section, we describe how we extend the DLH model to field retrieval in order to improve the top-ranked documents from the first-pass retrieval.

Unlike the BM25 model (see Eq. (1)), the DLH model does not have an explicit term frequency normalisation component. Therefore, in order to adapt it to field retrieval, we can directly combine the term frequencies in different fields without normalising the term frequency.

In our DLHF field retrieval model, the term frequency tf in the document is a linear combination of the term frequencies in different fields of the document. Similarly to the BM25F model, each field is associated with a weight. The linear combination is given below:

$$tf = \sum_{\mathrm{f}} w_{\mathrm{f}} \cdot tf_{\mathrm{f}} \tag{7}$$

where w_{f} is the weight of a field f, and tf_{f} is the term frequency in field f of the document.

The DLHF field retrieval model is defined as Eq. (4), where the term frequency tf is given by the above linear combination. We use the DLHF model to perform both the first-pass retrieval, and the retrieval using the expanded query. Note that the DLHF model has fewer parameters than the BM25F model (see Section 2.1), in the sense that the DLHF model does not have any term frequency normalisation hyper-parameters (b_{f} in BM25F).

3.2. Query expansion framework on fields

In this section, we propose a query expansion mechanism on fields. We aim to improve the quality of the query term reweighting by refining the statistics of the query terms. We suggest that in addition to the quality of the top-ranked documents, the quality of the query term reweighting is also related to the performance of query expansion. Combining field evidence can provide better statistics for the query terms. Therefore, it may

[1] In Macdonald et al. (2005), we called the process that uses an external resource for query expansion the external expansion.

improve query term reweighting and achieve a better performance of query expansion. This assumption will be evaluated in Section 3.3.

In order to combine the statistics from different fields, we could compute a term weight for each field separately, then combine them afterwards. As stressed in Robertson et al. (2004), this method suffers from the risk of breaking the non-linear saturation of term frequency in the weighting model. Therefore, we combine the frequencies of the query term in the top-ranked documents directly, and then weight the term based on the combined statistics.

We apply a linear combination to sum the term frequencies in different fields. The term frequency in each field of the top-ranked documents is $tf_{x\mathrm{f}}$, and each field is associated with a weight wq_{f}. The term frequency tf_x in the top-ranked documents is given by:

$$tf_x = \sum_{\mathrm{f}} wq_{\mathrm{f}} \cdot tf_{x\mathrm{f}} \tag{8}$$

where wq_{f} is the weight of a field f in the top-ranked documents. Each weight wq_{f} reflects the relative importance of the associated field in the top-ranked documents. tf_{xf} is the term frequency in field f in the top-ranked documents.

We then define the Bo1F term weighting model on fields as Eq. (5), where the term frequency in the top-ranked documents tf_x is given by Eq. (8).

Using the Bo1F model, the proposed query expansion mechanism on fields can be described as follows:

- First, we run a first-pass retrieval and obtain the top-ranked documents.
- Second, we compute the weights of the terms in the top-ranked documents using the Bo1F term weighting model on fields.
- Third, we extract the terms with the highest term weights from the top-ranked documents, merge them with the original query terms, and then eweigh the terms in the expanded query using Eq. (6).

We evaluate the proposed query expansion mechanism on fields in the next section.

3.3. Evaluation of the query expansion mechanism on fields

In Section 3.3.1, we introduce the experimental setting of the evaluation of our query expansion mechanism on fields. We present the results in Section 3.3.2.

3.3.1. Experimental setting

Our experiments are conducted using the Terrier platform. Terrier is a modular platform for the rapid development of large-scale Information Retrieval applications, providing indexing and retrieval functionalities. It is developed at the University of Glasgow.[2]

In this section, we present experiments that aim to evaluate our query expansion mechanism on fields. We use a Web retrieval context where a document is represented using three fields, namely its title, the anchor text of its incoming links, and the body of its text.

