On the topical structure of the relevance feedback set

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

Published in:

Citation for published version (APA):
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 $D = \{d_i\}_{i=1}^{M}$ and a threshold $\theta$, the coherence score (co) is defined as:

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

(1)

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

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

(2)

Table 1 shows the average coherence scores for clusters containing different number of topics, each setting is sampled 100 times with cluster size of 60 documents.

<table>
<thead>
<tr>
<th>Source</th>
<th>1-topic clusters</th>
<th>2-topic clusters</th>
<th>3-topic clusters</th>
<th>Random clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP89+88 mean</td>
<td>0.7205</td>
<td>0.4773</td>
<td>0.3900</td>
<td>0.0557</td>
</tr>
<tr>
<td>(var)</td>
<td>0.0351</td>
<td>0.0199</td>
<td>0.0284</td>
<td>0.0002</td>
</tr>
<tr>
<td>robust04 mean</td>
<td>0.6947</td>
<td>0.4490</td>
<td>0.3365</td>
<td>0.0457</td>
</tr>
<tr>
<td>(var)</td>
<td>0.0463</td>
<td>0.0114</td>
<td>0.0064</td>
<td>0.0002</td>
</tr>
<tr>
<td>trec78 mean</td>
<td>0.7121</td>
<td>0.4619</td>
<td>0.3508</td>
<td>0.0518</td>
</tr>
<tr>
<td>(var)</td>
<td>0.0450</td>
<td>0.0161</td>
<td>0.0071</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

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 function:

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

(3)

where

$$\lambda = \frac{tf_c}{TF_c}$$

(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:

$$w(t \in q') = \frac{w(t)}{w_{\max}(t)}$$

(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 100 queries 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 $\omega$ 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 $\omega$. 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](image1)

![Figure 2: Number of documents in the top-10 Relevance Feedback set replaced by random documents versus the drop of corresponding coherence scores](image2)

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 $\Delta$MAP 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.
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. 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


<table>
<thead>
<tr>
<th>X = 3</th>
<th>X = 5</th>
<th>X = 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\omega \leq \omega_1)</td>
<td>(\omega &gt; \omega_1)</td>
<td>(\omega \leq \omega_2)</td>
</tr>
<tr>
<td>AP89*+88</td>
<td>0.0012</td>
<td>0.0266</td>
</tr>
<tr>
<td>Robust04</td>
<td>0.0227</td>
<td>0.0396</td>
</tr>
<tr>
<td>TREC78</td>
<td>0.0189</td>
<td>0.0297</td>
</tr>
<tr>
<td>Overall</td>
<td>0.0089</td>
<td>0.0334</td>
</tr>
</tbody>
</table>

Table 2: \(\Delta\)MAP 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.