Active Feedback in Ad Hoc Information Retrieval

Xuehua Shen
Department of Computer Science
University of Illinois at Urbana-Champaign
xshen@cs.uiuc.edu

ChengXiang Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
czhai@cs.uiuc.edu

ABSTRACT

Information retrieval is, in general, an iterative search process, in which the user often has several interactions with a retrieval system for an information need. The retrieval system can *actively* probe a user with questions to clarify the information need instead of just passively responding to user queries. A basic question is thus how a retrieval system should propose questions to the user so that it can obtain maximum benefits from the feedback on these questions. In this paper, we study how a retrieval system can perform *active* feedback, i.e., how to choose documents for relevance feedback so that the system can learn most from the feedback information. We present a general framework for such an active feedback problem, and derive several practical algorithms as special cases. Empirical evaluation of these algorithms shows that the performance of traditional relevance feedback (presenting the top K documents) is consistently worse than that of presenting documents with more diversity. With a diversity-based selection algorithm, we obtain fewer relevant documents, however, these fewer documents have more learning benefits.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance Feedback, Search Process, Clustering

General Terms

Algorithms

Keywords

Active Feedback, Ad Hoc Information Retrieval

1. INTRODUCTION

In ad hoc information retrieval, a user often needs to interact with the retrieval system several times to obtain satisfactory results for one information need, which provides opportunities for the retrieval system to actively participate in this iterative retrieval process. Most traditional retrieval systems just passively respond to user queries and put the responsibility of refining/improving the search solely on the user. But there has been evidence showing that a retrieval system can play an active role in this process, e.g., obtaining user feedback explicitly or implicitly when the user browses these documents, and exploiting such information to improve the performance in the next round of search [6, 8]. Ideally, a retrieval system should collaborate with the user in the whole interactive search period to improve the accuracy and reduce the number of interactions.

When explicit feedback is possible, a natural way for the retrieval system to actively participate in the retrieval process is to clarify the user's information need by probing the user with well-designed questions. A question could be whether a document or passage is relevant, or whether a term describes the user's information need.

In this scenario, a basic question is how the retrieval system should intelligently propose the questions so that it can learn most from the user's answers to these questions. In this paper, we study how a retrieval system can perform *active feedback*, i.e., how to choose documents for relevance feedback so that the system can learn most from the feedback information.

Relevance feedback is known to be effective for improving retrieval performance [16, 18, 4]. Previous work on relevance feedback focuses on query updating techniques such as query term reweighting and query expansion. The issue of choosing documents for relevance feedback has not been well addressed. Traditionally, relevance feedback methods just choose the top ranked documents for feedback, which is not necessarily the best strategy from the learning perspective. For example, if the top two documents have identical contents, the learning benefits of these two documents will be nearly equal to that of any one of them. Thus a very interesting research question is how to select appropriate documents for user judgment to maximize the learning benefits, which is the focus of the study in this paper.

Active feedback is essentially an application of active learning in ad hoc information retrieval. Active learning has been extensively studied in machine learning [17, 22, 3]. It has been applied to text categorization in several previous studies [12, 14, 23], and recently to adaptive information filtering [29]. But there has been little work on applying it to ad hoc retrieval, partly because there are two special challenges in applying active learning to ad hoc retrieval. First, in ad hoc retrieval, we do not have any training examples available to guide the retrieval system for actively selecting the documents for feedback; the query is the only information that can be exploited. Second, it is unclear how we can define an objective function that optimizes *ranking* performance rather than classification accuracy. An interesting recent work on applying active learning to ad hoc retrieval is [5], where a user is assumed to iteratively choose clusters, and the active learning task for the system is to design good clusters, a different task from active feedback. The TREC HARD Track [1] has stimulated some recent work along the line of active feedback including [15, 20].

In this paper, we frame the problem of active relevance feedback as a statistical decision problem, and examine several special cases in refining the framework. We derive several practical algorithms for active feedback, including the Top K, Gapped Top K and K Cluster Centroid algorithm. We empirically evaluate these three algorithms using the TREC2003 HARD data , AP88-89 and AP90. The results show that the performance of the Top K algorithm (i.e., the traditional way of relevance feedback) is consistently worse than that of Gapped Top K algorithm and K Cluster Centroid algorithm which present documents with more diversity. In general, with a diversity-based selection algorithm, we obtain fewer relevant documents, but these fewer documents have more learning benefits.

The remaining sections are organized as follows. In Section 2, we present the active feedback framework and derive several practical algorithms. In Section 3, we describe our evaluation methods and three algorithms we tested. We discuss the experiment results in Section 4 and conclude our work in Section 5.

