Active Feedback – UIUC TREC-2003 HARD Experiments

Xuehua Shen, ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Abstract

In this paper, we report our experiments on the HARD (High Accuracy Retrieval from Documents) Track in TREC 2003. We focus on active feedback, i.e., how to intelligently propose questions for relevance feedback in order to maximize accuracy improvement in the second run. We proposed and empirically evaluated three different methods, i.e., top-k, gapped top-k, and k-cluster centroid, to extract a fixed number of text units (e.g. passage or document) for feedback. The results show that presenting the top k documents for user feedback is often not as beneficial for learning as presenting more diversified documents.

1 Introduction

For interactive information retrieval such as Web search, a user may need to interact with the search engine many times because of the mismatch of the returned results and the information need. In this case, the user often has to initiate a refined query to do the retrieval several times. But the search engine just uses the current query as the only clue about the user's information need and neglects other apparently useful information such as the user's previous queries in the same search session [10]. In this sense, the search engine responds to a user's query passively.

We believe that a search engine can actively participate in this interactive information retrieval process so that the user's effort can be reduced and retrieval performance can be improved. One interesting way for a search engine to actively participate in this process is to decide what retrieval results the search engine should present to the user during an interactive information retrieval process. Since there are several interactions in this process, when the search engine decides which documents to present to the user, it need consider not only the relevance of the documents to the user's query, but also whether presenting these documents will help the system gain feedback information from the user to improve the next search activity. In this case, the search engine should actively learn which are the best candidate documents to show to the user at any specific moment.

The HARD track of TREC 2003 makes it possible to explore this direction. In the HARD track, the number of times of information retrieval interaction is set to 2. So a search engine would have an opportunity to make use of the first interaction to improve the performance of the second (also final in this case) interaction. In the end of the first interaction, the search engine can propose questions to the user to clarify the user's information need. The search engine can then obtain answers to these questions (e.g., whether a passage is relevant) and some metadata about the information need (e.g., the purpose of the user's search activity), which can presumably be exploited to improve the performance in the second round of retrieval. An interesting and challenging research question is thus how we can best utilize the first interaction to maximize the performance improvement in the second interaction.

We focus our exploration on *active feedback*, i.e., how to intelligently propose passages/documents for user feedback. More specifically, we propose and study three methods, i.e., top-k, gapped top-k, and k-cluster centroid, to extract a fixed number of documents or passages from the initial retrieval results and present them to the user for feedback. Then we use the obtained relevant documents or passages from the feedback process to update our query model and do the second-time retrieval.

The organization of this paper is as follows. In Section 2, we briefly introduce the HARD track. Then we introduce the active feedback in Section 3. In Section 4, we describe how active feedback is used in HARD track. In Section 5, we describe our experiments and result analysis. Section 6 gives our conclusions.

2 HARD Track

The HARD track in TREC2003 is an exploration of how to achieve high accuracy retrieval from documents by leveraging additional information about the searcher and/or the search context, through techniques such as passage retrieval and using very targeted interaction with the searcher [1].

There are two runs to submit in HARD track. In the first (baseline) run, just like the traditional TREC track, given a document database and topics (each topic consists of title, description and narrative), participants use their search engines to do retrieval and submit the retrieval results. At the same time, participants submit a clarification form for each topic, which is used to solicit answers to some questions from the assessors who originally initiated the information need described by the topic. A search engine can freely propose all kinds of questions, e.g., whether some document is relevant to this topic or not. The constraint on the clarification form is that clarification form should be held in a small web page and the assessor will spend no more than 3 minutes filling out the form. We consider this step as the first interaction.

After half a month, participants get the filled out clarification forms with answers from the assessor. At the same time, some metadata about each topic such as relevant terms and searching purpose of the user are also distributed. Search engines can make full use of such information to improve retrieval performance, and submit the second-run retrieval results for the assessor to evaluate. We consider this step as the second interaction.

