On the topical structure of the relevance feedback set

He, J.; Larson, M.; de Rijke, M.

Publication date
2008

Published in
Proceedings: Workshop Information Retrieval 2008, 6.-8. October 2008, University of Würzburg, Germany

On the Topical Structure of the Relevance Feedback Set*

Jiyin He, Martha Larson and Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
{j.he, m.a.larson}@uva.nl, mdr@science.uva.nl

Abstract

We investigate the topical structure of the set of documents used to expand a query in pseudo-relevance feedback (PRF). We propose a coherence score to measure the relative topical diversity/compactness of this document set, which we call the relevance feedback set. The coherence score captures both the topical relatedness of the members of a document set as well as the set's overall clustering structure. We demonstrate the ability of the coherence score to capture topical structure of a document set using tests performed on synthetic data. Then, experiments are performed which show that when queries are divided into two sets, one with high and one with low relevance feedback set coherence, two different levels of retrieval performance are observed. In particular, our results suggests that if a query has a relevance feedback set with a high coherence score, it is apt to benefit from PRF.

1 Introduction

Query expansion (QE) using pseudo relevance feedback (PRF) [5] is an important approach to improving information retrieval. However, expanding queries across the board does not always lead to better overall performance. Many studies have investigated factors impacting QE and developed query predictors for selective expansion [2, 4, 6]. If a predictor is reliable, it will allow the system to pass over those queries that would not benefit from PRF and expand only those that would, leading to an overall improvement in system performance.

One important issue of PRF query expansion is the quality of the set of documents used to expand the original query. We will refer to this set as the relevance feedback set (or the "RF-set"). Measures that capture RF-set quality have high potential to serve as indicators for the success of query expansion. In PRF, the RF-set is built by selecting the top X documents from the list of documents that are retrieved the first time the query is submitted to the system. Intuitively, the RF-set contains documents that are similar to each other due to the fact that they have been selected by the retrieval algorithm as relevant to the query and are therefore biased to the query. We consider scenarios that can possibly result from the initial retrieval run and distinguish two extreme cases. First, the RF-set can contain exclusively relevant documents. In this case, the topical structure of the RF-set is tight and the self-similarity of the RF-set documents can be attributed to topical relatedness to the information need expressed by the query. Second, the RF-set can contain documents with varied topicality, either drawn from aspects of the query topic that are not related to the underlying information need or else topically entirely unrelated to the query. Here, the topical structure is loose and the documents are self-similar only because they are biased by the query and not because they are consistently related to the underlying information need that gave rise to the query. In reality, most RF-sets will fall along the spectrum between the two extremes. However, we conjecture that measures capable of detecting loose topical structure will also reflect RF-set quality and thereby aid in selecting queries for expansion. The underlying assumption is that expansion terms selected from RF-sets containing topically diverse documents would cause problems such as topical drift.

Previous work that has investigated the best way to choose documents for query expansion [14] as well as the impact of the quality of the document set used for PRF on retrieval performance, e.g., [15] and [12]. In particular, Macdonald and Ounis [12], who investigate the task of expert search, build on insights similar to those providing the motivation for our own investigation. In an expert search system, a user query returns a ranked list of people who are experts on a topic. In a candidate-centric model [3], profiles of top candidates in the ranked list are used to expand the query. It is shown in [12] that this approach runs a particular risk of topical drift, since top candidates tend to be experts in multiple topics and using their profiles for expansion introduces words related to extraneous topics into the query. A cohesiveness measure is proposed which is designed to identify candidates with topically diverse profiles and eliminate them from the expansion set. The cohesiveness measure captures the average distance between each document and the overall candidate profile. Our proposed method is motivated by a similar insight, namely, retrieval performance on queries with RF sets having low topical homogeneity may not be helped, and in fact may be hurt, by query expansion. In the work reported here we investigate ad hoc retrieval and look for topical diversity in the set of documents used for pseudo relevance feedback.

Our experiments with coherence scores as measures of the quality of the relevance feedback set used for query expansion is an outgrowth of previous work in which we have demonstrated that coherence scores can be used as a pre-

*This research was supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 017.-001.190, 640.001.501, 640.002.501, 612.066.512, STE-07-012, 612.061.814, 612.061.815.

dictor of query difficulty [10]. Here, we chose documents from the collection to build up a set of documents that reflected the different aspects contained in the query. The coherence of this document set was shown to correlate with retrieval performance measured in terms of Mean Average Precision (MAP). This work suggested that coherence-based scores can capture important information about the topical homogeneity of a document set and motivated us to search for other aspects of retrieval in which the coherence score could be applied. Coherence scores provide a viable alternative to the often more computationally intensive methods proposed in other work on diagnosing query difficulty/ambiguity [7–9, 15].