2. ACTIVE FEEDBACK FRAMEWORK

The problem of active feedback is essentially a decision problem in which we choose the best subset of documents for relevance judgment by the user. To formalize this problem, we follow the risk minimization framework for retrieval [9] and treat it as the following optimization problem:

$$D^* = \arg\min_D \int_\Theta L(D,\mathcal{U},\theta)p(\theta|\mathcal{U},q,\mathcal{C})\,\mathrm{d}\theta$$

where $D = \{d_1, ..., d_k\}$ is a subset of the document collection $\mathcal{C}$, q is a query, $\mathcal{U}$ is a user variable, θ is the set of parameters of the query language model and document language models. $p(\theta|\mathcal{U},q,\mathcal{C})$ is the posterior probability distribution of all the parameters, and $L(D,\mathcal{U},\theta)$ is a loss function reflecting how much we can expect to learn by requesting relevance judgments on D from user $\mathcal{U}$. In general, the loss function may also depend on other factors such as any relevance judgments available from previous iterations of retrieval, but here we ignore those factors for the convenience of presentation.

Without refining the language models $p(\theta|\mathcal{U},q,\mathcal{C})$, which is not the focus of this paper, we study how to define the loss function for active feedback. Clearly, the actual value of a set of documents D for learning depends on not only D but also the judgments the user would make. Let $\mathcal{J} = \{0,...,m\}$ be the set of all possible relevance levels that a user may assign to each presented document (0 for "completely non-relevant"). For example, for binary judgments, $\mathcal{J} = \{0,1\}$. The loss function can now be written as

$$L(D,\mathcal{U},\theta) = \sum_{\vec{j}\in\mathcal{J}^k} l(D,\vec{j},\theta)p(\vec{j}|D,\theta,\mathcal{U})$$

where $\vec{j} = (j_1, ..., j_k)$ and j_i is a possible judgment for document d_i in D; $p(\vec{j}|D,\theta,\mathcal{U})$ is the probability that the user $\mathcal{U}$ would assign judgments $\vec{j}$ to all the documents in D; and $l(D,\vec{j},\theta)$ is a loss function that indicates how much we can learn from the judgments $\vec{j}$ on D. In other words, $l(.)$ tells us how good $(D,\vec{j})$ is as a set of labeled examples for learning.

Now assuming that the user would judge each document *independently*, we have

$$L(D,\mathcal{U},\theta) = \sum_{\vec{j}\in\mathcal{J}^k} l(D,\vec{j},\theta)\prod_{i=1}^{k} p(j_i|d_i,\theta,\mathcal{U})$$

Note that this assumption is reasonable if a user explicitly judges a document, but it is unlikely to hold when we infer a user's judgments based on, say, clickthrough data [6], as obviously a user would not open a redundant (but relevant) document.

Thus our general framework for active feedback is the following decision rule:

$$D^* = \arg\min_D \int_\Theta [\sum_{\vec{j}\in\mathcal{J}^k} l(D,\vec{j},\theta)\prod_{i=1}^{k} p(j_i|d_i,\theta,\mathcal{U})]p(\theta|\mathcal{U},q,\mathcal{C})\,\mathrm{d}\theta$$

In the remaining part of the section, we discuss some interesting special cases. We will assume that the relevance judgments are all binary, though most derivations can be easily generalized to multi-level judgments.

2.1 Independent Loss

Let us first simplify the loss function by assuming that the value of each judged example for learning is *independent* of each other. The total value of a set of examples $(D,\vec{j})$ can thus be written as the sum of the value of each individual example, i.e.,

$$l(D,\vec{j},\theta) = \sum_{i=1}^{k} l(d_i,j_i,\theta)$$

where $l(d_i,j_i,\theta)$ is the loss for a single judged document (d_i,j_i).

After some algebraic manipulation, we have

$$L(D,\mathcal{U},\theta) = \sum_{i=1}^{k}\sum_{j_i} l(d_i,j_i,\theta)p(j_i|d_i,\theta,\mathcal{U})$$

And the active feedback decision rule is

$$\begin{aligned} D^* &= \arg\min_D \int_\Theta \sum_{i=1}^{k}\sum_{j_i} l(d_i,j_i,\theta)p(j_i|d_i,\theta,\mathcal{U})p(\theta|\mathcal{U},q,\mathcal{C})\,\mathrm{d}\theta \\ &= \arg\min_D \sum_{i=1}^{k}\sum_{j_i}\int_\Theta l(d_i,j_i,\theta)p(j_i|d_i,\theta,\mathcal{U})p(\theta|\mathcal{U},q,\mathcal{C})\,\mathrm{d}\theta \end{aligned}$$

The optimal set D can thus be obtained by ranking all the documents according to the following risk function and taking the k documents with the least risk:

$$r(d_i) = \sum_{j_i}\int_\Theta l(d_i,j_i,\theta)p(j_i|d_i,\theta,\mathcal{U})p(\theta|\mathcal{U},q,\mathcal{C})\,\mathrm{d}\theta$$

which can be interpreted as the expected value of d_i for learning over all possible judgments.

We now examine the assumptions underlying two simple methods for defining $r(d_i)$ – "Top K" and "Uncertainty Sampling".

