Predicting the Usefulness of Collection Enrichment for Enterprise Search

Jie Peng, Ben He, and Iadh Ounis

Department of Computing Science,
University of Glasgow, G12 8QQ, UK
{pj, ben, ounis}@dcs.gla.ac.uk

Abstract. Query Expansion (QE) often improves the retrieval performance of an Information Retrieval (IR) system. However, as enterprise intranets are often sparse in nature, with limited use of alternative lexical representations between authors, it can be advantageous to use Collection Enrichment (CE) to gather higher quality pseudo-feedback documents. In this paper, we propose the use of query performance predictors to selectively apply CE on a per-query basis. We thoroughly evaluate our approach on the CERC standard test collection and its corresponding topic sets from the TREC 2007 & 2008 Enterprise track document search tasks. We experiment with 3 different external resources and 3 different query performance predictors. Our experimental results demonstrate that our proposed approach leads to a significant improvement in retrieval performance.

1 Introduction

Collections within enterprises are often characterised by their limited vocabulary, since they are written by a small number of people, following specific guidelines and aims. Therefore, while query expansion is usually effective in IR, the limited use of alternative lexical representations within enterprise collections could lead to poor pseudo-relevance sets. In this case, it seems intuitive to make use of the well-established Collection Enrichment (CE) technique, which performs query expansion on a larger and higher-quality external resource [1, 2]. The reformulated query is then used to retrieve documents from the local enterprise collection. However, the quality of the external collection is a key factor that affects the retrieval performance given by CE [2].

In this paper, we argue that the retrieval performance of document search within an enterprise can be further enhanced by applying CE in a selective manner on a per-query basis. The idea is that the usefulness of the local or external collection for QE varies from a query to another. Using query performance predictors, we propose a decision mechanism that indicates the appropriateness of the local and external collections for a given query. QE is then applied on the collection, either local or external, that is predicted to contain higher quality of relevant content for the query.

To the best of our knowledge, this is the first study that investigates the usefulness of selectively applying CE in an enterprise setting. The proposed selective CE mechanism is thoroughly evaluated on the standard CERC test collection and its corresponding topic sets from TREC Enterprise track 2007 & 2008. We apply three different external resources and three different query performance predictors. Our experimental results show that our selective application of CE provides a significantly better retrieval performance than an approach that systematically applies QE on either the local or external collection.

Table 1. The decision mechanism for the selective application of CE. *local*, *external* and *disabled* in the column *Decision* indicate expanding the initial query on the *local* resource, *external* resource and disabling the expansion, respectively.

$score_L > T$	$score_E > T$	$score_L > score_E$	Decision
True	True or False	True	local
True or False	True	False	external
False	False	True or False	disabled

2 Selective Collection Enrichment

Our decision mechanism enriches the enterprise collection only if the external resource is predicted to contain more relevant content to the query than the local collection. For a given query, we use query performance predictors to estimate the quality of the pseudo-relevance sets returned by either the local or the external collection. A lower score corresponds to a difficult query for that collection [3], while a higher score suggests a richer pseudo-relevance set. Using the query difficulty scores returned by the predictors, our decision mechanism applies QE on the collection that corresponds to the higher predictor score, i.e. the collection that is predicted to lead to a better retrieval performance.

In addition, if the predictor scores on both the local ($score_L$) and the external ($score_E$) collections are lower than a threshold (T), then query expansion is not applied for that given query. Table 1 summarises our proposed decision mechanism for the selective application of collection enrichment.

Table 2. Evaluation of the selective application of CE on the TREC 2008 enterprise document search task.

	Wikipedia			Aquaint 2			.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	-
MAX	0.4204 $\star$	0.5978 $\star$	100%	0.4022 $\star$	0.5832 $\star$	100%	0.4135 $\star$	0.5930 $\star$	100%
Selective CE by using Predictors									
AvICTF	0.3660 $\downarrow$	0.5567 $\downarrow$	40.98%	0.3576 $\downarrow$	0.5500 $\downarrow$	45.00%	0.3765 $\downarrow$	0.5570 $\downarrow$	55.73%
$\gamma 2$	**0.3959** $\uparrow$ $\star$	**0.5679** $\uparrow$	67.21% $*$	**0.3825** $\uparrow$	**0.5673** $\uparrow$	58.33%	0.3864 $\uparrow$	0.5658 $\uparrow$	59.01%
CS	0.3824 $\uparrow$	0.5653 $\uparrow$	49.18%	0.3658 $\downarrow$	0.5475 $\downarrow$	51.66%	**0.3941** $\uparrow$	**0.5782** $\uparrow$	67.21% $*$

3 Experimental Setting

Three popular query performance predictors, namely the Average Inverse Collection Term Frequency (AvICTF) and the $\gamma 2$ pre-retrieval predictors [3], and the Clarity Score (CS) post-retrieval predictor [4], are studied in this paper. These predictors have been widely applied in the literature and were shown to be generally effective in predicting query performance. Note that unlike the pre-retrieval predictors, the CS predictor involves a parameter that needs tuning. It is also of note that for the CS predictor, a lower score indicates a better retrieval performance [4]. Therefore, unlike for AvICTF and $\gamma 2$, the decision mechanism is reversed to favour collection with a lower predictor score. However, the principle of deciding on the use of the collection enrichment is the same.

