Predicting Document Effectiveness in Pseudo Relevance Feedback

Mostafa Keikha*, Jangwon Seo‡, W. Bruce Croft‡, Fabio Crestani*
* University of Lugano, Lugano, Switzerland
‡ CIIR, University of Massachusetts Amherst, Amherst, MA
mostafa.keikha@usi.ch, jangwon@cs.umass.edu
croft@cs.umass.edu, fabio.crestani@usi.ch

ABSTRACT

Pseudo relevance feedback (PRF) is one of effective practices in Information Retrieval. In particular, PRF via the relevance model (RM) has been widely used due to the theoretical soundness and effectiveness. In a PRF scenario, an underlying relevance model is inferred by combining language models of the top retrieved documents where the contribution of each document is assumed to be proportional to its score for the initial query. However, it is not clear that selecting the top retrieved documents only by the initial retrieval scores is actually the optimal way for query expansion.

We show that the initial score of a document is not a good indicator of its effectiveness in query expansion. Our experiments show that if we can estimate the true effectiveness of the top retrieved documents, we can obtain almost 50% improvement over RM. Based on this observation, we introduce various document features that can be used to estimate the effectiveness of documents. Our experiments on the TREC Robust collection show that the proposed features make good predictors, and PRF using the effectiveness predictors can achieve statistically significant improvements over RM.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Pseudo Relevance Feedback, Relevance Model

1. INTRODUCTION

Pseudo relevance feedback (PRF) is a well-studied technique for improving the performance of a retrieval system in Information Retrieval. Among a variety of instances of PRF, the relevance model is one of the widely used techniques because not only does it fit well in language modeling frameworks but also it has been shown to be effective for many tasks.

The relevance model (RM) is used to select the most appropriate terms (feedback terms) from the top retrieved documents by the initial search (feedback documents) to expand the original query [4]. Feedback terms for query expansion are selected based on their relevance to the initial query:

$$P(w|\theta_Q) \propto \sum_{D \in \mathbb{D}} P(w|D)P(Q|D) \quad (1)$$

where $\mathbb{D}$ is a set of feedback documents for the initial query Q.

One of the major issues in PRF is selecting the most appropriate feedback documents to be used in the expansion phase. Implementations of RM usually use the top retrieved documents based on their retrieval scores for the initial query. However, since not all the top retrieved documents are relevant, there is a high chance that we use non-relevant documents for expansion.

In order to investigate relations between the initial retrieval score of a document and its effectiveness as a source of query expansion, we conducted the following pilot study on the TREC Robust collection (see Section 3 for more details about the collection). First, we selected the top 100 documents based on the Dirichlet-smoothed language model scores for each topic. We then generated an expanded query based on each retrieved document. Note that the algorithm used for query expansion is RM where only one feedback document exists. This expanded query was submitted as a new query to the same query evaluation system. That is, we obtained 100 different expanded queries and their ranked lists. Average Precision (AP) of each expanded query is presumably the real effectiveness of the corresponding document for query expansion.

The Pearson correlation coefficient between the initial scores of the documents and the AP scores of their corresponding expanded queries tells us how good the initial scores of the documents are as effectiveness estimators. In our pilot study with 250 of the TREC Robust04 topics, the correlation coefficient is -0.32. This indicates a strong negative correlation meaning that the initial scores are poor indicators for selecting documents. Therefore, the initial scores need to be replaced by better estimators for better PRF.

In this work, we address the problem of finding better weighting of feedback documents for RM. Although good

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

document weighting for PRF is critical in RM, there has been relatively little work that addresses this issue. For example, Seo and Croft [7] changed the weighting scheme for combining multiple documents for RM. Other studies have exploited resampling methods to select better documents [1, 5]. However, they did not attempt to find better weights for documents. In this paper, we propose a novel discriminative technique estimating the effectiveness of documents for PRF. More specifically, we introduce various features for documents and a learning approach. By incorporating our estimated effectiveness values into the relevance model formulation as document weights, we demonstrate that our proposed technique can achieve significant improvements over the generic RM.

2. DOCUMENT SELECTION FOR QUERY EXPANSION

2.1 Importance of Document Effectiveness

Before trying to estimate the effectiveness of each document, it is important to know how good the PRF results would be if feedback documents are selected according to the real effectiveness values. To this end, we use the average precision score by each expanded query as the effectiveness value of the document used for generating that expanded query and analyze them.