2.1.1 Top K

Let us assume that the loss of any relevant example (document) and that of any non-relevant example (document) are both constants. We further assume that the former is smaller than the latter, which is to say that a relevant example is more useful for learning than a non-relevant one. Formally, $\forall d_i \in \mathcal{C}$, we have $l(d_i,1,\theta) = C_1$, $l(d_i,0,\theta) = C_0$, and $C_1 < C_0$.

The risk function now becomes

$$r(d_i) = C_0 + (C_1 - C_0)\int_\Theta p(j_i = 1|d_i,\theta,\mathcal{U})p(\theta|\mathcal{U},q,\mathcal{C})\,\mathrm{d}\theta$$

Since $C_1 - C_0 < 0$, clearly the optimal set D^* is precisely the k documents with the highest probabilities of being judged as relevant (i.e., with the highest expected values of $p(j_i = 1|d_i,\theta,\mathcal{U})$). That is, we should simply rank all the documents in $\mathcal{C}$ according to the estimated relevance status of each document and select the top k documents that are most likely relevant for feedback.

We have thus obtained the "Top K" method as a special case under three assumptions [1]: (1) independent loss function; (2) constant loss for any relevant (non-relevant) document; and (3) a relevant document has a smaller loss than a non-relevant one. The results are not really surprising because assumption 2 basically says that all relevant (non-relevant) examples are equally good for learning. However this analysis suggests that we may expect other methods to perform better than Top K if the underlying feedback algorithm does *not* satisfy all these three assumptions, e.g., independent loss function.

2.1.2 Uncertainty Sampling

In [12, 11], a similar document selection problem is studied, though a set of documents are selected for labeling to train a text classifier instead of a ranking function. Authors propose to select the most uncertain documents for labeling. In [28], a similar idea, i.e., selecting most uncertain objects, is used to guide the hidden annotation for content-based image information retrieval. Using our general active feedback framework, we can derive the uncertainty sampling method by assuming the following loss function:

$$\begin{aligned} l(d_i, 1, \theta) &= \log p(R = 1|d_i, \theta) \quad \forall d_i \in C \\ l(d_i, 0, \theta) &= \log p(R = 0|d_i, \theta) \quad \forall d_i \in C \end{aligned}$$

where $R \in \{0, 1\}$ is a binary relevance variable with 1 indicating "relevant". This loss function essentially says that a relevant example is more useful if the predicted probability of relevance is smaller according to our current model, and similarly, a non-relevant example is more useful if the predicted probability of relevance is larger. In other words, an example is more useful if our prediction has less confidence.

With such a loss function, and assuming $p(R|d_i, \theta)$ is an approximation of $p(j_i|d_i, \theta, \mathcal{U})$, the risk function becomes

$$r(d_i) = -\int_{\Theta} H(R|d_i, \theta) p(\theta|\mathcal{U}, q, \mathcal{C}) d\theta$$

where H is the entropy function. This means that, in order to obtain D, we should rank documents in the descending order of the expected entropy of the corresponding relevance variable R. That is, we would pick documents with the highest uncertainty.

We have thus obtained the "Uncertainty Sampling" method as a special case under two assumptions: (1) independent loss function; (2) an example is more useful for learning if our prediction of relevance is more uncertain. This method relies on explicitly predicting the probability of relevance, which is often not feasible in ad hoc retrieval.

2.2 Dependent Loss

Our assumption about an independent loss on each example is not realistic. For instance, if two examples are completely identical, their total value is clearly less than the sum of their individual values, and is probably close to the value of one of the examples. Thus we need to model the interactions between documents with a dependent loss function. Unfortunately, the exact form of such a loss function highly depends on the specific feedback algorithms. Nevertheless, intuitively, given a fixed size of D, increasing the representativeness of documents in D appears to be always desirable. At the same time, we can also reasonably assume that relevant examples are more useful than non-relevant examples. Thus one possible way to refine a dependent loss function is to associate the value of D for learning with the relevance status and diversity of D. That is, we write our loss function as

$$L(D, \mathcal{U}, \theta) \approx -\sum_{i=1}^{k} p(j_i = 1|d_i, \theta, \mathcal{U}) - \lambda \Delta(D, \theta)$$

where $\Delta(D, \theta)$ is a function that measures the diversity of documents in D and λ is a parameter indicating the tradeoff between the relevance and diversity.

According to this loss function, the active feedback decision rule is

$$\begin{aligned} D^* = \arg\min_{D} & - \int_{\Theta} \sum_{i=1}^{k} p(j_i = 1|d_i, \theta, \mathcal{U}) p(\theta|\mathcal{U}, q, \mathcal{C}) \, \mathrm{d}\theta \\ & - \lambda \int_{\Theta} \Delta(D, \theta) p(\theta|\mathcal{U}, q, \mathcal{C}) \, \mathrm{d}\theta \qquad (1) \end{aligned}$$

That is, we need to select D to simultaneously optimize both relevance (the first term) and diversity (the second term). A possible greedy algorithm is to first optimize the relevance term by selecting top N ($N > K$) documents according to relevance-based ranking, and then to further select K most diverse documents from the N documents. We now discuss several simple methods along this line.