We use two standard TREC Web collections, namely the WT10G and .GOV2 collections.[3] The WT10G collection is a small crawl of Web documents, which was used in the TREC 9, 10 Web tracks. It contains 1,692,096 Web documents, and has 10 G of data. The .GOV2 collection, which contains 25,205,179 Web documents and 426 G of data, is a crawl from the .gov domain. This collection has been employed in the TREC 13 and 14 Terabyte tracks (Clarke, Craswell, & Soboroff, 2004; Clarke, Scholer, & Soboroff, 2005). GOV2 is the largest TREC test collection. We use these two different standard Web collections to test the impact of the collection size on the query expansion mechanism on fields.

For the test topics, we use the TREC 9 & 10 Web ad-hoc tasks, and the TREC 13 & 14 Terabyte ad-hoc tasks, including the latest TREC ad-hoc topics in the TREC 14 Terabyte track. Table 1 lists the topic numbers

[2] More information about Terrier can be found at the following URL: http://ir.dcs.gla.ac.uk/terrier/.
[3] Related information on these two collections can be found from http://ir.dcs.gla.ac.uk/test_collections/.

Table 1
The TREC tasks and topic numbers associated with each collection

Coll.	Task	Topic#
WT10G	TREC 9 Web ad-hoc	451–500
WT10G	TREC 10 Web ad-hoc	501–550
.GOV2	TREC 13 Terabyte ad-hoc	701–750
.GOV2	TREC 14 Terabyte ad-hoc	751–800

associated with each TREC task on each Web collection used. Each task involves 50 topics, and we have 100 topics for each Web collection used.

Each TREC topic has three fields, namely title, description and narrative. In our experiments, we only use the titles. The reason for using only the titles is that it is a practical and realistic setting, since real queries on the Web are usually very short (Silverstein, Henzinger, & Marais, 1998).

We create indices for the three fields, i.e. body, title, and anchor text, using the WT10G and .GOV2 collections, respectively. Statistics of the three fields in the two collections are provided in Table 2. We can see that the size of the body field is much larger than the other two.

The tokens in the collections used are stopped using a standard stop-word list, and stemmed using Porter's stemming algorithm.

Our query expansion on fields involves six parameters, namely the weights (w_f) of the three used fields in the DLHF model, and the weights (wq_f) of the three used fields in the query expansion mechanism. Since it would be very time-consuming to optimise all six parameters, we make the following assumptions to reduce the number of parameters to two:

1. For a particular field f, we assume that $w_f = wq_f$. This is reasonable because the weight of a field reflects the contribution of the field in the retrieval, which should be consistent in document weighting and in query expansion.
2. Similarly to the work in Robertson et al. (2004), we set the weight of the body field to 1.

By making the above two assumptions, we reduce the number of parameters to two, namely the weights of the anchor text and title fields. In the rest of this paper, we use w_a and w_t to denote the weights of the anchor text and title fields, respectively.

To optimise the weights of the anchor text and title fields, we do a two-dimensional data sweeping within $[0, 2]$ with an interval of 0.1, and then choose the weights that give the best mean average precision (MAP) using relevance assessments on the test topics.

Our baseline is the content-based query expansion, which is equivalent to applying a normal query expansion over the three document fields. Using the WT10G and .GOV2 collections, we compare our query expansion mechanism on fields with the baseline. Moreover, since real user queries usually do not have relevance assessments available, we conduct experiments to find out what happens if the field weights are trained based

Table 2
The number of tokens and the average document length (*avg_l*) in each field of the two collections

Field	#tokens	*avg_l*
WT10G, N: 1,692,096		
Body	668,150,053	394.86
Anchor text	22,596,525	13.35
Title	7,479,058	4.42
Total	698,225,636	412.64
.GOV2, N: 25,205,179		
Body	16,331,848,283	647.96
Anchor text	966,003,355	38.32
Title	132,916,051	5.27
Total	17,430,767,689	691.56

N is the number of documents in each collection.

on a sample of test queries. Each of the two collections used is associated to 100 topics used in two TREC ad-hoc tasks. On each collection used, we conduct a two-fold holdout evaluation. In each fold, we train the field weights on the 50 topics used in one associated ad-hoc task, and test the effectiveness of our query expansion mechanism on the 50 topics used in another associated TREC ad-hoc task, using the field weights obtained in training. In the next section, we provide analysis on the experimental results.