The HARD track puts search into the context, which allows search engines to actively infer user's information need and improve retrieval performance. We focus on how to intelligently choose passages/documents for user feedback through the clarification forms. There is a limitation on the number of questions to be asked in a clarification form, which is also true in a real interactive retrieval scenario. We want to maximize the amount of feedback information that can be obtained subject to these constraints, in hope of maximizing the retrieval performance in the second run.

3 Active Feedback

Instead of considering information retrieval as only one independent query submission activity, we consider it as an iterative process, in which the user would initiate a query, get retrieval results, and refine the query and submit it again [3]. This provides opportunities for a search engine to actively participate in the retrieval process. For example, a search engine can obtain useful information from the interaction (e.g., inferring relevance of top ranked documents through clickthrough data [4] and/or extracting informative terms from query history [10]) and improve retrieval performance in later interactions of the same search session. Currently, most, if not all, search engines passively respond to user queries and ignore the search context. For example, most search engines only use the information in the current query to generate a ranked list of documents for the user. If the user is not satisfied with the result, (s)he generally has to refine the query and submit it again. Clearly, if the search engine can play a more active role and propose intelligent questions to probe the user's further information need, the user's effort will be reduced, and the final retrieval performance will be improved as well.

Here we consider search as an iterative process. During the retrieval interaction, the documents returned by a search engine have two roles [13]: one is to provide information to the user and the other is to obtain user feedback explicitly or implicitly when the user browses these documents [5]. A search engine can be expected to learn from such explicit or implicit feedback information to improve retrieval performance later in the same search session. In order to maximize the effectiveness of such learning, especially when explicit feedback is possible, the search engine should intelligently choose appropriate questions to ask the user. For example, a question could be whether a document/passage is relevant, or whether a term describes the user's information need. We refer to this problem as *active feedback*. Essentially, active feedback is a problem of applying active learning [9, 11] to ad hoc information retrieval. A similar problem is introduced for learning a text classifier in [7], where a sequential sampling method which chooses most uncertain examples is proposed.

4 Active Relevance Feedback Experiment Design

In HARD track, for each topic, participants can make a clarification form to probe the user with questions. The first decision we face is what kind of questions we want to ask in order to obtain information for active feedback. Perhaps the most natural question to ask is whether some text unit is relevant to the topic or not. The next decision to make is what kind of unit we should present to the user. Individual terms seem to be a good choice, but presenting individual terms has two disadvantages. One is that they are often ambiguous and it is often hard for an assessor to judge precisely whether a particular term is relevant to the topic or not. The other disadvantage is judging individual terms is a boring work for the assessor since the assessor benefits very little from judging individual term relevance. However, presenting documents may not be a good choice either, because a document is generally long and sometimes the assessor may not be able to finish reading a single long document in 3 minutes. Therefore, passages appear to be a good compromise. Accordingly, we make each clarification form contain several passages (6 in this HARD track due to the limited size of the form), so that we can obtain relevance feedback on these passages. An additional benefit of presenting passages is that the assessor can benefit from reading these passages while judging their relevance.

Considering the computation efficiency, we presegment each document into several passages of similar lengths (average is 68.8 words and maximum is 208 words). We build an inverted index for all the passages, do passage retrieval and get a ranked list of passages for each topic (We do not submit this result since it is only used to create clarification forms. Instead we submit document retrieval results in the baseline run).