2.2.1 Gapped Top K

Suppose we let $N = (G+1)K$, where G is a small positive integer. To capture the diversity, we partition the N documents into K clusters based solely on the relevance scores so that our first cluster would have the first $G + 1$ documents and the second one have the next $G + 1$ documents, and so on so forth. From each cluster, we then select a document with the highest relevance score to form our feedback document set D. We refer to this method as "Gapped Top K" since it corresponds to selecting the top K documents with a gap of G documents in between any two documents. An interesting property of this method is that when $G = 0$, it is essentially the regular Top K method.

2.2.2 Maximal Marginal Relevance (MMR)

Maximal Marginal Relevance (MMR) ranking is a greedy algorithm for ranking documents based on relevance and at the same time avoiding redundancy [2, 26]. Specifically, we iteratively select a document which optimizes the following MMR function:

$$r(d|D) = \alpha s(d) + (1 - \alpha) \max_{d' \in D} sim(d, d')$$

where $s(d)$ is a relevance scoring function, $sim(d, d')$ is a similarity function, and α is a parameter for trading off between relevance and redundancy.

This method can also be regarded as performing an *implicit* clustering and then selecting a document from a cluster with the highest relevance value. The first document selected will be the top ranked one based on relevance. Since the next document to be selected must be far from this selected first document, we can interpret the first document as implicitly defining a cluster with the first document being the centroid, and none of the other documents in the cluster will be selected since they are all too close to the selected first document. As we select the second document, we again have another cluster which further excludes some documents from being selected. However, it is unclear what the clustering boundary is exactly as it is affected by not just the similarity function, but also the relevance scores of documents and the parameter α. The MMR method can also cover the Top K method as a special case when $\alpha = 1$.

---
[1] Top K as an active feedback method was first discussed in [10].

2.2.3 Cluster Centroid

A more direct method to maximize diversity is to perform explicit clustering. Specifically, we can first select the top N documents according to the relevance scores. Then we partition these N documents into K clusters and construct D by selecting one representative document from each cluster.

To optimize the relevance term in equation 1, we restrict N to a relatively small number. In this way, we ensure that each of the N documents has a reasonably high probability of relevance. The diversity is ensured through clustering and choosing only one document from each cluster. There are several different ways to choose a representative document from each cluster. One is to choose the centroid document, which maximizes the average similarity between the chosen document and other documents in the cluster. Another choice is to choose the document with the highest relevance score.

3. EXPERIMENT METHODOLOGY

3.1 Data Set

We use two data sets for experiments. One is the Associated Press (AP) news data on TREC disks 1, 2, and 3. The other is the TREC2003 HARD (High Accuracy Retrieval from Documents) track data set [1]. TREC2003 HARD track puts search into context, which allows a retrieval system to actively infer a user's information need and improve retrieval performance [20]. Our experiment process simulates two runs of the HARD track experiment setup [1]. For the HARD track data set, we use 48 topics that have relevance judgments. For the AP data set, we use 92 topics from topics 1-50 and 101-150, which have relevance judgments on both the AP88-89 and AP90 data set. We use the title of each topic as the query.

3.2 Experiment Setup

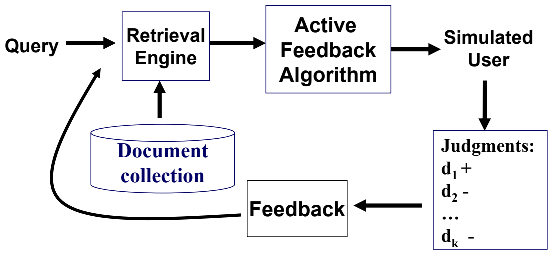

Figure 1: Evaluation Procedure

We use the Lemur toolkit as our retrieval system [24] and the KL-Divergence language retrieval model as our retrieval model [9, 25]. K is fixed to 6 in most experiments, and all parameters are set to default values [24] unless otherwise stated. Our baseline run is regular retrieval without any feedback. It allows us to test whether we can improve performance by performing feedback. From the baseline retrieval results, we use different active relevance feedback algorithms to select a set of documents for relevance feedback. Using the known relevance judgments available from these TREC data to simulate a user's judgments, we obtain relevance judgments on the selected documents. These judgments are then used to perform feedback using the mixture model approach implemented in Lemur [27]. This method only uses relevant documents for query model updating, which can be a limitation of our study. The retrieval results in the second run using different active feedback algorithms are compared for evaluation. This experiment procedure is illustrated in Figure 1.