3.3.2. Experimental results

Table 3 compares the mean average precision (MAP) obtained by the content-based query expansion (QE) and by our query expansion mechanism on fields (QEF). w_a and w_t are the optimal weights of anchor text and title, which result in the best obtained MAP by the query expansion mechanism on fields. $\Delta\%$ indicates the difference between the two MAP values in percentage. The p-value is given by the Wilcoxon matched-pairs signed-ranks test.

From Table 3, we observe an improvement brought by our query expansion mechanism on fields. According to the Wilcoxon matched-pairs signed-ranks test, the improvement is statistically significant at 0.05 level on both collections. In our experiments, we do not observe a remarkable impact of the collection size on the performance of the query expansion mechanism on fields.

Moreover, we find that anchor text is not very useful in refining query term reweighting. Indeed, the optimal weight of the anchor text field is 0 or 0.1 for the four associated TREC ad-hoc tasks, (see column w_a of Table 3). This indicates that the anchor text information makes no ($w_a = 0$), or only a little contribution ($w_a = 0.1$), to improving query term reweighting. We suggest that this is because a query term usually repeatedly occurs in the documents where the term appears. Consequently, the frequency of the query term in the anchor text field is so high that a low anchor text field weight is required to reduce the contribution of anchor text on the document ranking. On the other hand, it is worth giving a relatively large weight to the title field (see column w_t of Table 3). Therefore, the title field is a good evidence for improving query expansion.

Table 4 contains the experimental results of the two-fold holdout evaluation, which uses a sample of the associated TREC topics for training, and uses the other ones for testing. Table 4 shows that when the topics

Table 3
The mean average precision obtained by content-based query expansion (QE) and by our query expansion mechanism on fields (QEF), respectively

TREC	QE	QEF	w_a	w_t	$\Delta\%$	p-Value	+	−	=
TREC 9	0.1792	0.1991	0.0	0.0	+11.10	2.986e − 06*	40	7	3
TREC 10	0.2142	0.2421	0.1	0.4	+13.02	2.480e − 06*	43	6	1
TREC 9 & 10	0.1967	0.2206	0.0	0.1	+10.83	2.200e − 11*	83	13	4
TREC 13	0.2420	0.2639	0.1	1.0	+9.05	3.624e − 04*	38	11	1
TREC 14	0.3248	0.3482	0.1	0.5	+7.20	1.601e − 04*	35	14	1
TREC 13 & 14	0.2838	0.3048	0.1	0.4	+7.40	1.06e − 05*	68	31	1

w_a and w_t are the optimal weights of anchor text and title, respectively. The p-value is given by the Wilcoxon matched-pairs signed-ranks test. A star indicates a statistically significant difference at 0.05 level. +, −, and = indicate for how many queries QEF results in improvement, degradation and no change in the retrieval performance, respectively.

Table 4
The mean average precision obtained by the content-based query expansion (QE) and by our query expansion mechanism on fields (QEF), respectively

TREC	QE	QEF	$\Delta\%$	p-Value
TREC 9	0.1811	0.1921	+6.07	5.576e − 04*
TREC 10	0.2205	0.2376	+7.76	5.371e − 04*
TREC 13	0.2420	0.2576	+6.45	2.045e − 03*
TREC 14	0.3248	0.3468	+6.77	6.555e − 04*

The field weights are obtained by training on a sample of the TREC topics. The p-value is given by the Wilcoxon matched-pairs signed-ranks test. A star indicates a statistically significant difference at 0.05 level.

used for training and testing are different, our query expansion mechanism on fields can still outperform the baseline. Overall, our query expansion mechanism on fields has achieved effective performance in our experiments on the two TREC Web collections used.

4. Adaptive query expansion

In this section, we propose a low-cost adaptive query expansion mechanism, which applies both the collection enrichment and selective query expansion techniques. The idea is to choose an appropriate resource rather than using a unique collection, either the local collection or an external collection, for query expansion. The proposed adaptive mechanism also uses a query performance prediction technique. We describe the proposed mechanism in Section 4.1, and present the related evaluation in Section 4.2.

