On Bias Problem in Relevance Feedback

Qianli Xing[*]
State Key Lab of Intelligent Tech. & Sys.
Tsinghua National Lab for
Information Science & Technology
Dept. of Computer Science & Technology
Tsinghua University, Beijing 100084, China
xingqianli@gmail.com

Yi Zhang, Lanbo Zhang
School of Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064, USA
{yiz,lanbo}@soe.ucsc.edu

ABSTRACT

Relevance feedback is an effective approach to improve retrieval quality over the initial query. Typical relevance feedback methods usually select top-ranked documents for relevance judgments, then query expansion or model updating are carried out based on the feedback documents. However, the number of feedback documents is usually limited due to expensive human labeling. Thus relevant documents in the feedback set are hardly representative of all relevant documents and the feedback set is actually biased. As a result, the performance of relevance feedback will get hurt. In this paper, we first show how and where the bias problem exists through experiments. Then we study how the bias can be reduced by utilizing the unlabeled documents. After analyzing the usefulness of a document to relevance feedback, we propose an approach that extends the feedback set with carefully selected unlabeled documents by heuristics. Our experiment results show that the extended feedback set has less bias than the original feedback set and better performance can be achieved when the extended feedback set is used for relevance feedback.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Relevance feedback, Retrieval models

General Terms: Experimentation, Algorithms.

Keywords: Relevance feedback, bias, novelty, heuristic.

1. INTRODUCTION

In typical relevance feedback methods, feedback documents are usually the top documents retrieved against the initial query. The number of feedback documents is limited due to the expensive user efforts of making relevance judgments. On one hand, the insufficient relevant documents will make the feedback set less representative, namely biased, and thus the performance of relevance feedback will be limited. On the other hand, the redundant information among feedback documents will aggravate the bias of feedback set. For these reasons, the feedback set of top-ranked documents are very likely to be biased because of both insufficient quantity and the similarity among documents.

To recall as many relevant documents as possible, an effective way is to include diverse information in the feedback set. Some previous work studied how to actively select documents which are both relevant and novel for true relevance feedback[5, 6] and pseudo relevance feedback[1]. However in the problem setting that a set of feedback documents is given, which is more likely to happen in reality, these active methods are not available.

In reality, the feedback set is usually a set of top-retrieved documents. Given the feedback set consisting of relevant/irrelevant documents, how to reduce the bias of the feedback set before carrying out relevance feedback is still an open problem. As far as we are concerned, there's little work studying the bias problem of the feedback set. In this paper, we focus on the properties of the feedback set and study how bias is caused. Then an approach that selects unlabeled documents to extend the feedback set is proposed. The documents to select are supposed to be helpful to reduce the bias of feedback set so several heuristics are used for the document selection.

2. BIAS OF FEEDBACK SET

As a feedback set consisting of top-ranked documents tends to be biased, the result of relevance feedback will probably be worse than using a feedback set that contains the same number of relevant documents with better representativeness. In this section, we first show the bias of the feedback set generally exists regardless of retrieval method and feedback method by a control experiment, then we study which feedback sets suffer most from the bias problem.

2.1 Experiment Setting

We conduct our experiments on two standard TREC data sets : the AQUAINT data set used in TREC 2005 HARD track and the RCV1 (Reuters Corpus Volume 1) data set used in TREC 2002 filtering track. There are 50 topics for each data set and only titles are used for retrieval. BM25 and KL-Divergence Language Model are used as our retrieval models. Rocchio algorithm[3] and Collection Mixture Model[7] are the feedback methods for the two retrieval models respectively. We implement the retrieval system and the models using Lemur toolkits with default parameters. Porter stemming and stop words removal are adopted be-

[*]This work was done when the author was a visiting student at UC Santa Cruz

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

Table 1: Performance of relevance feedback. A star (*) and two stars (**) indicate the improvement over Top K is significant according to t-test with p-value threshold of 0.1 and 0.05 respectively.