We proposed and explored three strategies to choose passages for the clarification form. The first one is to choose *top-k* passages from the passage retrieval results, which is the most natural method and is what an existing retrieval system would do. The second one is to choose *gapped top-k* passages from the results. For example, if we set the gap to 3 and k to 6 as we actually did in this HARD track, we will end up choosing the 1st, 4th, 7th, ..., 16th passage from the retrieval results. The third one is to choose *k-cluster centroid* passages from the results. We cluster top N passages of passage retrieval results (we set N to be 100 in this HARD track) into k clusters and choose centroid of k clusters. We use the k-medoid clustering algorithm to do clustering of top N document. And we choose J-Divergence [8] of two passages as the distance function. J-divergence is a divergence metric similar to KL-divergence. But unlike the non-symmetry of KL-divergence, J-divergence is symmetric. The underlying hypothesis for choosing gapped top-k and k-cluster centroid is that the top-k passages may be very similar so that they have redundant information. If search engines instead choose diversified top passages for the clarification form, search engines can also benefit from the active feedback similarly or even more.

When we get the feedback from the assessor, we select relevant passages/documents from the feedback and update the query model. We use mixture model [12] to update the original query used in the baseline run. Then we do the second run retrieval and get the ranked document list.

5 Experiments and Results

We use the Lemur toolkit as our search engine[2] and the KL-Divergence language retrieval model as our retrieval method[6, 12].

In the evaluation stage of HARD track, two judgment files are distributed. One is the hard evaluation judgment file and the other is the soft evaluation judgment file. In the hard evaluation judgment file, a document is relevant if not only the document is relevant to the topic but also the document matches the metadata of the topic. For the soft evaluation judgment file, a document is relevant if the document is relevant to the topic. We pick the soft evaluation judgment file as our judgment file. Since we do not use any metadata information, soft evaluation judgment file is more fair to our experiment evaluation, and is the judgments that we use in all our evaluation.

We submitted five runs. The first one is the baseline run. Then we do passage retrieval to get a ranked passage list. We use gapped top-k and k-cluster centroid strategies to create two clarification forms. After we get answers from the clarification form, we extract relevant passages(we only use relevant passages in this HARD track). We use two methods to update query model. One is to use passage index to update the query model. The other is to assume documents which have relevant passages are relevant and use document index to update the query model. The feedback method is the mixture model as presented in [12]. Then we do document retrieval and obtain results for four runs.

We summarize the mean average precision and pr@20docs in Table 1. From Table 1, we can see retrieval performance using active feedback is significantly better than that of baseline, which indicates that our feedback method is effective. It is also clear that the performance of using relevant *passages* to update the query model is better than that of using relevant *document* to update the query model. The performance of our four active feedback methods is all higher than the median performance of all HARD submissions. Among our four active feedback submissions, the UIHARD4 submission is best for average precision, which uses gapped top-k and updates the query model using passage index. But for pr@20docs, our UIHARD3 submission is the best, which uses k-cluster centroid and also updates the query model using passage index.

Submission		avg prec	pr@20docs
Baseline		0.3077	0.4854
Cluster	doc (UIHARD1)	0.3286	0.5015
	passage (UIHARD3)	0.3465	**0.5219**
	improvement(3 vs. 1)	5.4%	4.0%
Gap	doc (UIHARD2)	0.3321	0.5031
	passage (UIHARD4)	**0.3510**	0.5167
	improvement(4 vs. 2)	5.7%	2.7%

Table 1: Mean Average Precision and pr@20docs of 5 HARD track submissions.The best performance is shown in bold.

To test the hypothesis that diversified documents may be better for feedback than the natural top-k documents, we can use the top-k method as a baseline and compare it with the other two methods. Unfortunately, in our official HARD submissions, we forgot to include the results using the top-k strategy, which makes it impossible to do such analysis with the official runs.

Thus we ran some post-TREC experiments and use the judgments provided by NIST to simulate user feedback. Specifically, we do regular document retrieval and use three active feedback strategies to select k documents for relevance feedback. Then we use relevant documents to update the query; our feedback method can only learn from relevant documents. We then use the updated query to do a second retrieval. The results are shown in Table 2 and Figure 1. We can see that the k-cluster centroid method performs better than the gapped top-k method, which in turn is better than the top-k method in both average precision and precision at 20 documents, though the difference is generally small. Table 2 also shows the total number of relevant documents obtained from the feedback process for all the 48 topics for each method. It is interesting to see that the best performing method – k-cluster centroid – actually has obtained least number of relevant document examples. This suggests that the quality of the examples obtained by k-cluster centroid is probably higher than that of the examples obtained by the other two methods.