4.1. Decision mechanism

We propose a new low-cost adaptive decision mechanism for the application of query expansion on a per-query basis. The decision mechanism is based on a pre-retrieval performance prediction technique. We hypothesise that a high quality external collection could bring more useful information for query expansion and lead to a better query term reweighting. Therefore, it could retrieve more relevant documents. The expansion of the query can be either *local*, using documents from the local collection, or *external*, using collection enrichment. The adaptive mechanism predicts the benefit of query expansion using either option, and uses the most effective one. If both options are predicted to lead to the degradation of the query performance, then query expansion is *disabled*.

We use the Average Inverse Collection Term Frequency (AvICTF) (He & Ounis, 2005) to predict the quality of the top-ranked documents, which is the average of the inverse collection term frequency (ICTF) (Kwok et al., 1996) of each query term. The definition of AvICTF is as follows:

$$\text{AvICTF} = \frac{\log_2 \prod_Q \frac{\text{token}_{\text{coll}}}{F}}{ql} \tag{9}$$

In the above definition, Q refers to the query, $\text{token}_{\text{coll}}$ the number of tokens in the whole collection, ql the query length, and F is the frequency of the term in the collection. By definition, a high AvICTF value predicts a good query performance, and a low AvICTF value predicts a poor query performance. From this definition, given a query, the AvICTF's value is comparable for different collections. Therefore, we can devise a unique thresholding mechanism over the two collections for the adaptive query expansion.

We could have used other query performance predictors, such as the query difficulty (Amati, 2003) and the clarity score (Cronen-Townsend et al., 2002). However, unlike these predictors, using AvICTF avoids the actual retrieval on both local and external collections before making a decision for query expansion. It uses only the collection statistics of the query terms to predict the query performance. Therefore, our adaptive query expansion mechanism is very low-cost. Besides, as shown in He and Ounis (2005), it is a reliable query performance predictor.

Using the AvICTF predictor, the decision mechanism operates on a per-query basis, and can be described as follows:

- First, given a local collection and an external one, we compute the AvICTF on both local and external collections.
- If the AvICTF values on both collections are lower than a threshold, we disable query expansion.
- Otherwise, we apply query expansion on the collection that has the highest AvICTF value.

In the next section, we conduct experiments to evaluate our adaptive query expansion mechanism.

4.2. Evaluation of the adaptive query expansion mechanism

In this section, we evaluate our adaptive query expansion mechanism. We introduce the experimental setting in Section 4.2.1 and present the results in Section 4.2.2.

Table 5
The obtained mean average precision (MAP) by the query expansion on fields (QEF), query expansion using external collection only (ExQEF) and the adaptive query expansion (AdapQEF), respectively

Coll.	QEF	ExQEF	AdapQEF	Δ%	*p*-value
WT10G	0.2197	0.2261	0.2362	+4.47	2.7e − 10*
.GOV2	0.3528	0.3265	0.3528	0	–

The *p*-values are given by the sign test between AdapQEF and the best between QEF and ExQEF. Δ% indicates the difference between AdapQEF and the best retrieval performance between QEF and ExQEF. A star indicates a statistically significant difference at 0.05 level.

4.2.1. Experimental settings

In this section, the used local collections and topics, the weighting models and their settings are the same as in Section 3.3.

To create an external collection for the collection enrichment, we need to make sure that the external collection is of a high quality so that it can bring useful information for query expansion. We suggest that the TREC newswire collections and the English Wikipedia domain are good resources. Therefore, we form an external collection by merging the TREC disks 1–5 collections[4] and a crawl of the English Wikipedia[5] in early July, 2005. The TREC disks 1–5 collections contain newswire articles from various sources, e.g. the Financial Times, the Wall Street Journal, etc. The English Wikipedia contains explanations of various concepts, and is usually considered to be of a good quality, making it an appropriate external resource. The merged external collection consists of 3,445,344 documents, which is approximately twice as large as the WT10G collection, while being much smaller than the .GOV2 collection. As stressed in Kwok and Chan (1998), a relatively large collection is more likely to contain more relevant documents. In other words, if the external collection is much smaller than the local collection, it is unlikely to have a positive impact on the query expansion's performance. Hence we expect the adaptive query expansion mechanism to improve the retrieval performance on WT10G, while having little impact on the large-scale .GOV2 collection.

