Finding Good Feedback Documents

Ben He
Department of Computing Science
University of Glasgow
Glasgow, The United Kingdom
ben@dcs.gla.ac.uk

Iadh Ounis
Department of Computing Science
University of Glasgow
Glasgow, The United Kingdom
ounis@dcs.gla.ac.uk

ABSTRACT

Pseudo-relevance feedback finds useful expansion terms from a set of top-ranked documents. It is often crucial to identify those good feedback documents from which useful expansion terms can be added to the query. In this paper, we propose to detect good feedback documents by classifying all feedback documents using a variety of features such as the distribution of query terms in the feedback document, the similarity between a single feedback document and all top-ranked documents, or the proximity between the expansion terms and the original query terms in the feedback document. By doing this, query expansion is only performed using a selected set of feedback documents, which are predicted to be good among all top-ranked documents. Experimental results on standard TREC test data show that query expansion on the selected feedback documents achieves statistically significant improvements over a strong pseudo-relevance feedback mechanism, which expands the query using all top-ranked documents.

Categories and Subject Descriptors: H.3.3 [Information Storage & Retrieval]: Information Search & Retrieval
General Terms: Performance, Experimentation
Keywords: Relevance feedback, Feedback document classification

1. INTRODUCTION

Relevance feedback is a technique that improves query representation using feedback information. A classical relevance feedback algorithm was proposed by Rocchio in 1971 [9] for the SMART retrieval system [10]. *Pseudo-relevance feedback (PRF)* automatically uses the top-ranked documents in the first-pass retrieval for relevance feedback. A strong assumption behind PRF is that the top-ranked documents are mostly relevant and informative, from which important terms that are closely related to the topic can be extracted. Despite the marked improvement in the retrieval performance over the first-pass retrieval (e.g. [1, 8]), PRF can also fail, leading to a decreased retrieval performance. As suggested by many previous works, the quality of the feedback document set is a key factor that affects query expansion effectiveness (e.g. [1]). A poor feedback document set can be very noisy, so that off-topic expansion terms are added to the query, leading to a degraded retrieval performance.

In the literature of information retrieval (IR), there have been many studies on PRF's effectiveness. For example, a wide range of predictors were proposed to indicate the query performance, which is usually highly correlated with PRF's effectiveness (e.g. [2]). Recently, Cao et al. proposed to refine PRF at the term level [4]. They apply Support Vector Machine (SVM) to select good expansion terms using a list of term features, such as the proximity of the expansion term and the original query terms, or the co-occurrences of the expansion term and the original query terms in the collection. While the expansion term selection approach in [4] has been shown to be effective, we suggest that PRF can also be improved by choosing the right documents for relevance feedback, from which expansion terms are extracted.

In this paper, we argue that the quality of feedback documents is a crucial factor that affects PRF's retrieval performance. We aim to refine PRF at the document level by differentiating between "good" and "bad" feedback documents. We apply standard classification methods to pick up the high-quality feedback documents, or in other words, to remove the low-quality ones. A list of novel feedback document features, including the Entropy of query terms in a feedback document, the similarity between a single feedback document and the whole feedback document set, are applied in our study. In addition, we adapt some of the expansion term features used in [4] to the document level, which are also used in our study.

The main contributions of this paper are as follows. We propose a feedback document filtering mechanism based on standard classification algorithms with various document features. Using our proposed feedback document filtering mechanism, only documents predicted to be of good quality by the classifiers are used for relevance feedback. By extensive experiments on standard TREC test collections, we show that the proposed feedback document filtering mechanism provides statistically significant improvement in the retrieval performance over a PRF baseline, which uses all top-ranked documents for pseudo-relevance feedback.

2. FEEDBACK DOCUMENT FILTERING MECHANISM

In this section, we present our proposed feedback document filtering mechanism. We define the selection of feed-

CIKM'09, November 2–6, 2009, Hong Kong, China.

back documents from the pseudo-relevant set as a binary classification problem, where each candidate feedback document in the pseudo-relevant set is predicted to be either good or bad.