BM25	No RF	Top 10	Random
AQUAINT	0.183	0.290	0.310*
RCV1	0.269	0.425	0.515**

Language Model	No RF	Top 10	Random
AQUAINT	0.149	0.241	0.243
RCV1	0.286	0.345	0.372**

fore building index. As earlier study has shown that relevant documents are more valuable than irrelevant documents[4], only relevant documents are used by the feedback methods. The relevance judgments are provided by TREC assessors and the recall-based metric MAP (Mean Average Precision) is used for evaluation.

2.2 How the Bias Exists?

In order to show the bias of feedback set using top-ranked documents, we conduct experiments on two feedback sets, which are described as:

Top k: Select the relevant documents from top k retrieved documents as feedback set.

Random: Select the same number of relevant documents as Top from the set of all relevant documents to the query.

$Random$ can be viewed as an approximation of unbiased feedback set because of the random sampling among all relevant documents. The results of relevance feedback are shown in Table 1. As we can see, $Random$ consistently performs better than Top on different data sets and retrieval methods. It indicates that feedback set used in Top is biased. The performance gap between Top and $Random$ can be viewed as the degree of the bias.

2.3 Where the Bias Exists?

As bias generally exists in feedback sets, which feedback sets suffer most from the bias problem? To answer the question, we investigate the relationship between bias and some of the features of the feedback set. The features are listed in Table 2 and related notations are: S_r is the feedback set of relevant documents; S_{nr} is the feedback set of irrelevant documents; S_{res} is the residual set of relevant documents with respect to S_r in the space of all relevant documents. Here $Sim(d_i, d_j)$ is calculated by cosine similarity.

The ground truth of the degree of bias is defined as the performance gap between Top and $Random$ mentioned in previous section. Correlation coefficients are then calculated between the bias and each feature over two data sets with BM25 retrieval model and Rocchio feedback. Experiment results show that $InterSimAvg$ and $InnerSimMin$ are two most correlated features. It indicates that the bias tends to be serious when 1) the similarity between S_r and S_{res} is small; 2) the documents in S_r are similar to each other ($InnerSimMin$ is large).

3. EXTENDING FEEDBACK SET

Since bias exists in feedback set due to lack of diversity, one way to reduce it is to extend the feedback set by adding additional documents. The additional documents are supposed to be both relevant and novel so that novel information could be introduced into the feedback set. In this sections, we study which documents to use for the extension.

Table 2: Features of feedback set S_r

Feature	Calculation
RelDocNum	$\lvert S_r \rvert$
InnerSimAvg	$\frac{\lvert S_r\rvert(\lvert S_r\rvert-1)}{2}\sum_{d_i,d_j\in S_r;i<j} Sim(d_i,d_j)$
InnerSimMin	$\min_{d_i,d_j\in S_r;i<j} Sim(d_i,d_j)$
InnerSimMax	$\max_{d_i,d_j\in S_r;i<j} Sim(d_i,d_j)$
InterSimAvg	$\frac{1}{\lvert S_r\rvert\lvert S_{res}\rvert}\sum_{d_i\in S_r}\sum_{d_j\in S_{res}} Sim(d_i,d_j)$
InterSimMin	$\min_{d_i\in S_r,d_j\in S_{res}} Sim(d_i,d_j)$
InterSimMax	$\max_{d_i\in S_r,d_j\in S_{res}} Sim(d_i,d_j)$

3.1 Usefulness of Unlabeled Documents

As we want to select documents that are useful to reduce the bias of the feedback set and thus boost the performance of relevance feedback, we define the ground-truth of usefulness of document d_i as the MAP improvement achieved by adding d_i to S_r for relevance feedback. In order to find such documents, we extract some features of document in Table 3, then we evaluate how these features are correlated with usefulness. Here the similarity score $Sim(d_i, d_j)$ is calculated in the similarity space of the retrieval method used.