For the optimisation of the threshold setting of the adaptive query expansion mechanism, we do a data sweeping in the range of $[10, 27]$ with an interval of 1. For all TREC topics used, the AvICTF values on both local and external collections are within this range. We choose the threshold setting that gives the best mean average precision (MAP) on the test topics using relevance assessments.

Our baseline is the best performing query expansion on fields using either the local collection (QEF) or the external collection (ExQEF). This allows us to assess whether the adaptive query expansion mechanism has the potential for further improvement of the retrieval performance based on our query expansion mechanism on fields. Section 3.3, for comparison purposes, we compare the performance of our adaptive query expansion with the performance achieved by the best TREC submitted title-only runs on a per-task basis.

4.2.2. Experimental results

Table 5 presents the evaluation results. The *p*-value is computed by the sign test, which indicates if the adaptive query expansion mechanism provides a positive improvement, compared with the baseline, for a statistically significant number of queries.

On the WT10G collection (see row WT10G in Table 5), we observe a 4.47% statistically significant improvement on the baseline. The optimal threshold setting is $\text{AvICTF} = 10$. The adaptive query expansion mechanism applies the local query expansion for 59 queries, and the external query expansion for 41 queries. It does not disable query expansion for any query. The performance of the adaptive query expansion is better than expanding all the queries locally or externally, which indicates that our adaptive query expansion mechanism successfully selects the appropriate collection resource for query expansion. The adaptive query expansion mechanism makes the correct decision for 81 out of 100 queries. The sign test shows that the mechanism makes the correct decisions for a statistically significant number of queries at the 0.05 level.

However, for the large-scale .GOV2 collection, we do not observe an improvement using the adaptive query expansion mechanism. It results in the best MAP when the local query expansion is applied for all the queries.

[4] Related information can be found from http://www-nlpir.nist.gov/trec/data/docs_eng.html.
[5] Related information can be found from http://en.wikipedia.org/wiki/Main_Page.

We suggest that this is because the .GOV2 collection is so large that the used external resource cannot render more relevant documents and better collection statistics for query term reweighting. This is consistent with the suggestion in Kwok and Chan (1998), and our expectation in Section 4.2.1.

In summary, our adaptive query expansion is shown to be effective when the external collection is larger than the local collection. Our adaptive query expansion mechanism is shown to have the potential to further improve the retrieval performance based on the query expansion mechanism on fields. However, the experimental results suggest not to apply collection enrichment if the external collection is much smaller than the local one. Further research is required to study whether a huge collection, such as the Web, would benefit from the adaptive query expansion mechanism. In particular, the choice of the external collection may change the performance, whatever the size of the local collection.

5. Experiments with BM25F

In this section, we conduct experiments with the BM25F model to see the impact of using a different document weighting model on the proposed query expansion techniques. In the query expansion mechanism on fields and the adaptive query expansion, we replace DLHF with BM25F for both the first-pass retrieval and the retrieval using the expanded query. The experimental settings, including the collections, topics, and the term weighting model (i.e. Bo1F) used, are the same as in Sections 3.3.1 and 4.2.1.

Using BM25F, we need to optimise five parameters, which are the hyper-parameter b_f for the three fields (see Eq. (2)) and the weights of anchor text (w_a) and title (w_t) fields. To optimise the hyper-parameter b_f, for each field f, we set w_f to 1 and the weights of the other two fields to 0. Then, we choose the b_f value that gives the best mean average precision (MAP) on the test topics in a data sweeping within $[0,1]$ with an interval of 0.05. The optimal b_f values are given in Table 6. When the hyper-parameters b_f of all the three fields are optimised, we perform a two-dimensional data sweeping to optimise w_a and w_t in $[0,2]$ with an interval of 0.1. If the optimal weights are not found in this range, we increase the upper bound of the range. Again, we choose the weights that give the best MAP on the test topics.