Generally, standard classification methods, such as Naive Bayes classification or Logistic Regression, can be used to yield a prediction confidence value k, from 0 to 1, for each classification instance. In our case, the classification instance is each candidate feedback document in the pseudo-relevant set. Such a k value indicates to which degree the classifier is confident in the prediction outcome.

In this paper, we denote *(+)k* the confidence value of a feedback document predicted to be good, and *(-)k=1-(+)k* that of a feedback document predicted to be bad. The higher k is, the more confidence the classifier has on the prediction outcome. For example, a k value of *(+)0.90* shows that the feedback document is highly likely to be good, and a k value of *(-)0.60* is equivalent to *(+)0.40*, which shows that the feedback document is likely to be bad, but with a less confidence value than (+)0.90 has.

In our study, we propose to use the confidence value as a threshold to develop a feedback document filtering mechanism. Using this mechanism, only a feedback document with a confidence value above the given threshold is used for relevance feedback.

For example, setting the threshold to (+)0.90 implies a hard classifier, which only uses documents that are highly likely to be good for relevance feedback. On the other hand, setting the threshold to (-)0.70 implies a relatively soft classifier, which includes not only the documents in the "good" class, but also the documents in the "bad" class with an absolute confidence value smaller than 0.70.

3. CLASSIFIERS AND DOCUMENT FEATURES

In this paper, we apply Naive Bayes classification (NB) and Logistic Regression (LR) to classify feedback documents. Although any classification method can be applied for this task, we use these two methods for their excellent trade-off between effectiveness and efficiency. In our experiments, we use Weka's implementation of the above two classifiers with default parameter settings [11]. For the NB classifier, the kernel density estimator instead of the normal distribution is empirically applied for a better effectiveness.

We apply a list of features to assist the classification of feedback documents. The applied features take into account the statistics of the expansion terms and feedback documents in different ways, in an attempt to capture the salient characteristics of the good feedback documents. The applied features are described as follows:

- *Relevance Score*. This intuitive feature uses the relevance score produced by the weighting model for each feedback document. The use of the relevance score feature implies that the higher a document is ranked in the first-pass retrieval, the more chance it can be a good feedback document.

- *Entropy*. The *Entropy* feature measures how the query terms are spread over a given feedback document. The PRF process extracts the most informative terms from the feedback documents. In many cases, there might be only a small part of the feedback document that contains relevant information. Thus, off-topic terms are possibly added to the query, resulting in a decrease in the retrieval performance. Therefore, it is necessary to examine the distribution of query terms in the feedback documents to see to which degree the feedback documents are related to the topic. In our work,*Entropy* is defined as follows [7]:

$$Entropy(t,d) = -\sum p_i \cdot \log_2 p_i \tag{1}$$

  where p_i is the probability of observing the query term in the ith subset of the document in tokens. In this paper, we empirically fix the number of subsets in a document to 14. In order to avoid assigning zero probability to subsets where the query term does not appear, we apply Laplace smoothing as follows:

$$p_i = \frac{tf_i + 1}{tf + n} \tag{2}$$

  where tf_i is the term frequency in the ith subset of the document, and tf is the term frequency in the whole document. n is the number of subsets that the document is divided into, which is fixed to 14 as mentioned above. Note that when the query term is uniformly distributed in the document, i.e. p_i is the same across all subsets of the document, the entropy measure is maximised, indicating a dedicated interest of the document in the query topic.

- $sim(d, D)$: The similarity between a given feedback document d and the whole feedback document set D. Ideally, when the feedback document set is of a high quality, the composed feedback documents have a dedicated interest in the query topic. In this case, the most informative terms in different feedback documents should be highly similar, and the feedback document set is highly coherent. In this paper, we define $sim(d, D)$ as the cosine similarity between the $|expT|$ most weighted terms from d and D. In this paper, $|expT|$ is empirically set to 40.

We also apply some of the features that were used in [4]. As the features in [4] were defined at the term level, we transform them to the document level by considering all most weighted expansion terms in a feedback document.