3.3 Algorithm Description

As a first step of studying active feedback, we evaluate three representative active feedback algorithms discussed in Section 2. The first one is Top K, which chooses top K documents from the baseline run retrieval, and is also what existing retrieval systems would normally do. The second one is Gapped Top K , which is to choose gapped top K documents from the baseline run results. For example, if we set the gap to 3 and K to 6, we will end up choosing the 1st, 5th, 9th, ..., 21st documents from the retrieval results. The third one is K cluster centroid, which represents the most direct way of modeling diversity. We use the K-Medoid clustering algorithm [7] to cluster the top N documents. And we use J-Divergence [13] of two documents 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 formula of J-Divergence is as follows.

$$J(d_i||d_j) = \sum_w p(w|\theta_i) \log \frac{p(w|\theta_i)}{p(w|\theta_j)} + \sum_w p(w|\theta_j) \log \frac{p(w|\theta_j)}{p(w|\theta_i)}$$

Evaluation of these methods allows us to examine whether presenting a diverse set of documents for feedback leads to more effective feedback than presenting the top k documents with the highest relevance values.

3.4 Evaluation Method

To measure the performance of a ranking method, we use two standard ad hoc retrieval measures: (1) Mean Average Precision (MAP): This is the commonly used non-interpolated average precision and serves as a good measure of the overall ranking accuracy since it is sensitive to the rank of every relevant document. (2) Precision at 10 documents (pr@10): This measure does not average well and only gives us the precision at one single cutoff point. But it reflects the utility perceived by a user who may only read up to top 10 documents. In all cases, the reported figure is the average over all the topics.

Since the task of active feedback involves identifying a certain number of relevant documents by the user, an interesting question is whether we should include such relevant documents when computing the retrieval precision of an active feedback algorithm. While this is also a problem for relevance feedback evaluation, it is especially a challenge for evaluating active feedback algorithms because the set of relevant documents used for feedback can usually be controlled in regular relevance feedback evaluation, but must vary in evaluating active feedback algorithms.

In our evaluation, we decided to include all the judged documents, including both relevant and non-relevant documents, because if we exclude them, we would have a potentially *different* test set for each method. In particular, it would be unfair for a method that tends to present more "easy" relevant documents for feedback; indeed, the retrieval task would become artificially harder for such a method due to the fact that more "easy to retrieve" relevant documents would be excluded.

However, including such judged documents also has a problem – it does not accurately reflect the actual utility of a method as perceived by a user. Indeed, a user would presumably not really care about where the judged feedback documents are ranked because the user has already seen them. Thus any improvement in the ranking of a seen relevant document does not really bring any real benefit to the user.

Gap		0	2	3	4	5	6	8	10	15	20
HARD	MAP	0.3247	0.3277	0.3275	0.3289	0.3285	**0.3300**	0.3298	0.3262	0.3289	0.3267
	pr@10	0.5271	**0.5563**	0.5438	0.5396	0.5521	0.5479	0.5479	0.5500	0.5417	0.5292
	#APRel	3.0	3.1	3.0	2.7	2.8	2.6	2.7	2.7	2.5	1.9
AP88-89	MAP	0.2284	0.2332	0.2320	0.2317	0.2323	0.2303	0.2284	**0.2344**	0.2303	0.2243
	pr@10	0.3511	0.3837	**0.3913**	0.3826	0.3880	0.3859	0.3761	0.3891	0.3826	0.3609
	#AFRel	2.1	1.9	1.8	1.8	1.6	1.5	1.7	1.5	1.2	0.9

Table 1: Average Performance of Gapped Top K with different gaps. The best performance is shown in bold.

N		6	20	40	60	80	100
HARD	MAP	0.3247	0.3280	0.3303	0.3277	0.3289	**0.3318**
	pr@10	0.5271	0.5563	0.5479	0.5583	0.5500	**0.5646**
	#AFRel	3.0	2.9	2.6	2.4	2.5	2.4
AP88-89	MAP	0.2284	0.2310	**0.2368**	0.2279	0.2318	0.2341
	pr@10	0.3511	0.3804	**0.3934**	0.3804	0.3739	0.3826
	#AFRel	2.1	1.9	1.3	1.3	1.4	1.2

Table 2: Average Performance of K Cluster Centroid with N. The best performance is shown in bold.

In order to see more clearly how much a method can improve the ranking of unseen documents, we can run the active feedback algorithms on one document database (i.e., the training database) to obtain relevance judgments and then use another *similar* document database (i.e., the testing database) to test the retrieval performance [21]. Thus, in addition to the regular evaluation on the HARD track data set and AP88-89 with all the judged documents included, we also use AP88-89 for training and AP90 for testing to compare different methods, assuming that the contents in these two databases are sufficiently similar.