We use the standard CERC enterprise test collection [5], and its corresponding title-only topics from the TREC Enterprise track 2007 & 2008, respectively. There are 42 and 63 judged topics from TREC 2007 & TREC 2008, respectively. We experiment with three external resources, namely Wikipedia[1], Aquaint 2[2], and the TREC .GOV collection. For indexing and retrieval, we use the Terrier IR platform[3], and apply standard stopword removal and the Porter's stemming algorithm for English. For the CERC and .GOV collections, we index the body, anchor text and titles of the documents as separate fields. For the Wikipedia and the Aquaint 2 collections, we do not use the anchor text field as our initial experiments show that it is not beneficial for retrieval. We use the Bo1 term weighting model for query expansion [6]. Documents are ranked using the PL2F field-based DFR document weighting model [7]. The parameters that are related to the PL2F document weighting model, the CS predictor and the threshold T of the decision mechanism are set by optimising MAP on the TREC 2007 dataset, using a simulated annealing procedure [8]. We evaluate our method using the TREC 2008 topics.

4 Experimental Results

Table 2 presents the evaluation results of our proposed method. As shown in the table, the use of query expansion on the enterprise collection (PL2F+QE) outperforms PL2F, as well as a system that systematically applies QE on the external collection (denoted PL2F+CE) across three different external resources. Hence, in order to compare our proposed method with a strong baseline, we use PL2F + QE as our baseline. $\uparrow$ & $\downarrow$ denote that the obtained retrieval performance by using the predictor is better (resp. worse) than the baseline. The *Acc.* column shows the accuracy of our proposed method, which is given by the number of queries that has been appropriately applied with CE divided by the total number of queries. The symbol $*$ denotes that the predictor makes a correct prediction for a statistically significant number of queries, according to the Sign Test ($p < 0.05$). Values that are statistically better than the baseline are marked with $\star$ (Wilcoxon Matched-Pairs Signed-Ranks Test, $p < 0.05$).

Firstly, we assess how important it is to selectively apply CE on a per-query basis, by estimating the MAP and nDCG upper bounds (highlighted with underline). In this case, if query expansion is deemed helpful for the query, we manually select the most appropriate collection, from which to build the pseudo-relevance set. From Table 2, it is clear that the retrieval performance with the manual selective application of CE leads to a significant improvement over the systems that apply systematically either QE or CE, across three different external resources. This suggests that identifying the most appropriate collection for query expansion on a per-query basis is useful.

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download

[2] http://trec.nist.gov/act_part/tracks/qa/qa.07.guidelines.html

[3] http://terrier.org

Secondly, we test how effective our proposed selective application technique is, by comparing the performance obtained by using the query performance predictor to the baseline that systematically applies QE on the local collection (PL2F+QE). From Table 2, we can see that the best retrieval performance (excluding the upper bounds), highlighted in bold, in each column is obtained by using our proposed method. We also observe that, in some cases, a statistically significant number of queries have been correctly applied with CE. In particular, the $\gamma2$ predictor has led to a significant improvement in MAP. This suggests that our proposed approach is an effective method for selectively applying CE.

Finally, we investigate the importance of the choice of external resources and predictors. From Table 2, we can see that the systematic application of CE can harm the retrieval performance (e.g. Aquaint 2), compared to the results obtained by using PL2F only; while the $AvICTF$ predictor constantly decreases the retrieval performance. This suggests that the choice of an appropriate external resource and predictor before the application of selective CE is very important. In addition, we find that the $\gamma2$ predictor constantly enhances the retrieval performance across 3 different external resources. In fact, the highest MAP score is achieved by using the $\gamma2$ predictor on the Wikipedia collection. Besides, the $\gamma2$ predictor is parameter-free and only relies on the statistics of the collection.

5 Conclusions

We have proposed the use of query performance predictors to selectively apply CE on a per-query basis for document search within an enterprise. The experimental results show that the retrieval performance can be significantly improved when the external resource and the predictor have been appropriately chosen. In particular, the $\gamma2$ predictor is the most efficient, effective and robust predictor for the enterprise document search. In the future, we plan to deploy our proposed method for blog search as collections from the blogosphere contain many spam documents and other noisy vocabulary, meaning that query expansion might benefit from the use of high-quality external resources.

References

1. Diaz F., Metzler D.: Improving the Estimation of Relevance Models using Large External Corpora. In Proceedings of SIGIR 2006.
2. Kwok K. L., Chan M.: Improving two-stage ad-hoc retrieval for short queries. In Proceedings of SIGIR 1998.
3. He B., Ounis I.: Query Performance Prediction. *Information Systems*, (2004).
4. Cronen-Townsend S., Zhou Y., Croft W. B.: Predicting Query Performance. In Proceedings of SIGIR 2002.
5. Bailey P., Craswell N., de Vries A. P., Soboroff I.: Overview of the TREC 2007 Enterprise Track. In Proceedings of TREC 2007.
6. Amati G.: Probabilistic Models for Information Retrieval based on Divergence from Randomness. PhD thesis, UK (2003).
7. Macdonald C., Plachouras V., He B., Lioma C., Ounis I.: University of Glasgow at WebCLEF 2005: Experiments in Per-field Normalisation and Language Specific Stemming. In Proceedings of CLEF 2005.
8. Kirkpatrick S., Gelatt C., Vecchi M.: Optimization by simulated annealing. *Science*, 220(4598) (1983)