- **dist**: The distance between the expansion terms in the given feedback document and the original query terms. Cao et al. suggested that good expansion terms should appear in close proximity to the query terms, since they are likely to be within the context of the query topic [4]. In their work, the distance between an expansion term and the query terms is defined as the weighted minimal distance from the expansion term to any of the query terms within a given window size. While they define the feature at the term level, we interpret the *dist* feature at the document level by the mean of the weighted minimal distance between an expansion term and the query terms:

$$log_2 \frac{\sum_{t_i \in expT} tf_p \cdot dist(t \in Q, t_i)}{|expT| \cdot \sum_{t_i \in expT} tf_p} \tag{3}$$

where tf_p is the number of co-occurrences of an expansion term t_i with any query term t in the query Q within a given window size. $dist(t \in Q, t_i)$ is the minimum distance between the expansion term t_i and any query term t in the query Q within a given window size. We empirically set the window size to 50, which is suitable for the two collections used, where the feedback documents are usually very long.

- $DF(expT, Q)$: Number of documents in the collection containing each of the expansion terms and all original query terms. This feature measures if the co-occurrence of the expansion terms with the original query terms in the feedback document is by chance or not. Cao et al. defined a similar feature at the term level, which is the number of documents containing a given expansion term and all original query terms.
- *ExpW*: The sum of the Kullback-Leibler divergence (KLD) weights of the expansion terms in the feedback document. This feature implies that a good feedback document should contain informative words, which have high KLD weights. A similar feature was also used in [4] at the term level.

In this paper, we train our classifiers in a supervised manner. In the next section, we introduce our methods for creating the training data for the supervised learning.

4. WHAT IS A GOOD FEEDBACK DOCUMENT?

An initial step of our experiments is to create a ground truth, where each candidate feedback document is labelled as either "good" or "bad". Our classifiers for the feedback documents are then trained based on this ground truth through supervised learning. With respect to this issue, an interesting research question arises: What is a good feedback document?

An intuitive solution is to consider a feedback document to be "good" when it provides an improvement in average precision (AP), compared to the first-pass retrieval. In other words, let AP be the first-pass retrieval performance, and PRFAP(d) be the AP obtained by PRF using document d for feedback, if $\Delta = PRFAP(d) - AP$ is larger than zero, we consider d to be a good feedback document, and a bad one otherwise.

A potential problem of the above naive definition of a good feedback document is that it assumes a linear relation between AP and Δ, which may not be the case in practise. We suggest that the improvement in AP that we expect from relevance feedback is not linearly related to the first-pass AP. If the first-pass AP is too low, the query expansion mechanism will not have a good enough pseudo-relevant set to extract useful expansion terms [2]. On the other hand, if the first-pass AP is too high, there might be only little room for potential improvement. Therefore, the relation between the first-pass AP (AP) and the improvement in AP brought by query expansion (Δ) can be non-linear.

In this paper, we assumes a quadratic function for the expected decrease in AP brought by a bad feedback document:

$$\overline{\Delta} = f(AP) = \alpha(AP - \lambda)^2 - \beta \quad (4)$$

where α, β and λ are again the parameters of the quadratic function. A feedback document is considered to be good when it does not cause a decrease in the retrieval performance that exceeds the expectation.

5. EXPERIMENTS

In the following, we perform experiments with our classifiers for feeback documents. We use Terrier [1] for both indexing and retrieval. We apply the DPH model [6], derived from the Divergence From Randomness (DFR) framework [1], for the first-pass retrieval. Note that DPH is a parameter-free model. All variables in its formula can be directly obtained from the collection statistics. No parameter tuning is required to optimise DPH, and we can rather focus on studying PRF. We mainly report the experimental results obtained by using the 50 top-ranked documents for pseudo-relevant feedback for brevity. We have also experimented with different numbers of candidate feedback documents used for each query, for which the related results are summarised in Section 6.