Figure 1 shows the precision-recall curve for these three methods. We can see again that at low recall level(0, 0.1, 0.2 and 0.3), performance of gapped top k strategy and k cluster centroid strategy are better than that of top k strategy. In high recall level, performance of top k strategy are slight better.

Active Feedback	avg prec	pr@20docs	#rel
top-k	0.3247	0.4979	146
gapped top-k	0.3278	0.5042	150
k-cluster centroid	**0.3299**	**0.5135**	105

Table 2: Mean Average Precision and pr@20docs of Post-HARD track.The best performance is shown in bold.

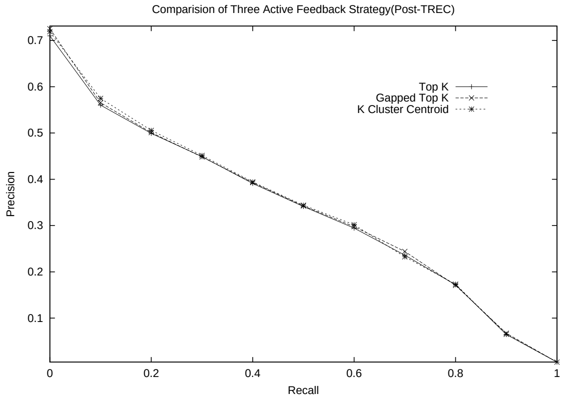

Figure 1: Average Precision at different recall levels.

Active Feedback	avg prec	pr@20docs	#rel
top-k	0.3016	0.4698	146
gapped top-K	0.3114	0.4770	150
k-cluster centroid	**0.3255**	**0.5031**	105

Table 3: Mean Average Precision and pr@20docs of Post-HARD track excluding documents used for active feedback.The best performance is shown in bold.

The results shown in Table 2 are generated based on all the relevance judgments, which means the judgments obtained from the feedback process are also included, which may be biased. Intuitively, the user really does not care where the feedback documents are ranked because the user has already seen these documents. Thus another more meaningful way to evaluate these methods is to exclude any document presented to the user in the feedback stage. This gives us the results shown in Table 3. Here we see again the same order of methods in terms of their relative performance. In fact, the difference between the methods appears to be amplified. Note that this evaluation strategy might be unfair for a method that has obtained more "easy" relevant documents in feedback, since the task becomes harder as more "easy to retrieve" relevant documents are excluded.

These results strongly suggest that just presenting the top-k documents is not optimal for active feedback. Methods that intend to return k diverse documents, such as k-cluster centroid, can be more effective.

The main difference between the experiments that we have just described and our official HARD track submissions is that we use documents instead of passages for judging relevance. Since there is no way for us to obtain equivalent feedback information judged by the official assessors based on passages for the top-k method, we decide to generate results of an approximate top-k baseline.

For the UIHARD1 and UIHARD2 official submissions, we do passage retrieval to get top k passages. Then we use documents which contain at least one of the top k passages for relevance feedback. This top-k results obtained in this way are comparable (not strictly) with UIHARD1 and UIHARD2. The three active feedback methods are compared in Table 4 and Figure 2. This time, we see that the top k method has slightly better performance in both average precision and pr@20docs,and the gapped top-k method obtained the largest number of relevant passages.

Active Feedback	avg prec	pr@20docs	#rel
top-k	**0.3373**	**0.5125**	134
gapped top-k	0.3319	0.5021	155
k-cluster centroid	0.3286	0.5063	121

Table 4: Mean Average Precision and pr@20docs of HARD track using document index to update query model. The best performance is shown in bold.

