A Study of Selective Collection Enrichment for Enterprise Search

Jie Peng, Craig Macdonald, Ben He, Iadh Ounis
Department of Computing Science
University of Glasgow, Scotland, UK
{pj,craigm, ben, ounis}@dcs.gla.ac.uk

ABSTRACT

Enterprise intranets are often sparse in nature, with limited use of alternative lexical representations between authors, making query expansion (QE) ineffective. Hence, for some enterprise search queries, it can be advantageous to instead use the well-known collection enrichment (CE) method to gather higher quality pseudo-feedback documents from a more diverse external resource. However, it is not always clear for which queries the collection enrichment technique should be applied. In this paper, we study two different approaches, namely a predictor-based approach and a divergence-based approach, to decide on when to apply CE. We thoroughly evaluate both approaches on the TREC Enterprise track CERC test collection and its corresponding topic sets, in combination with three different external resources and nine different query performance predictors. Our results show that both approaches are effective to selectively apply CE for enterprise search. In particular, the divergence-based approach leads to consistent and marked retrieval improvements over the systematic application of QE or CE on all external resources.

Categories & Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval

General Terms: Experimentation, Performance

Keywords: Query expansion, Selective application, Collection enrichment, Enterprise search

1. INTRODUCTION

Query expansion (QE) usually refers to a technique that expands the original query with new terms, by taking into account a set of top ranked documents, called the pseudo-relevant set [12]. The effectiveness of query expansion is correlated with the quality of the pseudo-relevant set returned by the first-pass retrieval. The better the pseudo-relevant set is, the more likely that useful expansion terms are added to the query and the better the performance that query expansion can achieve [2].

CIKM'09, November 2–6, 2009, Hong Kong, China.

In general, query expansion is effective for enterprise document search [1]. However, for some enterprise search queries, query expansion may fail. This is explained by the fact that an intranet collection generally reflects the view of the organisation that it serves and that content generation often tends to be autocratic or bureaucratic rather than democratic [5]. This leads to a limited use of alternative lexical representations, restricting the usefulness of query expansion. For the aforementioned queries, it may be advantageous to use instead the well-known collection enrichment (CE) method [4, 8], which performs query expansion using a larger and higher-quality external resource and then retrieves from the local collection using the expanded query.

However, it is not always clear when and for which queries CE should be applied instead of QE to improve retrieval effectiveness in enterprise search. To the best of our knowledge, these issues have never been investigated in the enterprise search literature. In this paper, we study two different selective application methods for enterprise search, with the aim to develop a decision mechanism informing on when CE is more likely to be beneficial than QE. In particular, we investigate the effectiveness of both approaches on three different external resources. We conduct our experiments on the TREC Enterprise CERC test collection. There are two key contributions of this work. First, we study the performances of two different approaches for selectively applying CE, namely the predictor-based and the divergence-based approaches, on three different external resources for enterprise search. Second, we show that the effectiveness of CE is highly dependent on the selected external resource.

2. SELECTIVE CE APPROACHES

We study two different selective application methods for enterprise search. Macdonald et al. [9] proposed the use of query performance predictors to selectively apply CE for ad-hoc Web search. In particular, this approach uses a query performance predictor to predict a given query's performance on both the internal and external resources. Then, it decides whether or not to apply CE based on the predicted performances. We evaluate this approach in the context of enterprise search. Peng & Ounis [11] proposed a method that uses the divergence between relevance score distributions prior to and after applying a given document feature, to selectively apply an appropriate document feature on a per-query basis. In this paper, we adapt the divergence-based approach to selective CE by examining the divergence between relevance score distributions prior to, and after the application of CE.

2.1 Predictor-based Approach

A query performance predictor predicts a given query's performance on a specific collection. Many predictors have been previously proposed in information retrieval (IR) [7, 6, 13]. They are generally classified into two types: pre-retrieval predictors and post-retrieval predictors. Generally speaking, pre-retrieval predictors only rely on the statistics of the collection while post-retrieval predictors are more reliant on the statistics of the top ranked documents for the query [7].