4. EXPERIMENT RESULT

4.1 Gapped Top K

As we mentioned in Section *2.2.1*, Top K can be considered as a special case of Gapped Top K (i.e. when the gap equals to 0). We do experiments varying the gap to test whether a non-zero gap can perform better than Top K. The results on the HARD data set and AP88-89 data set are shown in Table 1, where we show the MAP, the precision at 10 documents, and the number of judged relevant documents per query.

From the results, we can see Top K ($gap = 0$) is clearly not the best strategy. Actually, when we choose small gaps ($gap \leq 6$), the performance is consistently better than Top K, which strongly suggests that top K is really a poor choice for active relevance feedback. We may also note that, as we increase the gap, we obtain fewer relevant documents than we could obtain with Top K. But using these fewer relevant documents for feedback can achieve better retrieval performance, which means these fewer relevant documents have more learning benefits. The same phenomenon is also observed when active learning is applied in the classification problem [19]. One explanation of this phenomenon is that when we increase the gap, we obtain more diverse documents, thus the judgments become more informative.

4.2 K Cluster Centroid

Here we use the clustering algorithm to select more diverse documents for active relevance feedback. We cluster the top N documents into K clusters and choose the K cluster centroid for relevance feedback. When $N = K$, we again have Top K as a special case. We vary N for fixed K ($= 6$) to test if presenting documents with higher diversity is beneficial. The results are shown in Table 2.

The variation of N causes a different tradeoff point for relevance and diversity. If we choose a bigger N, we pay more attention to diversity, while if we choose a smaller N, we pay more attention to relevance. We see that the optimal values are different for the two databases. Comparing Top K ($N = K$) with other results in the Table again shows that Top K is mostly the worst among all the results, suggesting that the relevance judgments obtained with clustering are more effective for feedback than those obtained using Top K. Moreover, with a large N, we actually obtain fewer judged relevant documents, but these fewer relevant documents are better examples for learning.

4.3 Comparison of Different Algorithms

Since the effectiveness of the underlying feedback mechanism (the mixture model method in our case) is an important factor that may affect our evaluation, we compare several different feedback algorithms with the non-feedback baseline in Table 3. The performance for the Gapped Top K and the K Cluster Centroid is the best performance from Table 1 and Table 2, respectively.

From these results, we can see that the performance of both active feedback and pseudo feedback are better than that of baseline retrieval. We also see that the Top K relevance feedback performs better than using the top K documents for pseudo feedback. All these results show that the underlying feedback mechanism is effective.

Among active feedback algorithms, K cluster centroid outperforms Gapped Top K algorithm, which in turn outperforms Top K algorithm, although the improvement appears to be quite small. A very interesting observation is that the K cluster centroid algorithm obtains the fewest number of relevant documents from user feedback, yet its performance is the best. This suggests that selecting diverse documents leads to more effective learning.

As mentioned in Section 3.4, when comparing different active feedback algorithms, it is more reasonable to use one document database for active feedback (training), and the other document database for measuring retrieval performance (testing). Thus we further compare these methods using AP88-89 as the training set and AP90 as the testing set. Specifically, we perform baseline retrieval on AP88-89 database, select a document subset for relevance

Method		Baseline	K pseudo feedback	Top K	Gapped Top K	K Cluster Centroid
HARD	MAP	0.3076	0.3195	0.3247	0.3300**	**0.3318**
	pr@10	0.5014	0.5146	0.5271	0.5479*	**0.5646**
	#AFRel	/	/	3.0	2.6	2.4
AP88-89	MAP	0.2007	0.2184	0.2284	0.2344*	**0.2368****
	pr@10	0.3255	0.3426	0.3511	0.3891**	**0.3934****
	#AFRel	/	/	2.2	1.5	1.3

Table 3: Average performance of different active learning algorithms. The best performance is shown in bold. A double star (**) and a single star (*) indicate that the performance is significantly better than that of Top K according to Wilcoxin signed rank test at the level of 0.05 and 0.1, respectively.

Method	Baseline	K pseudo FB	Top K	Gapped Top K	K Cluster Centroid
MAP	0.2026	0.2196	0.2203	0.2219	**0.2232**
pr@10	0.2946	0.3174	0.3207	**0.3261****	0.325

Table 4: Average performance of different retrieval algorithms on AP90 data set. The best performance is shown in bold. A double star (**) indicates that the performance is significantly better than that of Top K according to Wilcoxin signed rank test at the level of 0.05.

feedback using different active relevance feedback algorithms, update the query model, all on AP88-89, and then retrieve documents from AP90. The experiment results are shown in Table 4.

The results again show that the results of the Top K algorithm is the worst among three active relevance feedback algorithms. Although the performance difference is mostly insignificant according to the Wilcoxin signed rank test except in the case of pr@10 for Gapped Top K, there are more topics for which Gapped Top K and K Cluster Centroid are better than Top K than the other way in all cases. In the case of MAP, it is 42 topics vs. 31 topics (with 19 cases tied) for both Gapped Top K and K Cluster Centroid. In the case of pr@10, it is 12 topics vs. 3 topics (with 77 cases tied) and 9 topics vs. 5 topics (with 78 cases tied) for Gapped Top K and K Cluster Centroid, respectively. The large number of tied cases indicates that our query expansion feedback mechanism is conservative. Indeed, as we show later in Table 6, when we change the query expansion parameter to perform more aggressive query expansion, the performance improvement is generally amplified. The performance of all active feedback algorithms is also better than that of pseudo feedback and baseline retrieval.