Using BM25F, Table 7 compares the performance between our query expansion mechanism on fields and the content-based query expansion (i.e. the baseline). Compared with the baseline, we observe a statistically

Table 6
The optimised values of the parameter b_f of BM25F for the body (b_{body}), anchor (b_{anchor}), and title (b_{title}) fields

TREC	b_{body}	b_{anchor}	b_{title}
TREC 9	0.20	1	0.25
TREC 10	0.40	0.85	0.30
Merged	0.30	0.95	0.30
TREC 13	0.40	0.70	0.25
TREC 14	0.40	1.00	0.40
Merged	0.40	0.70	0.35

Table 7
The mean average precision obtained by the content-based query expansion (QE) and by our query expansion mechanism on fields (QEF), respectively

TREC	QE	QEF	w_a	w_t	Δ%	p-Value	+	−	=
TREC 9	0.2192	0.2368	0.1	1.1	+8.03	1.114e − 06*	40	5	5
TREC 10	0.2431	0.2475	0.2	0.7	+1.40	3.623e − 04*	36	11	3
Merged	0.2311	0.2368	0.0	1.0	+2.34	4.708e − 09*	79	12	9
TREC 13	0.2715	0.2840	0.3	5.0	+4.60	0.09255	29	19	2
TREC 14	0.3620	0.3709	0.2	2.7	+2.46	0.0672	30	19	1
Merged	0.3172	0.3234	0.5	0.5	+1.95	0.7147	49	46	5

w_a and w_t are the optimal weights of anchor text and title, respectively. The p-value is given by the Wilcoxon matched-pairs signed-ranks test. A star indicates a statistically significant difference at 0.05 level. +, −, and = indicate for how many queries QEF results in improvement, degradation and no change in the retrieval performance, respectively.

Table 8
The mean average precision obtained by the content-based query expansion (QE) and by our query expansion mechanism on fields (QEF), respectively

TREC	QE	QEF	Δ%	*p*-Value
TREC 9	0.2263	0.2329	+2.92	0.1583
TREC 10	0.2370	0.2405	+1.48	0.6407
TREC 13	0.2715	0.2855	+5.16	0.0105*
TREC 14	0.3620	0.3666	+1.27	0.2444

w_a and w_t are the optimal weights of anchor text and title, respectively. The *p*-value is given by the Wilcoxon matched-pairs signed-ranks test. A star indicates a statistically significant difference at 0.05 level.

Table 9
The obtained mean average precision (MAP) using *BM25F* for the evaluation of the adaptive query expansion mechanism

Coll.	QEF	ExQEF	AdapQEF	Δ%	*p*-Value
WT10G	0.2376	0.2374	0.2460	+3.54	6.15e − 11*
.GOV2	0.3251	0.3191	0.3251	0	–

The notations in this table are the same as in Table 5.

significant improvement on the WT10G collection. However, on .GOV2, unlike the obtained results using DLHF (Table 3), we do not observe a significant improvement using BM25F. Moreover, it seems that using BM25F, it is necessary to give higher weights to the anchor text and title fields. In particular, the optimal weight of the title (w_t) field is much larger than that of DLHF (see columns w_t in Tables 3 and 7).

Table 8 contains the evaluation results of a two-fold holdout evaluation, which uses half of the test queries for training, and the other half for testing. From the results, we find that the query expansion mechanism on fields outperforms the baseline in all the four cases. However, the difference between the obtained MAPs are statistically significant in only one case (see the *p*-value with a star in Table 8). Overall, compared with our experimental results using DLHF in Section 3.3.2, we find that BM25F leads to a relatively small improvement in the retrieval performance, brought by the query expansion mechanism on fields. This may be due to the fact that BM25F employs a per-field normalisation component, while DLHF does not have a term frequency normalisation component. Consequently, in BM25F, the field evidence is more exploited than in DLHF, which results in a better retrieval performance of the baseline.

Table 9 evaluates the adaptive query expansion mechanism using BM25F. From the table, we have similar observations with those obtained using DLHF. Again, we obtain a significant improvement on WT10G, but not on the large-scale .GOV2 collection.