Using query performance predictors, Macdonald et al. [9] proposed a decision mechanism for ad-hoc Web search to decide whether or not to apply collection enrichment on a per-query basis. The approach is based on the predicted performance score of a given query on the local and external resources. In particular, the decision mechanism applies collection enrichment if and only if the predicted query performance score obtained on the external resource is higher than a threshold, as well as the predicted query performance score obtained using the local resource.

It is of note, however, that for some query performance predictors, the lower the predictor score for the external resource, the higher the query performance on that resource is predicted to be, and hence the more beneficial collection enrichment using that resource should be. For example, for the clarity score predictor [3], a lower query performance score on the external resource means a higher similarity between the query language model and the external collection's language model, suggesting that applying CE could bring more useful expansion terms. Hence, in applying the predictor-based approach, the nature of the used predictor is taken into account when comparing the query performance scores.

2.2 Divergence-based Approach

Different from the predictor-based approach, which is based on the predicted query performance scores on internal or external resources, Peng & Ounis proposed a divergence-based approach (DBA) [11] for the selective application of query-independent features. The approach estimates the divergence between the relevance score distributions prior to and after applying a given query-independent feature. Then, it uses the estimated divergence scores to choose an appropriate query-independent feature for a given query based on a previously learnt divergence distribution. Peng & Ounis showed that the distribution of these estimated divergence scores can be used to selectively choose an appropriate query-independent feature on a per-query basis [11].

In this work, we apply the divergence-based approach for selectively applying CE. Instead of selectively choosing a query-independent feature, we use the distribution of divergence scores to decide whether or not to apply CE for a given query. To achieve this, we learn the distribution of divergence scores, which are estimated between two different lists of ranked documents obtained with and without the application of QE or CE, using training data. Then, for a new query, we apply CE when it exhibits the divergence score with a higher training performance compared to that of QE.

3. EXPERIMENTAL SETTING

In the following, we address two research questions: firstly, we investigate how effective the two aforementioned approaches are for selectively applying CE; secondly, we investigate the importance of the used external resource for CE.

We use the standard TREC CERC test collection and its corresponding TREC 2007 & 2008 Enterprise track document search tasks, which have 42 & 63 query topics with relevance assessments, respectively. These topics have title and narrative fields. However, for a realistic setting, we use query terms from only the title field. We experiment with three different external resources, namely Wikipedia, and the standard Aquaint2 & .GOV collections. The Wikipedia corpus used in this paper is a snapshot from August 2008. We use both pre-retrieval and post-retrieval query performance predictors, which include: $AvICTF$, $\gamma 1$, $\gamma 2$ and QS [7]; $AvIDF$ and $AvPMI$ [6]; WIG and QF [13]; CS [3].

For indexing and retrieval, we use the Terrier IR platform [10]. All corpora are indexed by removing standard stopwords and applying Porter's stemming algorithm for English. We index the body, anchor text and titles of documents as separate fields. For Wikipedia, we ignore the anchor text field as our initial experiments found that it does not improve the retrieval performance, while the Aquaint2 newswire corpus does not have any anchor text. Documents are ranked using the PL2F field-based Divergence From Randomness document weighting model [9].

In the experiments, we conduct testing on the TREC 2008 dataset after training on the TREC 2007 dataset. The parameters that are related to the PL2F document weighting model, the predictors and the threshold of the decision mechanism are set by optimising Mean Average Precision (MAP) on the training dataset, using a simulated annealing procedure. Finally, for the post-retrieval predictors, the number of top ranked documents is also set by optimising MAP over the training dataset, using a large range of different settings. The evaluation measures used in all our experiments are the MAP and the normalised Discounted Cumulative Gain (nDCG). We report the obtained results, and their analysis in the next section.