4.4 Performance Sensitivity of K

The results shown so far are all obtained by fixing $K = 6$. We now examine how choosing a different K may affect our conclusions. We compare Top K, Gapped Top K (gap=3), and K cluster centroid ($N = 100$) for several different values of K in Table 5. The results show that our conclusion, i.e., the performances of Gapped Top K and K Cluster Centroid are better than that of Top K, is relatively insensitive to the choice of K. Indeed, the Top K results are almost always the worst among the three methods. Also, on the HARD data, the K cluster centroid method consistently outperforms the other two methods with fewer judged relevant documents.

4.5 Mixture Feedback Algorithm Parameter α Factor

In the results shown so far, the improvement of Gapped Top K and K Cluster Centroid over Top K is not so significant. We find that the feedback algorithm parameter is an important factor. In [27], the new query model $\hat{\theta}_{Q'}$ is

$$\hat{\theta}_{Q'} = (1-\alpha)\hat{\theta}_Q + \alpha \times \hat{\theta}_F$$

Here, α controls how much weight we give to feedback documents. In all the previous results, we set α to 0.5. But since the feedback documents are judged to be relevant by users, we can give more weight to these feedback documents. So we did another set of experiments by varying α and keeping all other parameters fixed. The results are shown in Table 6. From these results, we can see clearly that the α can amplify the effect of feedback. And when α is increased, the improvement of Gapped Top K and K Cluster Centroid over Top K is also amplified.

5. CONCLUSIONS AND FUTURE WORK

This paper presents the first serious study of the problem of active relevance feedback, in which a retrieval system actively chooses the best documents for relevance feedback. Ad hoc information retrieval is largely an interactive process. Active relevance feedback allows a retrieval system to actively probe a user and clarify the user's information need, thus can improve retrieval performance.

We formulate the problem of active feedback as a statistical decision problem and study several special cases. We analyze the assumptions made in each case. We derive three specific algorithms for active relevance feedback, i.e., Top K, Gapped Top K, and K Cluster Centroid algorithm. We evaluate these algorithms using the TREC2003 HARD data set, AP88-89 and AP90 data set. Experiment results show that the Top K algorithm, which is what an existing retrieval system normally uses for relevance feedback, is not optimal for active relevance feedback, and is actually often worse than both the Gapped Top K algorithm and the K Cluster Centroid algorithm. Compared with the Top K algorithm, Gapped Top K algorithm and K Cluster Centroid algorithm emphasize returning more diversified documents. The results show that with fewer judged relevant documents, both Gapped Top K and K Cluster Centroid outperform the Top K algorithm, suggesting that the diversity in the presented documents is a desirable property. Although the difference is generally small, the overall consistency strongly supports our conclusions.

Method		HARD			AP88-89		
		Top K	Gapped Top K	K Cluster Centroid	Top K	Gapped Top K	K Cluster Centroid
K=2	MAP	0.3235	0.3239	0.3204	0.2216	0.2184	0.2145
	pr@10	0.5167	0.5146	0.5333	0.3576	0.3457	0.3533
	#AFDoc	1.1	1.1	0.8	0.8	0.7	0.4
K=4	MAP	0.3253	0.3263	0.3299	0.2228	0.2301	0.2261
	pr@10	0.5271	0.5292	0.5480	0.3521	0.3837	0.3620
	#AFDoc	2.0	2.0	1.8	1.5	1.3	1.0
K=6	MAP	0.3247	0.3275	0.3264	0.2285	0.2320	0.2249
	pr@10	0.5271	0.5438	0.5458	0.3511	0.3913	0.3740
	#AFDoc	3.0	3.0	2.3	2.1	1.8	1.3
K=8	MAP	0.3248	0.3270	0.3360	0.2307	0.2346	0.2334
	pr@10	0.5250	0.5396	0.5708	0.3532	0.3902	0.3859
	#AFDoc	4.0	3.8	3.0	2.7	2.1	1.6
K=10	MAP	0.3249	0.3274	0.3304	0.2319	0.2375	0.2304
	pr@10	0.5271	0.5500	0.5563	0.3663	0.3924	0.3740
	#AFDoc	5.1	4.6	3.6	3.3	2.5	1.9
K=12	MAP	0.3256	0.3282	0.3341	0.2339	0.2374	0.2363
	pr@10	0.5396	0.5438	0.5521	0.3859	0.3880	0.3957
	#AFDoc	6.1	5.3	4.4	3.9	2.8	2.2

Table 5: Sensitivity of average performance of different active learning algorithms on K.