In summary, for the query expansion mechanism on fields, BM25F seems to give more importance to the anchor text and title fields than DLHF. The optimal weights of these two fields are much larger than those of DLHF. Moreover, using BM25F, the adaptive query expansion mechanism improves the retrieval performance on WT10G, but not on the large-scale .GOV2 collection, which is similar with what we obtained using DLHF. Overall, the proposed query expansion techniques can achieve an improvement on the baselines using both document weighting models, although the improvement depends on the collections used.

6. Conclusions and future work

In this paper, we have improved query expansion by combining evidence from different fields. We suggest that the performance of query expansion is related not only to the quality of the top-ranked documents, but also to the quality of the reweighting of the query terms. Therefore, combining field evidence can refine the statistics for query term reweighting, and improve query expansion. The proposed query expansion mechanism on fields is evaluated on two standard TREC Web collections, namely WT10G and the large-scale .GOV2 collection. The experiments are conducted using two different document weighting models, namely DLHF and BM25F. The evaluation results show that our query expansion mechanism on fields has achieved

a significant retrieval performance improvement compared with the content-based query expansion, except when using BM25F on WT10G, where its performance is comparable with the content-based query expansion. Also, we have shown that the proposed query expansion mechanism on fields has a good potential for enhancing the retrieval performance, and can outperform the best official TREC runs, if the weights of the fields are properly set.

Moreover, we have devised an adaptive query expansion mechanism that automatically selects the appropriate collection resource for query expansion. The proposed adaptive mechanism applies query performance prediction, selective query expansion, and collection enrichment techniques. In our experiments, we observe that the performance of the adaptive query expansion mechanism depends on the relative size of the external collection. This is consistent with the suggestion in Kwok and Chan (1998). In addition, on the WT10G collection, the adaptive query expansion mechanism gives consistent retrieval performance improvement, regardless of the document weighting model used.

In the future, we aim to further investigate the adaptive query expansion mechanism on very large-scale collections and using various external collections. For example, for .GOV2, we can use the SPIRIT collection (Joho & Sanderson, 2004) as an external resource for query expansion. The SPIRIT collection contains 94,552,870 documents and is approximately four times larger than the .GOV2 collection.

Moreover, this paper has focused on showing that query expansion on fields is beneficial and has a considerable potential for improving the retrieval performance. On the other hand, the setting of the fields' weights is crucial to the retrieval performance of the proposed query expansion techniques. For example, BM25F and DLHF require different optimal weights for the anchor text and title fields. Moreover, for the same document weighting model, the optimal weights are different on the two collections used. In the future, we plan to investigate a method for automatically setting the weights of the fields without assuming that the relevance assessments are readily available. This allows the proposed techniques to be used in an operational setting, where little relevance information is available.

Finally, in our experiments, we found that the effectiveness of the adaptive query expansion mechanism depends on the size of the external collection used for query expansion. In the future, we plan to study possible reasons that explain this finding. For example, a large external collection may contain more relevant documents, or may have a better global statistics than the local collection, which improves the query term reweighting process.

References

Amati, G. (2003). Probabilistic Models for Information Retrieval based on Divergence from Randomness. PhD thesis, Department of Computing Science, University of Glasgow.

Amati, G. (2006). Frequentist and bayesian approach to information retrieval. In *Advances in information retrieval, 28th European conference on IR research, ECIR 2006* (pp. 13–24). London, UK.

Amati, G., Carpineto, C., & Romano, G. (2004a). Query difficulty, robustness, and selective application of query expansion. In *Advances in information retrieval, 26th European conference on IR research*, ECIR 2004 (pp. 127–137). Sunderland, UK.

Amati, G., Carpineto, C., & Romano, G. (2004b). Fondazione Ugo Bordoni at TREC 2004. In *Proceedings of the 13th text retrieval conference (TREC 2004)*. Gaithersburg, MD.

Attardi, G., Esuli, A., & Patel, C. (2004). Using Clustering and Blade Clusters in the Terabyte Task. In *Proceedings of the 13th text retrieval conference (TREC 2004)*. Gaithersburg, MD.