Loose topical structure in the RF-set of a query can result either when the query is an inadequate representation of the original information need or when the document collection as a whole fails to provide adequate on-topic documents for use in expansion. We are not explicitly concerned with the source of loose RF-set structure, but rather with the impact of PRF on the performance of the query on the collection.

First, we introduce the coherence score, a score that measures the relative tightness of the clustering structure in a certain set of documents measured with respect to a larger collection. We perform experiments that demonstrate that the coherence score reflects the topical structure of a set of documents faithfully enough to distinguish RF-sets with loose topical structures. Second, we design and carry out experiments with synthetic data to confirm our conjecture concerning the relationship between retrieval improvement gained by PRF and the topical structure of the RF-set used for query expansion. Queries whose RF-sets have loose topical structure as reflected by the coherence score are shown to behave differently under query expansion as measured by change in MAP. Our investigation focuses on one of the successful QE approaches from the divergence from randomness (DFR) framework [1].

2 Experiments

2.1 Topical structure of document sets

To capture the topical structure of a document set we use the coherence score. We make the usual assumption applied when using the Vector Space Model in information retrieval that proximity in a space defined by collection vocabulary reflects semantic relatedness. We choose the coherence score because it captures not only the overall semantic similarity of the documents in the RF-set, but also measures the tightness and structure of the sub-topics existing as document clusters within the set. The *query coherence score* we propose to measure the topical structure of the RF-set is inspired by the *expression coherence* score that is used in the genetics literature [13].

The coherence score is defined as that proportion of pairs of documents in a document set that are "coherent" pairs. A document pair is "coherent" if the similarity of the two documents is greater than a certain threshold. More formally defined, given a set of documents $D = \{d\}_{d=1}^{M}$ and a threshold θ, the coherence score (co) is defined as:

$$co(D) = \frac{\sum_{i \neq j \in \{1,\ldots,M\}} \delta(d_i, d_j)}{\frac{1}{2}M(M-1)}. \quad (1)$$

where, for $i \neq j \in \{1, \ldots, M\}$,

$$\delta(d_i, d_j) = \begin{cases} 1 \text{ if } sim(d_i, d_j) \geq \theta, \\ 0 \text{ otherwise,} \end{cases} \quad (2)$$

	scores	1-topic clusters	2-topic clusters	3-topic clusters	random sets
AP89+88	mean	0.7205	0.4773	0.3900	0.0557
	(var)	(0.0351)	(0.0199)	(0.0284)	(0.0002)
robust04	mean	0.6947	0.4490	0.3365	0.0457
	(var)	(0.0463)	(0.0114)	(0.0064)	(0.0002)
trec78	mean	0.7121	0.4619	0.3508	0.0518
	(var)	(0.0450)	(0.0161)	(0.0071)	(0.0002)

Table 1: The mean value and variance of the coherence scores for clusters containing different number of topics, each setting is sampled 100 times with cluster size of 60 documents

where $sim(d_i, d_j)$ can be any similarity metric. In our experiments, we use cosine similarity, but we could have chosen any convenient distance measure without affecting the usefulness of the coherence score for capturing topical structure.

The threshold θ in Eq. 2 is set so that similar document pairs are required to be more similar than random pairs of documents drawn from the background collection. In our experiments, we randomly sample 100 documents for 30 runs. We calculate the similarities between each pair of documents and rank them in descending order, and take the similarity value at the top 5% for each run. Then we take the mean values of those 30 runs as the value of our threshold θ.

The threshold θ is an inherent feature of the background collection rather than a free parameter. For this reason, we do not need a large amount of labeled data to train the system, which is often the case for machine learning based techniques. Moreover, since we measure the topical diversity in an implicit way, we avoid the necessity of making a presumption of the number of clusters contained in the document set, as most of the clustering techniques do.

In order to test the power and reliability of the coherence score in measuring the topical structure of a set of documents, we perform tests using synthetic data. We artificially construct four data sets each containing 60 documents. The first three data sets are generated by randomly sampling 1, 2, and 3 queries from the TREC topics (i.e., TREC queries), and extracting the relevant documents from the TREC qrels (i.e., TREC relevance judgment sets). In this way we control the topical structure of the document set by varying the number of topics it contains. The fourth data set is a random set, sampled directly from the background collection. We calculate the coherence score for each data set. The construction procedure for the four data sets is repeated 100 times. Table 1 shows the average coherence score for these 100 runs on different TREC collections. We use the following TREC collections and topics for experiments: AP89+88 (topics 1–200), Robust04 (topics 301–450, 601–700), TREC78 (topics 351–450).

The results in Table 1 reveal that on average, data sets with 1-topic cluster have significantly higher coherence scores than the data sets with 2- and 3- topic clusters and the random data set. This experiment promises that the coherence score does indeed reflect the topical structure of a set of documents.