As is shown in Table 4, we notice that over the set of all documents, usefulness is not strongly correlated with any of these features. However, the correlations are much stronger over the set of relevant documents. It implies the importance of relevance. Without the precondition of relevance, it's hard to tell whether an unlabeled document is useful or not by using these features directly.

To identify the relevance, we find that $AvgSimPos$ and $MaxSimPos$ are both good features over the set of all documents. However, it's interesting to notice that over the set of relevant documents, $AvgSimPos$ and $MaxSimPos$ become negatively correlated with usefulness but $Novelty$ becomes a good feature. These facts indicate that when relevance is guaranteed, documents with novelty are more useful to reduce the bias.

3.2 Heuristics for Document Selection

According to the analysis of usefulness of documents, we propose several heuristics to select useful unlabeled documents for the extension of feedback set.

3.2.1 H1:Single-link

Typical centroid-based relevance feedback algorithms prefer to recall documents that are close to the centroid of the feedback documents, while some relevant documents that are only similar to some of the relevant documents will be missed. As $MaxSimPos$ and $MaxSimNeg$ are shown to be well correlated with relevance in Table 4, we utilize these two features to find relevant candidates. The idea is if a document is close to any document in S_r and far away from each relevant in S_{nr}, we believe the document tends to be relevant. Because the relevance is measured by two individual documents in S_r and S_{nr} respectively, we call this heuristic 'Single-link'. On the other hand, novelty is the key factor of usefulness for relevant documents, so we also introduce the novelty component into this heuristic. For a document d,

Table 3: Features of document d

Feature	Description	Calculation
DocLength	the number of words in the document.	$\|d\|$
QuerySim	the similarity with the initial query q	$Sim(q,d)$
AvgSimPos	the average similarity with each document in S_r	$\frac{1}{\|S_r\|}\sum_{d_i \in S_r} Sim(d,d_i)$
AvgSimNeg	the average similarity with each document in S_{nr}	$\frac{1}{\|S_{nr}\|}\sum_{d_i \in S_{nr}} Sim(d,d_i)$
MaxSimPos	the largest similarity with the document in S_r	$\max_{d_i \in S_r} Sim(d,d_i)$
MaxSimNeg	the smallest similarity with the document in S_{nr}	$\max_{d_i \in S_{nr}} Sim(d,d_i)$
Novelty	the portion of new appearing words with respect to S_r	$\left\| W_d \cap \overline{\bigcup_{d_i \in S_r} W_{d_i}} \right\| / \|W_d\|$

Table 4: The correlation between the features and the relevance/usefulness over the set of all/relevant documents

	Relevance	Usefulness	
	all docs	all docs	rel docs
DocLength	0.027	0.011	0.004
QuerySim	0.116	-0.033	-0.119
AvgSimPos	0.299	0.004	-0.254
AvgSimNeg	-0.036	0.066	-0.245
MaxSimPos	0.394	-0.099	-0.197
MaxSimNeg	-0.133	-0.168	-0.142
Novelty	-0.339	-0.182	0.251

the heuristic of usefulness is calculated as:

$$\begin{aligned} H_1(d) = \quad & (MaxSimPos(d) - MaxSimNeg(d)) \\ & \times Novelty(d) \end{aligned} \tag{1}$$

3.2.2 H2:Stretch-out

Compared to Single-link, we also propose a more conservative heuristic called 'Stretch-out'. As we have found that the density, namely the average similarity between any two documents in the feedback set, is an important factor that causes the bias. This heuristic targets to select documents which will dilute the density while maintaining relevance. The selected documents should be as novel as possible but still in the range of S_r. In other words, the radius of the S_r in similarity space should not be increased after the extension. Radius of a feedback set is measured by $InnerSimMin$ (in Table 2). The heuristic is denoted as:

$$H_2(d) = \begin{cases} -density(S_r \cup \{d\}) & , \quad r(S_r \cup \{d\}) \leq r(S_r) \\ 0 & , \quad r(S_r \cup \{d\}) > r(S_r) \end{cases} \tag{2}$$

A non-zero high score for this heuristic means the selected document d is in the acceptable range of relevance and will dilute the density of the feedback set after d is added.