In order to analyze the effect of using the real effectiveness, we must decide how to use it. For example, RM aggregates the document models of feedback documents by the weighted average and generates a single model that is assumed to represent the relevance probability of each term for the original query. Similarly, we can use the effectiveness of each document as its weight and estimate a single relevance model as follows:

$$P(w|Q) \propto \sum_{D \in \mathbb{D}} P(w|D) \mathit{Effectiveness}(D, Q) \quad (2)$$

where $\mathbb{D}$ is a set of the top 10 documents with the highest effectiveness values among the top 100 documents retrieved by the initial query Q. *Effectiveness(D,Q)* is the effectiveness of document D for query expansion via RM and in the ideal case will be the average precision score of the expanded query by the relevance model estimated from D. Note that in the original RM formulation as shown in (1), the query-likelihood of the original query is used instead of the effectiveness. We call this method $\mathbf{ORACLE}_S$.

Another approach would be to generate multiple ranked lists based on different expanded queries and aggregate the ranked lists to produce a single ranked list afterward. The intuition behind this approach is that there can be some useful terms in some good documents which do not appear in other documents. That is, when generating a single relevance model, these useful terms might be discarded in the final expanded query. To avoid these pitfalls, we try to combine different ranked lists, weighted by the effectiveness of its corresponding document:

$$\mathit{Score}(D', Q) = \sum_{D \in \mathbb{D}} \mathit{Score}(D', Q_D) \mathit{Effectiveness}(D, Q) \quad (3)$$

where Q_D is an expanded query by document D and $\mathit{Score}(D', Q_D)$ is the query evaluation score of document D' for the expanded query. We call this method $\mathbf{ORACLE}_M$.

Table 1: Evaluation results using the true document effectiveness

Model	MAP	P@10	Bpref
LM	0.2794	0.4434	0.2687
RM	0.3085	0.4505	0.2853
$\mathbf{ORACLE}_M$	**0.4582**	**0.6303**	**0.4459**
$\mathbf{ORACLE}_S$	0.4284	0.5909	0.4020

We conducted experiments using the true effectiveness values over the TREC Robust04 queries. Throughout this paper, we fixed the number of feedback terms and the number of feedback documents to 10. These values were selected since they showed the best performance for RM and we did not separately tune them for other methods.

Table 1 shows the performance of the methods using the real effectiveness of documents in comparison to two baselines, i.e., **LM** (the Dirichlet-smoothed language model) and **RM** (the traditional relevance model). As we see, if we know the true effectiveness of documents, we can achieve significant improvements over **RM**. The improvements by the *Oracle_Multiple* are greater and can be up to 50%. These results give us a good motivation for estimating the document effectiveness for query expansion.

2.2 Predicting the Document Effectiveness

In order to detect the most useful documents for query expansion, we extract a set of features that can help us predict the effectiveness of each document. These features are mainly based on the content of each document and its relation with other top retrieved documents. Table 2 summarizes features that we use.

Type1 features (f1-f3) are query independent features that capture some properties of a document and its relation with the collection.

Type2 features (f4-f6) are based on the initial retrieval score of the document and its similarity to the traditional relevance model of the query.

Type3 features (f7-f13) are based on similarity of the document with other top retrieved documents. For example, f7 is the average similarity of the document with other top retrieved documents. Since f7 can be biased toward a few highly similar documents, we define f8 as the number of the documents with higher similarity than a threshold. We set the threshold to be the average similarity between all pairs of the top retrieved documents.

For a term t, we define $\mathit{Com}(t, \mathbb{D})$ which shows how common t is in the top retrieved documents:

$$\mathit{Com}(t, \mathbb{D}) = \frac{\sum_{D \in \mathbb{D}} I(t \in D)}{|\mathbb{D}|} \quad (4)$$

Based on this function, we can define a set of new features for each document. The arithmetic mean, geometric mean and the minimum of the $\mathit{Com}(t, \mathbb{D})$ over all the terms in a document are three of the features that we use (f9-f11).

We assume that an effective document for query expansion has most of its terms in common with other top retrieved documents and so has high arithmetic mean of $\mathit{Com}(t, \mathbb{D})$ values. At the same time, the document would have few terms that are not very popular and accordingly has a low minimum or geometric mean values (those non-popular terms