4. RESULTS & ANALYSIS

4.1 Effectiveness of Selective CE

We investigate how effective the predictor-based and the divergence-based approaches are for selectively applying CE on a per-query basis by comparing them with the systematic application of QE on the enterprise collection, or the systematic application of CE on the external resource. In addition, in order to assess how important it is to selectively apply collection enrichment on a per-query basis for enterprise search, we simulate an oracle system, which applies CE only for the queries where CE brings more benefit to the retrieval performance compared to QE, while it applies QE on the enterprise collection for the remaining queries.

Table 1 presents the evaluation of the selective application of CE on the three used external resources by using the predictor-based and the divergence-based approaches on the TREC 2008 dataset. The best retrieval performance in each column is highlighted in bold. From the results, we note that the systematic application of QE consistently outperforms the systematic application of CE, regardless of the used external resource. $\uparrow$ denotes that the obtained retrieval performance by using the selective application technique is better than the systematic application of QE. Values that are statistically better than the systematic application of QE, according to the Wilcoxon Matched-pair Signed-Ranks test, are marked with $*$. The *Acc.* columns show the accuracy of the selective application techniques, obtained by calculating the proportion of queries with a correct application of CE.

Next, we examine selective CE using the predictors. From the results, we observe that $AvICTF$, $AvIDF$, $AvPMI$ and

TREC 2008 Enterprise Document Search Task									
	Wikipedia			Aquaint2			.GOV		
	MAP	nDCG	Acc.	MAP	nDCG	Acc.	MAP	nDCG	Acc.
PL2F	0.3629	0.5502	-	0.3629	0.5502	-	0.3629	0.5502	-
+QE	0.3811	0.5646	-	0.3811	0.5646	-	0.3811	0.5646	-
+CE	0.3684	0.5606	-	0.3402	0.5391	-	0.3583	0.5551	-
Oracle	0.4204	0.5978	-	0.4022	0.5832	-	0.4135	0.5930	-
Pre-retrieval Predictors									
AvICTF	0.3660	0.5567	40.98%	0.3576	0.5500	45.00%	0.3765	0.5570	55.73%
AvIDF	0.3691	0.5560	42.62%	0.3566	0.5509	46.66%	0.3768	0.5601	52.45%
$\gamma 1$	0.3944 $\uparrow$	0.5678 $\uparrow$	67.21%	0.3909 $\uparrow$	**0.5729** $\uparrow$	65.00%	0.3885 $\uparrow$	0.5687 $\uparrow$	62.29%
$\gamma 2$	0.3959 $\uparrow *$	0.5679 $\uparrow$	67.21%	0.3825 $\uparrow$	0.5673 $\uparrow$	58.33%	0.3864 $\uparrow$	0.5658 $\uparrow$	59.01%
AvPMI	0.3740	0.5613	47.54%	0.3700	0.5621	56.66%	0.3769	0.5587	44.26%
QS	0.3681	0.5590	40.98%	0.3416	0.5445	43.33%	0.3638	0.5555	49.18%
Post-retrieval Predictors									
CS	0.3824 $\uparrow$	0.5653 $\uparrow$	49.18%	0.3658	0.5475	51.66%	0.3941 $\uparrow$	**0.5782** $\uparrow$	67.21%
WIG	0.3773	0.5651 $\uparrow$	47.54%	0.3773	0.5507	55.00%	0.3706	0.5612	52.45%
QF	0.3814 $\uparrow$	0.5612	57.37%	0.3817 $\uparrow$	0.5585	55.00%	0.3807	0.5643	59.01%
Divergence-based Approach									
DBA	**0.4081** $\uparrow$	**0.5803** $\uparrow$	62.29%	**0.3914** $\uparrow$	0.5713 $\uparrow$	65.00%	**0.3996** $\uparrow *$	0.5765 $\uparrow$	67.21%

Table 1: Evaluation of the selective application of CE on the TREC 2008 Enterprise document search task.

QS consistently decrease the retrieval performance across 3 different external resources compared to the systematic application of QE. However, other predictors exhibit improvements. In particular, the $\gamma 1$ and $\gamma 2$ predictors exhibit consistent improvements over the systematic application of QE (this is statistically significant on the Wikipedia resource). Overall, these results suggest that the choice of query performance predictor is very important for the predictor-based approach.

