On Improving Pseudo-Relevance Feedback using Pseudo-Irrelevant Documents

Karthik Raman[1], Raghavendra Udupa[2],
Pushpak Bhattacharya[1], and Abhijit Bhole[2]

[1] Indian Institute of Technology Bombay, Mumbai
{karthikr,pb}@cse.iitb.ac.in
[2] Microsoft Research India, Bangalore
{raghavu,v-abhibh}@microsoft.com

Abstract. Pseudo-Relevance Feedback (PRF) assumes that the top-ranking n documents of the initial retrieval are relevant and extracts expansion terms from them. In this work, we introduce the notion of pseudo-irrelevant documents, i.e. high-scoring documents outside of top n that are highly unlikely to be relevant. We show how pseudo-irrelevant documents can be used to extract better expansion terms from the top-ranking n documents: good expansion terms are those which discriminate the top-ranking n documents from the pseudo-irrelevant documents. Our approach gives substantial improvements in retrieval performance over Model-based Feedback on several test collections.

Key words: Information Retrieval, Pseudo-Relevance Feedback, Query Expansion, Pseudo-Irrelevance, Linear Classifier

1 Introduction

Pseudo-Relevance Feedback (PRF) is a well-studied query expansion technique which assumes that the top ranking $n(> 0)$ documents of the initial retrieval are relevant and extracts expansion terms from them [1]. While several algorithms have been proposed for extracting expansion terms from the top ranking n documents of the initial retrieval, none of them leverage the empirical fact that many of the high scoring documents are actually irrelevant [2], [3]. In this paper, we make use of such documents to improve PRF substantially.

In order to make use of high-scoring irrelevant documents in PRF, we need to solve the following two problems:

1. **IDENTIFY**: Identifying irrelevant documents in the pool of high-scoring documents of the initial retrieval.
2. **EXTRACT**: Extracting good expansion terms from the pseudo-relevant documents with the help of irrelevant documents.

Identifying irrelevant documents among the top-ranking n documents of the initial retrieval can automatically improve retrieval performance. For instance, if we could remove the irrelevant documents from the top 10 results, we would be

automatically improving precision at 10 substantially in most cases. Further, we could use only the relevant among the top-ranking n documents for feedback and improve retrieval performance. Unfortunately, identifying irrelevant documents among the top-ranking n documents is not easy. However, given the set of top-ranking n documents, it is possible to identify high-scoring documents outside of the top n that are highly dissimilar to the top-ranking n documents. Most but not all of these documents are irrelevant in most cases in practice. We call such documents as *Pseudo-Irrelevant* documents. We propose a novel algorithm for identifying pseudo-irrelevant documents from the initial retrieval (Section 2).

Once the pseudo-irrelevant documents have been identified, extracting good expansion terms from the top-ranking n documents boils down to the problem of identifying terms that discriminate the top-ranking n documents from the pseudo-irrelevant documents. To see this, note that good expansion terms should a) increase the scores of the top-ranking n documents and documents similar to them and b) not increase the scores of high-scoring irrelevant documents. By selecting terms that discriminate the top-ranking n documents from the pseudo-irrelevant documents, we achieve both of these objectives. We propose a novel algorithm for extracting discriminative terms (Section 3).

2 Identifying Pseudo-Irrelevant Documents

Let F_R denote the set of top-ranking n documents of the initial retrieval, F_I denote the set of pseudo-irrelevant documents, X denote the the set of high-scoring documents outside of top n and Y denote the set of documents that are similar to any document in F_R. We note that pseudo-irrelevant documents are by definition a) high-scoring documents outside of top n and b) highly dissimilar to any of the top-ranking n documents. Therefore, an intuitive approach to find pseudo-irrelevant documents is to first intersect X with Y and then remove the intersection from the former: $F_I = X - (X \cap Y)$.

The above approach works only if we can extract the set Y from F_R. But we note that it is easy to form the set Y. For each document D in F_R, we only need to find documents in the collection that are similar to it. We form a query Q_D out of D by taking terms that have a collection frequency ≥ 5 and $Idf \geq \log 10$ and retrieve the top-ranking 10 documents for Q_D. These documents are deemed similar to D.

3 Extracting Discriminative Expansion Terms

As mentioned in Section 1, the expansion terms we are interested in are those which discriminate F_R from F_I. Such terms can be found via a classification problem in which each document in F_R is a $+ve$ instance and each document in F_I is a $-ve$ instance and the goal of classification is to learn a discriminant function w that correctly classifies the training instances. Feature vector for each instance is formed as follows: each term that appears in the document forms a

feature provided it is not a stop-word and its collection frequency ≥ 5 and $Idf \geq \log 10$ [3]. The value of a feature is the $tf * idf$ score of the associated term.