Table 2: Features used in the regression model for document D

Type	Feature	Description
Type1	f1	Size of D
	f2	Kullback-Leibler (KL) divergence between $p(w\|D)$ and collection model $p(w\|C)$
	f3	Entropy of multinomial distribution $p(w\|D)$
Type2	f4	Querylikelihood score $p(q\|D)$
	f5	KL divergence between $p(w\|D)$ and relevance model $p(w\|\theta_Q)$
	f6	Number of common terms between D and feedback terms selected by $p(w\|\theta_Q)$
Type3	f7	$1/\|\mathbb{D}\| \sum_{D' \in \mathbb{D}} similarity(D, D')$
	f8	$\|\{D' \in \mathbb{D} \| similarity(D', D) > threshold\}\|$
	f9	Arithmetic mean of $Com(t, \mathbb{D})$ over terms t's in D
	f10	Geometric mean of $Com(t, \mathbb{D})$ over terms t's in D
	f11	Minimum of $Com(t, \mathbb{D})$ over terms t's in D
	f12	f9 / f10
	f13	f9 / f11

are the reason that query expansion based on the single best document works better than the traditional RM). Based on this assumption, we define new features by dividing the arithmetic mean of the $Com(t, \mathbb{D})$ by its minimum or geometric mean. We expect these features to have high values for the effective documents.

To learn a document effectiveness prediction model using these features, we consider the following regression problem.

$$\arg\min_{\Phi} \sum_{Q \in T} \sum_{D \in \mathbb{D}_Q} ||\Phi(F(D)) - Effectiveness(D)||^2 \quad (5)$$

where T is a set of training topics, $\mathbb{D}_Q$ is a set of the top 100 retrieved documents for query Q, and F is a mapping from documents to feature spaces. The effectiveness to be learned is the average precision score of each document.

We solve this problem using the Stochastic Gradient Boosting Tree algorithm by Friedman [2]. Then, Φ is an additive model of multiple decision trees. Since some of our features showed non-linear relations with the AP scores in our preliminary experiments, non-linear models based on decision trees such as the Stochastic Gradient Boosting Tree are an appropriate choice.

The learned model on a training set is used to predict the effectiveness of documents on a test set. We select the top 10 documents based on their predicted values. These documents are used for query expansion according to techniques of Equation (2) and (3). Single or multiple expanded queries generated using these methods are called $\mathbf{PREDICT}_S$ and $\mathbf{PREDICT}_M$ respectively.

3. EXPERIMENTS

We evaluated our proposed methods on the TREC Robust04 collection which consists of newswire articles. The collection includes more than 500,000 documents and 250 queries. We used 150 queries for training (topic 301-450) and 100 queries for testing (topic 600-700). We used titles of the topics as queries.

We used the Terrier system[1] to index the collection with the default stemming and stopwords removal. The language modeling approach using the Dirichlet smoothing was used to retrieve and score the top documents for query expansion.

In our experiments, the estimated relevance model is interpolated with the original query model as follows:

$$P(w|\theta'_Q) = (1 - \alpha)P(w|\theta_Q) + \alpha P(w|Q) \quad (6)$$

α is a parameter and in our experiments is set to be 0.5 which has been shown to be effective in previous studies [6]. This interpolation approach is often used to improve retrieval performance.

To estimate language model $P(w|D)$ of the top retrieved documents in the feedback phase, we employ the parsimonious language model and filter out the collection model from the document model using the Expectation Maximization algorithm in order to obtain the most informative terms for each document [3].

For the Stochastic Gradient Boosting Tree (SGBT), we used the `gbm`[2] package implemented in `R`. The bagging option was turned on and the bagging portion parameter was set to 0.5. The number of decision trees and the depth of decision trees were set to 2000 and 7, respectively.

Table 3 shows the performance of the predicted values using the proposed methods. We can see that both $\mathbf{PREDICT}_S$ and $\mathbf{PREDICT}_M$ outperform LM and RM. Interestingly, as opposed to the oracle experiments, $\mathbf{PREDICT}_S$ demonstrates better performance than $\mathbf{PREDICT}_M$. This is mainly due to the sensitivity of $\mathbf{PREDICT}_M$ to the regression er-

Table 3: Evaluation results using the estimated document effectiveness. † and ‡ indicate statistically significant improvements on two baselines, i.e., LM and RM respectively (paired t-test with p-value < 0.05).

Model	MAP		P@10	Bpref	
LM	0.2794		0.4434	0.2687	
RM	0.3085	†	0.4505	0.2853	†
$\mathbf{PREDICT}_M$	0.3296	†‡	0.4586	**0.3130**	†‡
$\mathbf{PREDICT}_S$	**0.3317**	†‡	**0.4626**	0.3095	†‡

Table 4: The top five influential features

Feature	Relative Influence
f4	0.35
f9	0.21
f2	0.13
f1	0.09
f6	0.07

[1] `http://ir.dcs.gla.ac.uk/terrier/`

[2] `http://cran.r-project.org/web/packages/gbm/`

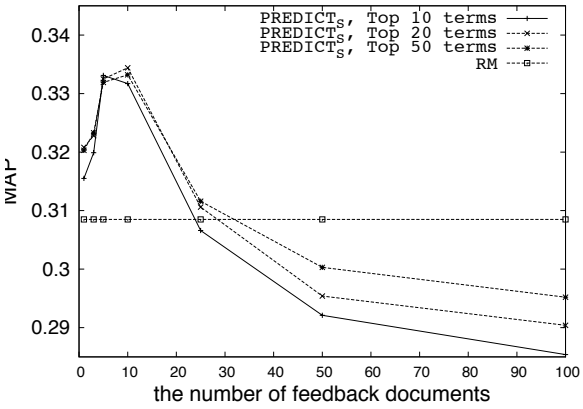

Figure 1: Sensitivity of PREDICT$_S$ to the number of feedback documents.