We also observe that the retrieval performance obtained by using the selective divergence-based approach using three different external resources always improves over the systematic application of QE in terms of MAP and nDCG. However, such improvements are not always statistically significant, probably because of the small size of the test dataset. Nevertheless, in most cases, the best retrieval performance on each external resource is achieved by the divergence-based approach, and the highest MAP score is also obtained by using the divergence-based model.

From the *Acc.* column, we can see that, in most cases, the accuracy of the selective application of CE by using the predictor-based and the divergence-based approaches is markedly higher than a random guess. In order to better understand the accuracy of the selective application techniques, we plot their accuracy in the Receiver Operating Characteristics (ROC) space based on different external resources for the TREC 2008 dataset, as shown in Figure 1. Y and X axes provide True Positive Rate (TPR) and False Positive Rate (FPR), respectively. The TPR determines a classifier or a diagnostic test performance on classifying positive instances (CE in this case) correctly among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples (QE in this case) available during the test. The dashed line represents a random guess. For points above the dashed line, the further the distance from the line is, the higher the accuracy achieved.

From Figure 1, we can see that most predictors are not always above the dashed line. For example, the QF predictor points are above the dashed line on the Wikipedia and .GOV collections, however, it performs worse than a random guess on the Aquaint2 collection. Moreover, the accuracy of some other predictors is close to a random guess, e.g. the accuracy of the CS predictor is close to the dashed line for both the Wikipedia and Aquaint2 collections. Among nine different predictors, only the $\gamma 1$, $\gamma 2$ and WIG predictors have consistently marked improvements over a random guess. In addition, we also observe that the accuracy of the divergence-based approach is consistently and markedly better than a random guess on all external resources. However, it is of note that as they are pre-retrieval predictors, $\gamma 1$ and $\gamma 2$ are more efficient than WIG, which requires a first-pass retrieval.

The above observations attest that both the predictor-based and the divergence-based approaches are effective for selectively applying CE for enterprise search. In particular, the γ predictors are the most efficient and effective among the nine used predictors. Moreover, the divergence-based approach is more robust than the predictor-based approach as it enhances the retrieval performance across all external resources.

4.2 Importance of the External Resource

The performance of collection enrichment is dependent on the usefulness of the expansion terms extracted from the pseudo-relevant set, which is typically obtained from the top ranked documents on the external resource. In this part, we investigate the importance of the external resource for the selective application of CE.

From Table 1, we notice that the systematic application of CE can decrease the retrieval performance over the PL2F baseline. For example, for the nDCG evaluation measure, the retrieval performance decreases from 0.5502 to 0.5391 after applying CE on the Aquaint2 collection for all queries. However, systematically applying CE on the Wikipedia collection always enhances the retrieval performance in terms of both MAP and nDCG.

Moreover, from Table 1, we observe that the performance of the predictor-based approach is dependent on the selected

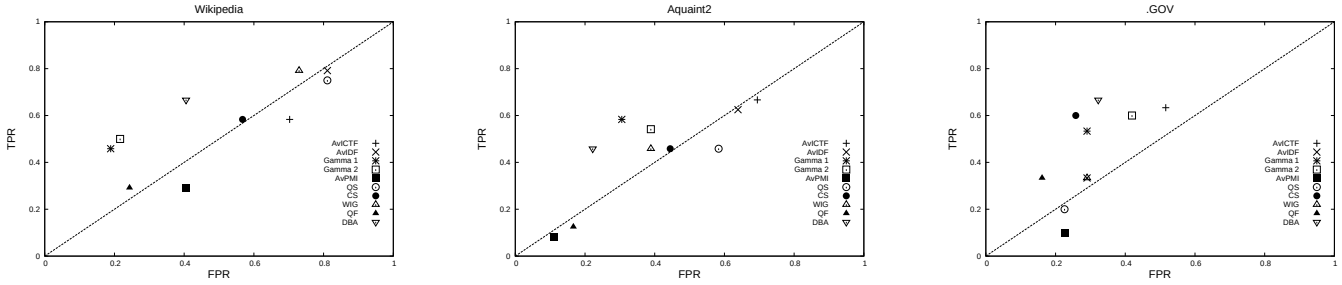

Figure 1: Comparison between different selective application techniques by plotting the ROC space for different external resources.