3.2.3 H3:Centroid-locate

As shown before, the distance to the center of relevant documents is a good indicator of relevance and representativeness[2]. Many typical feedback methods are also centroid-based, such as Rocchio algorithm. So for this heuristic, we naturally select documents that are close to the center of the feedback set for extension. The heuristic is simply calculated as:

$$H_3(d) = \frac{1}{|S_r|} \sum_{d_i \in S_r} Sim(d, d_i) \tag{3}$$

Note that in fact this heuristic works in the similar way as the centroid-based feedback methods such as Rocchio, so the relevance feedback on the extended feedback set can also be viewed as a 2-round relevance feedback.

4. EXPERIMENTAL RESULTS

In our experiments, we assume the relevance judgments of top 10 documents are given. Then the proposed heuristics are used to select unlabeled documents to extend the feedback set. In the extended feedback set, documents with explicit relevance judgments are given weight 1.0 while documents selected by heuristics are given a reduced weight λ (set to 0.5) in relevance feedback. Figure 1 shows the performance of relevance feedback (measured by MAP) and the percentage of relevant documents found by the heuristics when using different number of documents for extension.

The first row of Figure 1 shows MAP values, while the second row reflects the proportion of relevant documents in all documents found by the heuristics. According to these results, we have the following observations and analysis:

1)*Relevance is a necessity*: The baseline method (Pseudo) is always among the worst improving the performance of relevance feedback. The simple explanation is that it doesn't consider relevance when selecting documents so consequently it's among the worst in finding relevant documents. H2 also fails to locate relevant documents accurately at the beginning when k is small, so it doesn't help the overall performance as well. We can imagine when many irrelevant documents are used as relevant documents in relevance feedback, the performance drops anyway.

2)*Average similarity is not the best choice*: Usually given a set of relevant documents, we tend to believe that a document that is close to all relevant documents is relevant. H3 is the heuristic that follows this idea. But in fact, this assumption may not hold in different similarity spaces. Figure 1 shows that H3 does well in selecting relevant documents in BM25 model but fails in Language Model (even fails the baseline on RCV1 data set). On the contrary, H1 is always among the best in selecting relevant documents even with the novelty component that is negatively correlated to relevance (shown in Table 4). It indicates that the relevance component of H1 is quite effective. The reason why H1 is consistently good is the Single-link strategy we use, in which a document is believed to be relevant if it's close to any relevant document, instead of all relevant documents. In some similarity space, a document that has a large average similarity doesn't mean it's similar enough to any relevant document so that we can judge it as relevant. Instead, the Single-link assumption is much more robust in different similarity spaces according to this result.