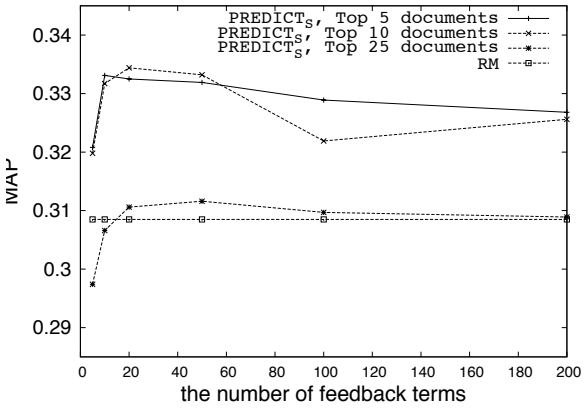

Figure 2: Sensitivity of PREDICT$_S$ to the number of feedback terms.

ror. Since **PREDICT**$_S$ combines multiple document models, it is less likely that a bad document has a high influence over the final term distribution.

We can calculate the relative influence of each feature from the learned SGBT according to [2]. Table 4 shows the top five influential features. Interestingly, while the initial score is not a good indicator by itself, it has the highest influence. Another noteworthy observation is that each of the three types of features (Type1, Type2 and Type3) have a representative in the top three influential features. This shows that all the three types of features need to be considered.

The Pearson correlation coefficient between the predicted values and the true values of effectiveness is +0.53, which indicates a strong correlation. However, there is still considerable room for improvement.

As mentioned before, the reported results are obtained using the top 10 terms extracted from the top 10 documents. In order to investigate the effect of these numbers, we run **PREDICT**$_S$, varying the numbers of feedback documents and feedback terms. Figure 1 and 2 show the effect of these parameters on the performance of the system. The performance appears more sensitive to the number of feedback documents than the number of feedback terms. In Figure 1, when selecting more than 25 documents, regardless of the number of feedback terms, the performance significantly drops. On the other hand, in Figure 2, with a proper choice of the number of feedback documents, e.g., 5 or 10, the effect of the number of feedback terms is minimized. Nevertheless, the best performance is achieved when the number of feedback terms is around 20.

4. CONCLUSION AND FUTURE WORK

In this paper, we investigated the effectiveness of the initial retrieved documents in terms of the performance of pseudo relevance feedback. We showed that the the initial score of a document is not a good indicator for its effectiveness. Based on this observation, we proposed various features that help us predict the effectiveness of documents. The predicted effectiveness values were used as weights of feedback documents for the proposed pseudo relevance feedback techniques. Our experiments showed that the proposed techniques using the predicted effectiveness can achieve statistically significant improvements over the relevance model.

Despite the significant improvements, the results are still far short of the ideal (oracle) situation yet. One direction for future work would be to define better features so that the effectiveness of documents can be more accurately predicted.

Also, in this work, a simple aggregation method is used for some proposed algorithms (e.g., **PREDICT**$_M$) requiring the aggregation of multiple ranked lists. Accordingly, another interesting extension of this work would be leveraging advanced aggregation techniques that have been well-studied in federated search.

5. ACKNOWLEDGEMENT

This work was supported by Swiss National Science Foundation (SNSF) for the "Temporal Analysis of User Generated Data" project as a grant for "Prospective Researcher" to visit CIIR at University of Massachusetts Amherst.

6. REFERENCES

[1] K. Collins-Thompson and J. Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of SIGIR 2007*, pages 303–310, 2007.

[2] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 1999.

[3] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of SIGIR 2004*, pages 178–185, 2004.

[4] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of SIGIR 2001*, pages 120–127, 2001.

[5] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of SIGIR 2008*, pages 235–242, 2008.

[6] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *Proceedings of SIGIR 2010*, pages 579–586, 2010.

[7] J. Seo and W. B. Croft. Geometric representations for multiple documents. In *Proceeding of SIGIR 2010*, pages 251–258, 2010.