We learn a linear discriminant function w from the labeled instances by training a Logistic Regression classifier [4]. The linear discriminant function associates a weight w_i to the term $t_i,\ i = 1, \ldots, N$. The linear discriminant function classifies a vector x as $+ve$ if $w^T x > 0$ and as $-ve$ if $w^T x \leq 0$. Ideally, $w^T x > 0$ for all documents in F_R and $w^T x \leq 0$ for all documents in F_I. Thus, terms $\{t_i : w_i > 0\}$ can be viewed as relevant expansion terms as their presence in a document contributes to the document being classified as $+ve$. Similarly, terms $\{t_i : w_i < 0\}$ can be viewed as non-relevant expansion terms as their presence in a document contributes to the document being classified as $-ve$. We pick the largest weighted $k > 0$ terms as the set of relevant expansion terms.

4 Empirical Investigation

4.1 Experimental Setup

We employed a KL-divergence based retrieval system with two stage Dirichlet smoothing for the initial retrieval [5]. We used model-based feedback (Mixture Model) as a representative PRF technique [3]. We formed the expanded query by interpolating the feedback model with the original query model with the interpolation factor being 0.5. For extracting the expansion terms, we used the top 10 documents fetched by the initial retrieval. We removed stop-words from topics as well as documents and stemmed the remaining words using the Porter stemmer. We trained the discriminant function using LibLinear with the default parameter settings [4].

We used the following test collections in our experiments:

1. CLEF:
   (a) LATimes 94, Topics 1 - 140 (CLEF 2000-2002).
   (b) LATimes 94 + Glasgow Herald 95, Topics 141-200 (CLEF 2003), 251-350 (CLEF 2005-2006).
2. TREC:
   (a) Associated Press 88-89, Topics 51 - 200 (TREC Adhoc Tasks 1, 2, 3).
   (b) Wall Street Journal, Topics 51 - 200 (TREC Adhoc Tasks 1, 2, 3).

4.2 Results

Table 1 compares the performance of the initial retrieval (LM), Mixture Model (MF), and our approach (PI). The performance measures are Mean Average Precision (MAP) and Precision at 5 (P@5). We see that our approach gives improvements in both MAP and P@5 for all the collections over the initial retrieval. Further, improvements in MAP over MF is substantial on the CLEF collections. In the case of TREC collections, the improvement in P@5 over MF is substantial although the MAP is the same. These preliminary results suggest that pseudo-irrelevant documents can improve retrieval performance of PRF.

[3] This throws away most of the noisy terms from the document which would otherwise interfere in the learning of the discriminant function.

Table 1. Retrieval Performance Comparison

Collection	LM		MF		PI	
	MAP	P@5	MAP	P@5	MAP	P@5
CLEF 00-02	0.43	0.49	0.44	0.50	**0.47**	0.50
CLEF 03-06	0.38	0.42	0.41	0.43	**0.43**	**0.47**
AP	0.28	0.47	0.33	0.50	0.33	**0.52**
WSJ	0.27	0.48	0.30	0.52	0.30	**0.53**

4.3 Analysis

While pseudo-irrelevant documents have a positive effect on the performance of expanded retrieval for all the collections we experimented with, the effect is rather varied. CLEF collections seem to benefit from our approach with respect to MAP whereas the TREC collections seem to benefit with respect to P@5. We investigated the causes for this varied effect and found out that TREC collections had more relevant documents per topic on an average than the CLEF collections (25 as against 100). As a consequence, the percentage of true irrelevant documents in the set of pseudo-irrelevant documents is much higher for CLEF topics (90%) than for TREC topics (75%). As a result, our extraction algorithm was losing some valuable expansion terms for some TREC topics as these terms were present in some of the pseudo-irrelevant documents. We are currently investigating ways of improving the percentage of true irrelevant documents amongst pseudo-irrelevant documents.

5 Future Work

In a related experiment, we observed that the mean pseudo-irrelevant distribution is close to the distribution of true irrelevant documents in the feedback document set. This opens up the possibility of leveraging pseudo-irrelevant documents in identifying irrelevant documents among the feedback documents. Such an approach is very likely to benefit PRF beyond what is reported in this work.

References

1. Efthimiadis, E.N. In: Query expansion. Volume 31. Annual Review of Information Science and Technology (1996) 121–187
2. Lavrenko, V., Croft, W.B.: Relevance based language models. In: Proceedings of SIGIR '01. (2001) 120–127
3. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of CIKM '01. (2001) 403–410
4. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. Journal of Machine Learning Research (August 2008)
5. Zhai, C.: Statistical language models for information retrieval a critical review. Found. Trends Inf. Retr. (3)