Active Feedback	avg prec	pr@20docs
top-k	0.3400	0.5177
gapped top-k	**0.3510**	0.5167
k-cluster centroid	0.3465	**0.5219**

Table 5: Mean Average Precision and pr@20docs of HARD track using passage index to update query model. The best performance is shown in bold.

For the UIHARD3 and UIHARD4 official submissions, we use passage retrieval to get top k passages, then we check relevance judgment file and consider a passage as relevant if the document containing this passage is relevant. Then we use passage index to update the query

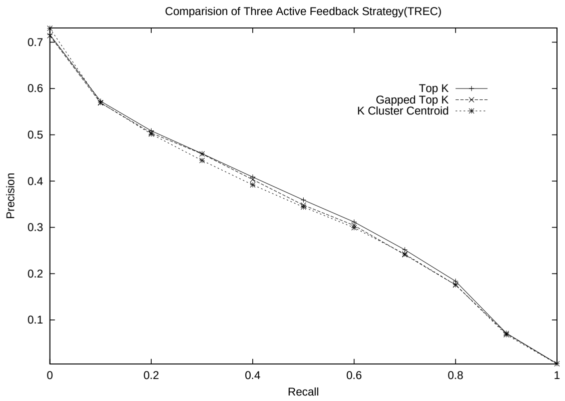

Figure 2: Average Precision at different recall levels.

model, which is used to retrieve the final results. We list our results in Table 5. The top-k results obtained in this way are comparable (again, not strictly) with UIHARD3 and UIHARD4. The three active feedback methods are compared in Table 5. This time, we see that the top K method is, again, inferior to the other two methods.

Since, strictly speaking, the top-k baseline results presented in Table 4 and Table 5 are not really comparable, it is actually hard to make any reliable inference from these two tables.

6 Conclusions

In HARD track of TREC 2003, we focused on the issue of active feedback. We proposed and evaluated three techniques for active relevance feedback, which are the top-k, gapped top-k, and the k-cluster centroid method. We found that the top-k method is not optimal for active feedback, and is worse than both the gapped top-k method and the k-cluster centroid method in a controlled design of experiments. The k-cluster centroid method, which emphasizes returning diversified documents, performs better than both the top-k and gapped top-k methods with fewer examples of relevant documents, suggesting that diversity in the presented documents may be a desirable property.

Clearly, our work represents only a very preliminary exploration of this important topic. We need to do more experiments on other data sets to draw more reliable conclusions. It would be very interesting to develop and test principled models for active feedback.

References

[1] http://ciir.cs.umass.edu/research/hard/.

[2] Lemur Toolkit 2003. http://www.cs.cmu.edu/ lemur.

[3] Tommi Jaakkola and Hava Siegelmann. Active information retrieval. In *Proceedings of Neural Information Systems(NIPS)*, 2001.

[4] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining(KDD)*, 2002.

[5] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: A bibliography.

[6] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of 24th Annual International ACM SIGIR Conference*, 2001.

[7] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of 17th International ACM SIGIR Conference*, 1994.

[8] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[9] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of 18th International Conf. on Machine Learning*, 2001.

[10] Xuehua Shen and ChengXiang Zhai. Exploiting query history for document ranking in interactive information retrieval (poster). In *Proceedings of 26th Annual International ACM SIGIR Conference*, 2003.

[11] Simon Tong and Daphne Koller. Active learning for parameter estimation in bayesian networks. In *Proceedings of Neural Information Systems(NIPS)*, 2000.

[12] Chengxiang Zhai and John Lafferty. Model-based feedback in kl divergence retrieval model. In *Proceedings of the 10th International Conference on Information and Knowledge Management(CIKM)*, 2001.

[13] Yi Zhang, Wei Xu, and James P. Callan. Exploration and exploitation in adaptive filtering based on bayesian active learning. In *Proceedings of the 20th International Conference on Machine Learning(ICML)*, 2003.