external resource, e.g. the CS predictor makes improvement when the external resource is Wikipedia or .GOV, whereas, the retrieval performance decreases when the Aquaint2 collection is used as the external resource. In addition, we also observe that the divergence-based approach is not highly dependent on the external resource chosen, as the retrieval performance can be consistently improved on all external resources. In particular, the best retrieval performance is obtained from Wikipedia. This is probably because of the scientific nature of the TREC 2008 Enterprise topics, which are possibly better covered in the encyclopedic Wikipedia.

The above observations suggest that choosing an appropriate external resource before applying selective collection enrichment is important and that Wikipedia seems to be the most appropriate one for the used enterprise collection.

5. CONCLUSION

In this paper, we have studied the effectiveness of two different selective CE approaches, namely the predictor-based and the divergence-based methods, for enterprise search. We have conducted the study on the TREC Enterprise CERC test collection and its corresponding topic sets, in combination with three different external resources. We used nine different query performance predictors, including both pre-retrieval and post-retrieval predictors.

Firstly, our experimental results showed that the retrieval performance for enterprise search can be markedly enhanced with the oracle selective application of CE. Secondly, by comparing the predictor-based and the divergence-based approaches with the systematic application of QE on the enterprise collection, we found that both the predictor-based and the divergence-based approaches are effective for selective CE. Moreover, for the predictor-based method, the γ predictors were the most efficient and effective among the nine used predictors. In addition, we also observed the robustness of the divergence-based approach as it always enhances retrieval performance on all external resources. Finally, we have investigated the importance of different external resources for selective CE. We found that Wikipedia is the most useful external resource for the used enterprise collection. This is probably because most of the TREC 2008 enterprise topics are of a scientific nature, which are well covered in Wikipedia articles.

6. REFERENCES

[1] K. Balog, K. Hofmann, W. Weerkamp and M. de Rijke. Query and document models for enterprise search. In *Proc. of TREC'07*, USA (2007).

[2] G. H. Cao, J. Y. Nie, J. F. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. of SIGIR'08*, Singapore (2008).

[3] S. Cronen-Townsend, Y. Zhou and W. B. Croft. Predicting query performance. In *Proc. of SIGIR'02*, Finland (2002).

[4] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proc. of SIGIR'06*, USA (2006).

[5] R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin and D. P. Williamson. Searching the workplace web. In Proc. of WWW'03, Hungary (2003)

[6] C. Hauff, V. Murdock and R. Baeza-Yates. Improved query difficulty prediction for the Web. In *Proc. of CIKM'08*, USA (2008).

[7] B. He and I. Ounis. Query performance prediction. *Information Systems*, 31(7):585-594 (2004).

[8] K.L. Kwok, L. Grunfeld, M. Chan, N. Dinstl and C. Cool. TREC-7 Ad-Hoc, high precision and filtering experiments using PIRCS. In *Proc. of TREC'98*, USA (1998).

[9] C. Macdonald, B. He, V. Plachouras and I. Ounis. University of Glasgow at TREC 2005: experiments in terabyte and enterprise tracks with Terrier. In *Proc. of TREC'05*, USA (2005).

[10] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *Proc. of OSIR'06*, USA (2006).

[11] J. Peng and I. Ounis. Selective application of query-independent features in Web information retrieval. In *Proc. of ECIR'09*, France (2009).

[12] J. Rocchio: Relevance feedback in information retrieval. In Gerard Salton (Ed.), The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, USA (1971).

[13] Y. Zhou and W. B. Croft. Query performance prediction in Web search environments. In *Proc. of SIGIR'07*, The Netherlands (2007).