2.2 RF-set structure and QE performance

In order to analyze the impact of the topical structure of the RF-set on the performance of query expansion, we de-

signed a second experiment using the same TREC collections and topic sets. For initial retrieval, we use BM25 with default parameter settings. For query expansion, we use the DFR [1] methods. The top ranked X documents are considered as the pseudo relevance feedback set (the RF-set). Terms appearing in the RF-set are weighted by their informativeness and those with highest weights are selected and added into the original query. Within the DFR framework, the informativeness of a term is measured by the divergence of the term distribution in the pseudo RF-set from a random distribution. Specifically, we chose the Bo1 model as the weighting model, which is based on Bose-Einstein statistics. In Bo1, the terms are weighted with the following weighting fuction:

$$w(t) = -\log_2\left(\frac{1}{1+\lambda}\right) - tf_{RF}\log_2\left(\frac{\lambda}{1+\lambda}\right) \quad (3)$$

where

$$\lambda = \frac{tf_c}{TF_c} \quad (4)$$

and tf_{RF} is the frequency of the term in the relevance feedback set, tf_c is the term frequency in the collection and TF_c is the total number of tokens in the collection.

Finally the terms in the query q' after expansion are weighted by:

$$w(t \in q') = tf_q + \frac{w(t)}{w_{max}(t)} \quad (5)$$

where tf_q is the term frequency in the original query.

In this experiment, for each query, we take the top 10 retrieved documents as the RF-set. In the first run, we use the original 10 RF documents to expand the query. Then, we create 10 runs of successive query expansions, which are variations on the first run. In each run, we randomly replace n documents in the RF-set with documents randomly sampled from the collection, where $n = 1, \ldots, 10$. The presence of randomly sampled documents introduces topical diversity, loosening the structure of the RF-set.

Figures 1 and 2 show the results. The plot in Fig. 1 shows the number of random replacements in the RF-set versus the average change in MAP. The plot in Fig. 2 illustrates the number of random replacements in the RF-set versus the coherence score. We can see that as the level of randomization goes up, both the change of MAP and the coherence score go down. The change of MAP is quite subtle at the beginning, and becomes more obvious after certain points, while the coherence score drops steadily and quickly at beginning, but becomes stable when it is very close to randomness. We performed the same experiments using different parameter settings, i.e., the size of the RF-set being 3 and 5, and found that they exhibit the same behavior.

As shown in Figures 1 and 2, the performance of QE is noticeably hurt starting from the point at which a certain level of randomness is present in the RF-set (Figure 1), and the coherence score is a faithful indicator of the level of randomness and thereby the level of topical diversity (Figure 2). Therefore we set a threshold ω on the coherence scores and analyze the behavior of QE for "structured" cases ($co > \omega$) and for "random" cases with loosely structured RF-sets ($co \leq \omega$). We consider the distribution of the coherence scores of the RF-sets for all queries in a test set and take the maximum coherence score of the bottom 5% RF-sets (sorted by coherence score) to be our threshold ω. Interestingly, Table 2 shows that queries in the two sets

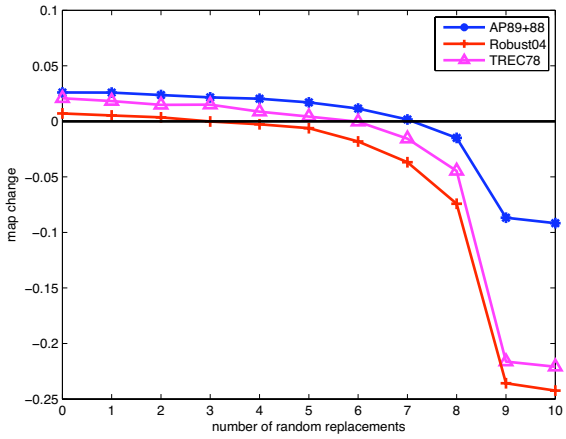

Figure 1: Plot of change in MAP against the number of documents in the top-10 Relevance Feedback set used for query expansion that were replaced by random documents from the collection

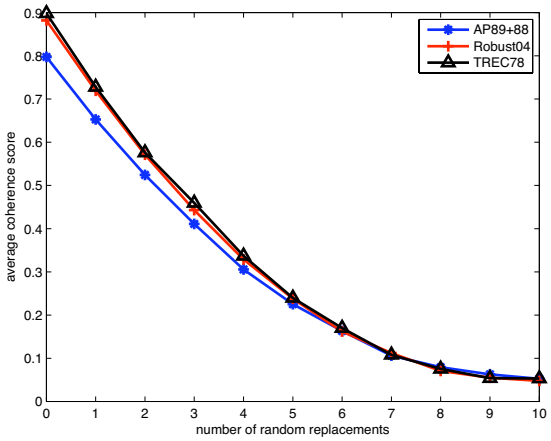

Figure 2: Number of documents in the top-10 Relevance Feedback set replaced by random documents versus the drop of corresponding coherence scores

behave differently with respect to the change of MAP: on average, for queries whose RF-set is random (with $co \leq \omega$), QE helps less (or even negatively) than for queries whose RF-set is structured (i.e., with $co > \omega$). In addition, for each RF-set size (3, 5, 10), we pooled the queries from all collections to attain adequate sample sizes and used a two-tailed Welch's t-test to determine whether the mean ΔMAP observations for the "random" and "structured" queries are equal, and found the difference to be significant (p-value =5.523e-05, 0.0425, 0.005, respectively).