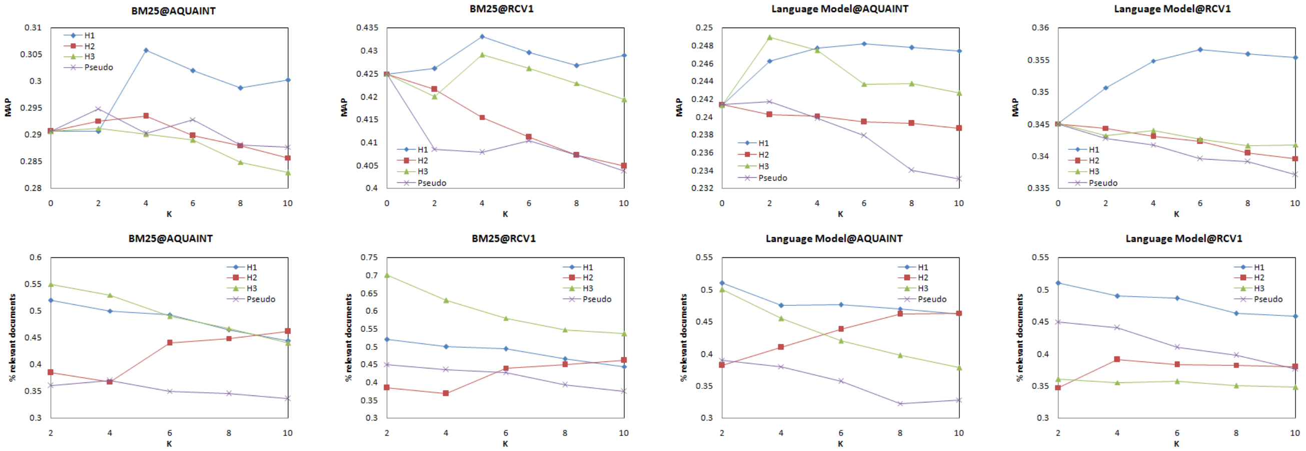

Figure 1: The performance of relevance feedback using the extended feedback set based on top 10 documents. K is the number of unlabeled documents added to the feedback set. $K = 0$ means the original feedback set is used. Pseudo is a baseline in which unlabeled documents are selected by the order of the original ranking.

3)*Novelty plays key role*: With respect to the performance after relevance feedback, H1 consistently works well on every combination of retrieval method and data set. It even has a trend of getting better when the percentage of relevant documents found begins to drop as more unlabeled documents are used. One of the reasons for the good performance is that H1 finds relevant documents well, which we described before. However, we notice that even sometimes with fewer relevant documents, H1 is still be able to outperform the other heuristics on the overall performance of relevance feedback. This fact implies the number of relevant documents is not the only reason that H1 improves the relevance feedback. Another important reason is the novelty. Without consideration of novelty, we see H3 is much less effective than H1 even with more relevant documents. BM25@AQUAINT in Figure 1 shows a very good example of the importance of novelty, in which H1 and H3 both find relevant documents very well. However as a result, H1 performs best for relevance feedback whereas H3 even loses the baseline. The reason why the documents found by H3 are so useless to relevance feedback is the lack of novelty. Redundant relevant documents won't help recall new relevant documents in relevance feedback.

Overall, introducing novelty is the key to reduce the bias of the feedback set. However, we should note that novelty is helpful only when it has the support from relevance. For example, H2 is designed to find documents that increase the diversity of feedback set. But as relevance part fails, the documents selected by H2 hurt the relevance feedback instead of doing any good.

5. CONCLUSION AND FUTURE WORK

This paper studied the bias problem in relevance feedback. We first showed the typical feedback set using top documents is biased by revealing the performance gap compared to a random sampling method. Then we found that feedback sets with high inner-similarity are more likely to be biased. Given a feedback set, we further studied how to reduce the bias by selecting unlabeled documents to extend the feedback set. Several heuristics are proposed for document selection, in which H1 is most effective. Although the improvement on MAP is not dramatic, our work still suggests that selectively using unlabeled documents can reduce the bias of a given feedback set and improve the performance of relevance feedback consequently. For the future work, we would consider developing a learning algorithm for document selection instead of using heuristics. We are also interested in adopting other intent-aware metrics for further evaluation.

6. ACKNOWLEDGMENTS

This research was supported by China Scholarship Council, Natural Science Foundation of China (60736044, 60903107) and Research Fund for the Doctoral Program of Higher Education of China (20090002120005).

7. REFERENCES

[1] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st ACM SIGIR*, SIGIR '08.

[2] F. Raiber and O. Kurland. On identifying representative relevant documents. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10.

[3] J. Rocchio. Relevance feedback in information retrieval. In *The SMART Retrieval System*, pages 313–323. PrenticeHall Inc., 1971.

[4] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1986.

[5] X. Shen and C. Zhai. Active feedback in ad hoc information retrieval. In *Proceedings of the 28th ACM SIGIR*, SIGIR '05.

[6] Z. Xu, R. Akella, and Y. Zhang. Incorporating diversity and density in active learning for relevance feedback. In *Proceedings of the 29th European conference on IR research*, ECIR'07.

[7] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01.