We experiment on the disk4&5 (minus the Congressional Record on disk4) of the TREC collections, and the large-scale DOTGOV2 TREC Web collection. The disk4&5 collection contains approximately half a million newswire articles from various sources, e.g. the Financial Times, the Los Angeles Times, etc. The 249 ad-hoc queries from the TREC 2004 Robust track are used. Out of the 249 topics, we use the 125 odd-numbered ones for training, and the 124 even-numbered ones for testing. DOTGOV2 is a very large crawl of the .gov domain, which has more than 25 million documents with an uncompressed size of 423 Gigabytes. There are 150 ad-hoc topics, from TREC 2004 - 2006 Terabyte tracks, associated to DOTGOV2. We use the 75 odd-numbered topics for training, and use 50 out of the 75 even-numbered topics for testing, which is the official setting in the TREC 2008 Relevance Feedback track [3]. All documents and queries are stemmed using Porter's stemmer. Standard stopword removal is also applied. We experiment with title-only queries because it is a realistic setting that reflects the concise nature of real users' queries.

We firstly optimise the threshold confidence value k. On the train topics, we test each k value from 0 to 1 with an interval of 0.1, and set the threshold to the k value with the best mean average precision (MAP) on the train topics. The obtained k values are then applied on the test topics to determine which documents are used for relevance feedback.

As the aim of this study is to improve PRF by classifying feedback documents, our baseline is Rocchio's PRF, which performs query expansion over all top-ranked documents. Table 1 provides the related evaluation results. In this table, each cell in the last row contains the obtained MAP, the threshold value obtained on the train topics, and the improvement over PRF in percentage. A star indicates a statistically significant improvement according to the Wilcoxon matched-pairs signed-ranks test at the 0.05 level. For example, "0.2824, (+)0.80, **4.94***" (see the result obtained by the Naive Bayes classifier learnt on disk4&5) indicates a MAP of 0.2824 obtained by relevance feedback. Feedback documents that are predicted to be poor with an absolute confidence value higher than 0.80 are filtered out from the pseudo-relevant set. Such a feedback document filtering mechanism provides a 4.94% statistically significant improvement over the PRF baseline. Moreover, a threshold value of 0.50 indicates that the feedback document filtering mechanism keeps

[1] http://terrier.org

Table 1: IR evaluation results on disk4&5 and DOTGOV2 with their corresponding test topics.

	Naive Bayes		Logistic Regression	
Method	disk4&5	DOTGOV2	disk4&5	DOTGOV2
First-pass	0.2510	0.3139	0.2510	0.3139
PRF Baseline	0.2691	0.3552	0.2691	0.3552
Our Method	0.2824, (+)0.80, **4.94***	0.3636, (0.50), **2.36***	0.2858, (+)0.70, **6.20***	0.3698, (0.50), **4.11***

Table 2: IR evaluation results on disk4&5 and its test topics with different pseudo-relevant set sizes. The threshold setting is the same as those in Table 1, namely (+)0.80 for Naive Bayes classifier and (+)0.70 for Logistic Regression.

$\|D\|$	PRF Baseline	Naive Bayes	Logistic Regression
10	0.2978	0.2982, ≈ 0	0.2962, ≈ 0
20	0.2758	0.2855, **3.52***	0.2824, 2.39
30	0.2550	0.2848, **11.69***	0.2787, **9.29***
50	0.2691	0.2824, **4.94***	0.2858, **6.20***
80	0.2165	0.2820, **30.25***	0.2715, **25.40***
100	0.2039	0.2668, **30.85***	0.2824, **38.50***

Table 3: IR evaluation results on DOTGOV2 and its test topics with different pseudo-relevant set sizes. The threshold setting is the same as those in Table 1, namely 0.50 for both classification methods.

$\|D\|$	PRF Baseline	Naive Bayes	Logistic Regression
10	0.3357	0.3403, 1.37	0.3429, 2.14
20	0.3139	0.3152, ≈ 0	0.3213, 2.36
30	0.2979	0.3124, **4.87***	0.3021, 1.41
50	0.3552	0.3636, **2.36***	0.3698, **4.11***
80	0.3405	0.3462, 1.67	0.3485, 2.35
100	0.3348	0.3429, **2.42***,	0.3417, 2.06

all feedback documents in the "good" class and removes all feedback documents in the "bad" class.