3 Conclusion

In this paper, we analyze the impact of the topical structure in the relevance feedback set on pseudo relevance feedback query expansion with designed experiments. We use the coherence score as the measure of the topical structure of a set of documents. Our initial experiments show that the coherence score can capture the topical diversity/compactness in a set of documents well, and that queries expanded with a highly coherent RF-set would benefit more than queries whose RF-sets have loose structures.

	$X = 3$		$X = 5$		$X = 10$	
	$co \leq \omega_1$	$co > \omega_1$	$co \leq \omega_2$	$co > \omega_2$	$co \leq \omega_3$	$co > \omega_3$
AP89+88	0.0012	0.0266	0.0071	0.0262	0.0002	0.0165
Robust04	0.0227	0.0396	-0.0580	0.0150	-0.0064	0.0184
TREC78	0.0189	0.0297	-0.0089	0.0205	-0.0026	0.0005
Overall	0.0089	0.0334	-0.0119	0.0198	-0.0016	0.0143

Table 2: ΔMAP for "random" and "structured" queries. X=$\{3, 5, 10\}$ is the size of RF-set, with threshold $\omega_1 = 0.33$, $\omega_2 = 0.4$, $\omega_3 = 0.3778$, respectively.

Additional directions for future work include determining whether the method that we propose here yields similar results with different retrieval frameworks and different methods of query expansion. We would also like to investigate whether the document set used for query expansion should be the same set of documents used to calculate the coherence score. It is possible that a larger set would provide a more stable estimate of the coherence score, while at the same time a smaller set is appropriate for expansion.

The work reported here is work in progress. The fact that the set of queries whose Relevance Feedback sets have coherence scores above the threshold have a higher MAP than the set of queries whose coherence scores are below the threshold still needs to be translated into improved MAP for the entire set of queries. The coherence score demonstrates clear potential in supporting the identification of queries that will benefit from query expansion, however, further work is necessary to develop a framework that will incorporate the coherence score into a selective query expansion strategy in such a way that an overall improvement of MAP is achieved.

We see additional applications of coherence scoring beyond pseudo relevance feedback and query difficulty prediction in retrieval settings where topical noise—the phenomenon where a document or set of documents discusses a broad range of issues on top of the topic at hand. An example is provided by He et al. [11], who integrate coherence scoring within a retrieval model based on language modeling to address the task of blog distillation: identifying blogs that show a recurrent interest in a given topic X. They show that significant improvements in retrieval effectiveness can be obtained.

References

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. on Info. Sys.*, 20:357–389, 2002.

[2] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness and selective application of query expansion. In *26th European Conference on Information Retrieval (ECIR 2004)*, pages 127–137, 2004.

[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, New York, NY, USA, 2006. ACM Press.

[4] C. Carpineto, R. D. Mori, G. Amati, and B. Bigi. An information theoretic approach to automatic query expansion. *ACM Trans. on Info. Sys.*, 19(1):1–27, 2001.

[5] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Readings in information retrieval*, pages 339–344, 1979.

[6] S. Cronen-Townsend and W. B. Croft. A language modeling framework for selective query expansion. IR 338, University of Massachusetts, 2004.

[7] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proceedings of the second international conference on Human Language Technology Research*, pages 104–109, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[8] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, New York, NY, USA, 2002. ACM.

[9] B. He and I. Ounis. Query performance prediction. *Inf. Syst.*, 31(7):585–594, 2006.

[10] J. He, M. Larson, and M. de Rijke. Using coherence-based measures to predict query difficulty. In *30th European Conference on Information Retrieval (ECIR 2008)*, pages 689–694, 2008.

[11] J. He, W. Weerkamp, M. Larson, and M. de Rijke. Blogger, stick to your story: modeling topical noise in blogs with coherence measures. In *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 39–46, New York, NY, USA, 2008. ACM.

[12] C. Macdonald and I. Ounis. Expertise drift and query expansion in expert search. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 341–350, New York, NY, USA, 2007. ACM.

[13] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatiorial analysis of promoter elements. *Nat. Genet.*, 29:153–159, 2001.

[14] X. Shen and C. Zhai. Active feedback in ad hoc information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59–66, New York, NY, USA, 2005. ACM.

[15] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 512–519, New York, NY, USA, 2005. ACM.