Method		HARD			AP88-89		
		Top K	Gapped Top K	K Cluster Centroid	Top K	Gapped Top K	K Cluster Centroid
α=0.5	MAP	0.325	0.328	0.326	0.228	0.232	0.225
	pr@10	0.527	0.544	0.546	0.351	0.391	0.374
α=0.6	MAP	0.332	0.335	0.340	0.239	0.244	0.236
	pr@10	0.529	0.552	0.556	0.370	0.407	0.390
α=0.7	MAP	0.339	0.344	0.357	0.251	0.259	0.250
	pr@10	0.544	0.575	0.594	0.387	0.418	0.409
α=0.8	MAP	0.348	0.355	0.348	0.264	0.277	0.267
	pr@10	0.552	0.577	0.581	0.404	0.442	0.431
α=0.9	MAP	0.356	0.368	0.388	0.275	0.295	0.273
	pr@10	0.544	0.602	0.640	0.421	0.472	0.442
α=0.95	MAP	0.350	0.367	0.341	0.276	0.300	0.274
	pr@10	0.548	0.602	0.577	0.428	0.479	0.429
α=0.98	MAP	0.337	0.350	0.307	0.270	0.293	0.263
	pr@10	0.527	0.598	0.546	0.423	0.471	0.436

Table 6: Average performance of different active learning algorithms on different α.

Our work represents only a very preliminary exploration of this important topic. There are several interesting directions to explore. (1) It would be interesting to study how to learn from non-relevant documents judged by the user so as to make full use of user efforts and feedback. (2) Another interesting question is how to optimize performance over the entire search session, rather than just one iteration. (3) We may explore other approaches for selecting documents. For example, MMR strategy is also a promising strategy. (4) We can try to combine pseudo feedback and active feedback. Since those highly ranked documents are very likely relevant, we do not really need to present them for judgments; instead, we can propose documents ranked below a few top documents for feedback (essentially the uncertainty sampling strategy). In this way, feedback can be based on those top-ranked documents, which we assume to be relevant, and the obtained relevance judgments through active feedback.

6. ACKNOWLEDGEMENTS

We thank Karen Spärck Jones, Stephen Robertson and the anonymous reviewers for their useful comments. This material is based in part upon work supported by the National Science Foundation under award numbers CAREER-IIS-0347933 and ITR-IIS-0428472.

7. REFERENCES

[1] J. Allan. HARD track overview in TREC2003. In *Proceedings of TREC 2003*, 2003.

[2] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summarires. In *Proceedings of SIGIR 1998*, 1998.

[3] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[4] D. Harman. Relevance feedback revisited. In *Proceedings of SIGIR 1998*, 1992.

[5] T. Jaakkola and H. Siegelmann. Active information retrieval. In *Proceedings of NIPS 2001*, 2001.

[6] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of SIGKDD 2002*, 2002.

[7] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[8] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[9] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR 2001*, pages 111–119, 2001.

[10] D. D. Lewis. Active by accident: Relevance feedback in information retrieval. *Unpublished Working Notes of 1995 AAAI Fall Symposium on Active Learning*, 1995.

[11] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Proceedings of ICML 1994*, 1994.

[12] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of SIGIR 1994*, pages 3–12, 1994.

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

[14] A. K. McCallum and K. Nigam. Employing EM in pool-based active learning for text classification. In *Proceedings of ICML 1998*, 1998.

[15] S. E. Robertson, H. Zaragoza, and M. Taylor. Microsoft Cambridge at TREC-12: HARD track. In *Proceedings of TREC 2003*, 2003.

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

[17] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of ICML 2001*, 2001.

[18] G. Salton and C. Buckley. Improving retrieval performance by retrieval feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[19] G. Schohn and D. Cohn. Less is more: Active learning with support vector machine. In *Proceedings of ICML 2001*, pages 839–846, 2001.

[20] X. Shen and C. Zhai. Active feedback–UIUC TREC2003 HARD experiments. In *Proceedings of TREC 2003*, 2003.

[21] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35(1):30–48, 1979.

[22] S. Tong. *Active Learning: Theory and Applications*. PhD thesis, Stanford University, 2001.

[23] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proceedings of ICML 2000*, 2000.

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

[25] C. Zhai. *Risk Minimization and Language Modeling in Text Retrieval*. PhD thesis, Carnegie Mellon University, 2002.

[26] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of SIGIR 2003*, pages 10–17, 2003.

[27] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM 2001*, pages 403–410, 2001.

[28] C. Zhang and T. Chen. An active learning framework for content-based information retrieval. *IEEE Transactions on Multimedia*, 4:260–268, 2002.

[29] Y. Zhang, W. Xu, and J. P. Callan. Exploration and exploitation in adaptive filtering based on Bayesian active learning. In *Proceedings of ICML 2003*, 2003.