Table 1 shows encouraging results. With appropriate threshold setting, our feedback document filtering mechanism significantly outperforms PRF, which uses all top-ranked documents for relevance feedback. Experiments in this section are conducted with 50 candidate feedback documents per query. In the next section, we vary the number of candidate feedback documents considered for each query to examine the impact the pseudo-relevant set size on the effectiveness of our proposed approach.

6. IMPACT OF THE FEEDBACK DOCUMENT SET SIZE

In this section, we conduct experiments with different numbers of documents in the pseudo-relevant set. Tables 2 and 3 provide the experimental results on disk4&5 and DOTGOV2, respectively. In the tables, each cell in the last two columns contains the obtained MAP value, and the improvement in percentage. A star indicates a statistically significant improvement over the baseline according to the Wilcoxon matched-pairs signed-ranks test at the 0.05 level.

According to the results in Tables 2 and 3, PRF shows a high sensitivity to $|D|$, the size of the pseudo-relevant set. PRF's retrieval performance varies strongly with the change of $|D|$. On the other hand, the retrieval performance of our feedback document filerting mechanism remains stable, particularly on disk4&5. This indicates that our proposed feedback document filtering mechanism is indeed able to pick up the good feedback documents for different sizes of the pseudo-relevant set.

Overall, our feedback document filtering mechanism has been shown to be robust and effective with a varying size of the pseudo feedback set. It provides a retrieval performance that is at least as good as PRF, even if an optimal pseudo-relevant set size is used. This is a very encouraging finding in that the size of the pseudo-relevant set is an important parameter of PRF, which has a direct impact on PRF's retrieval performance [5]. On the other hand, our proposed mechanism is able to achieve an effective retrieval performance without knowing what the actual optimal pseudo-relevant set size is.

7. CONCLUSIONS

In this paper, we have proposed a mechanism for filtering feedback documents that refines Pseudo-relevance feedback (PRF) at the document level. A variety of document features, including the distribution of query terms in the feedback document, the similarity between a single feedback document and all top-ranked documents, or the proximity between the expansion terms and the original query terms in the feedback document, are applied for facilitating the classification of the feedback documents. According to the extensive experimental results, our feedback document filtering mechanism provides effective retrieval performance compared to a strong PRF baseline that uses all top-ranked documents for relevance feedback.

Acknowledgements

This work is funded by SIMAP: Simulation modelling of the MAP kinase pathway. EC project 2006-2009.

8. REFERENCES

[1] G. Amati. *Probabilistic models for information retrieval based on divergence from randomness*. PhD thesis, Department of Computing Science, University of Glasgow, 2003.

[2] G. Amati, C. Carpineto, and G. Romano. Query difficulty, robustness, and selective application of query expansion. In *Proceedings of ECIR 2004*. Sunderland, UK.

[3] C. Buckley, S. E. Robertson. Relevance feedback track overview: TREC 2008. In *Proceedings of TREC 2008*. Gaithersburg, MD.

[4] G. Cao, J. Nie, J. Gao, and S. E. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of SIGIR 2008*. Singapore.

[5] C. Carpineto, G. Romano, and V. Gianini. Improving retrieval feedback with multiple term-ranking function combination. In *ACM Transactions on Information Systems (TOIS)*. 2002.

[6] B. He, C. Macdonald, I. Ounis, J. Peng, and R.L.T. Santos. University of Glasgow at TREC 2008: Experiments in Blog, Enterprise, and Relevance Feedback Tracks with Terrier. In *Proceedings of TREC 2008*. Gaithersburg, MD.

[7] B. He, and I. Ounis. Studying Query Expansion Effectiveness. In *Proceedings of ECIR 2009*. Toulouse, France.

[8] S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In *Proceedings of TREC 4*, 1995. Gaithersburg, MD.

[9] J. Rocchio. Relevance feedback in information retrieval. *The SMART Retrieval SystemâĂŤExperiments in Automatic Document Processing*. Prentice-Hall Englewood Cliffs, 1971.

[10] G. Salton. *The Smart Retrieval System*. Prentice Hall, New Jersey, 1971.

[11] I. Witten and E. Frank *Data Mining: Practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.