Clarke, C., Craswell, N., & Soboroff, I. (2004). Overview of the TREC 2004 Terabyte Track. In *Proceedings of the 13th text retrieval conference (TREC 2004)*. Gaithersburg, MD.

Clarke, C., Scholer, F., & Soboroff, I. (2005). The TREC 2005 Terabyte Track. In *Proceedings of the 14th text retrieval conference (TREC 2005)*. Gaithersburg, MD.

Cronen-Townsend, S., Zhou, Y., & Croft, B. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 299–306). Tampere, Finland.

Cronen-Townsend, S., Zhou, Y., & Croft, B. (2004). A framework for selective query expansion. In *Proceedings of the 2004 ACM CIKM international conference on information and knowledge management* (pp. 236–237). Washington, DC.

Grunfeld, L., Kwok, K. L., Dinstl, N., & Deng, P. (2003). TREC 2003 Robust, HARD and QA track experiments using PIRCS. In *Proceedings of the 12th text retrieval conference (TREC 2003)*. Gaithersburg, MD.

Hawking, D. (2000). Overview of the TREC-9 web track. In *Proceedings of the ninth text retrieval conference (TREC 9)*. Gaithersburg, MD.

Hawking, D., & Craswell, N. (2001). Overview of the TREC-2001 web track. In *Proceedings of the tenth text retrieval conference (TREC 2001)*. Gaithersburg, MD.

He, B., & Ounis, I. (2005). Query performance prediction. In *Information systems, special issue for the string processing and information retrieval: 11th international conference (SPIRE2004)*.

Joho, H., & Sanderson, M. (2004). The SPIRIT collection: an overview of a large web collection. *SIGIR Forum, 38*(2), 57–61.

Kwok, K. L. (1996). A new method of weighting query terms for ad-hoc retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 187–195). Zurich, Switzerland.

Kwok, K. L., & Chan, M. (1998). Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 250–256). Melbourne, Australia.

Kwok, K. L., Grunfeld, L., Sun, H. L., & Deng, P. (2004). TREC 2004 Robust track experiments using PIRCS. In *Proceedings of the 13th text retrieval conference (TREC 2004)*. Gaithersburg, MD.

Macdonald, C., He, B., Plachouras, V., & Ounis, I. (2005). University of Glasgow at TREC 2005: Experiments in terabyte and enterprise tracks with terrier. In *Proceedings of the 14th text retrieval conference (TREC 2005)*. Gaithersburg, MD.

Plachouras, V., He, B., & Ounis, I. (2004). University of Glasgow at TREC 2004: Experiments in web, robust, and terabyte tracks with terrier. In *Proceedings of the 13th text retrieval conference (TREC 2004)*. Gaithersburg, MD.

Rocchio, J. (1971). Relevance feedback in Information Retrieval. In G. Salton (Ed.), *The SMART retrieval system: Experiments in automatic document processing* (pp. 313–323). Prentice-Hall Englewood Cliffs.

Robertson, S. E., Walker, S., Beaulieu, M., Gatford, M., & Payne, A. (1995). Okapi at TREC-4. In *Proceedings of the fourth text retrieval conference (TREC-4)*. Gaithersburg, MD.

Robertson, S. E., Zaragoza, H., & Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *Proceedings of the 13th ACM international conference on information and knowledge management (CIKM 2004)* (pp. 42–49). Washington, DC.

Silverstein, C., Henzinger, M., & Marais, H. (1998). Analysis of a very large AltaVista query log. Digital SRC Technical Note #1998-014. Digital SRC.

Singhal, A., Choi, J., Hindle, D., & Lewis, D. (1998). AT&T at TREC-7. In *Proceedings of the seventh text retrieval conference (TREC 7)*. Gaithersburg, MD.

Yom-Tov, E., Fine, S., Carmel, D., & Darlow, A. (2005). Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 512–519). Salvador, Brazil.

Zaragoza, H., Craswell, N., Taylor, M., Saria, S., & Robertson, S. E. (2004). Microsoft Cambridge at TREC 13: Web and hard tracks. In *Proceedings of the 13th text retrieval conference (TREC 2004)*, Gaithersburg, MD.