A Bayesian Logistic Regression Model for Active Relevance Feedback

Zuobing Xu
School of Engineering
University of California
Santa Cruz, CA, USA, 95064
zbxu@soe.ucsc.edu

Ram Akella
School of Engineering
University of California
Santa Cruz, CA, USA, 95064
akella@soe.ucsc.edu

ABSTRACT

Relevance feedback, which traditionally uses the terms in the relevant documents to enrich the user's initial query, is an effective method for improving retrieval performance. The traditional relevance feedback algorithms lead to overfitting because of the limited amount of training data and large term space. This paper introduces an online Bayesian logistic regression algorithm to incorporate relevance feedback information. The new approach addresses the overfitting problem by projecting the original feature space onto a more compact set which retains the necessary information. The new set of features consist of the original retrieval score, the distance to the relevant documents and the distance to non-relevant documents. To reduce the human evaluation effort in ascertaining relevance, we introduce a new active learning algorithm based on variance reduction to actively select documents for user evaluation. The new active learning algorithm aims to select feedback documents to reduce the model variance. The variance reduction approach leads to capturing relevance, diversity and uncertainty of the unlabeled documents in a principled manner. These are the critical factors of active learning indicated in previous literature. Experiments with several TREC datasets demonstrate the effectiveness of the proposed approach.

Categories and Subject Descriptors

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

General Terms

Algorithms

1. INTRODUCTION

In information retrieval, it is well known that the original query formulation does not always capture user's semantic search intent. Relevance feedback [7, 17] can improve retrieval performance significantly. The relevance feedback approaches model the following retrieval process: user first sends an initial tentative query, which retrieves some useful documents for user to evaluate. Based on the relevance evaluation, the retrieval system modifies the query to retrieve more relevant documents in the next retrieval round. Query expansion [23, 14], which extracts terms relevant to the search topic from feedback documents to reformulate the original query, is an essential element in traditional relevance feedback.

"An inductive algorithm *overfits* the dataset if it models the given data too well and its predictions are poor[13]." The problem of overfitting has attracted attention in the machine learning community [13, 8] for a long period, but has not been as well studied in the context of the relevance feedback problem. Traditional relevance feedback algorithms which rely on query expansion suffer from a crucial drawback: *In relevance feedback the size of feedback (training) documents is much smaller than the size of the term space. Consequently, learning from limited feedback documents with many features will cause the overfitting problem [13].* Because of the existence of noise, some terms including the background noise terms can only discriminate between the relevant and non-relevant feedback documents, but cannot generalize to rank the relevancy of the remaining unlabeled documents, which is the *overfitting problem.* Overfitting is closely related to the bias-variance trade-off: if the algorithm is optimized to fit the training data too well, the variance term becomes too large. In the case where training data is limited, the variance term becomes even larger. The existing query expansion algorithms implicitly alleviate the overfitting problem [23] by filtering out the background noise terms that have a poor generalization power and only choosing the terms with largest probabilities in the feedback model, although they do not explicitly discuss the *overfitting problem.* Nevertheless, relevance feedback algorithms using the complete term space cannot avoid the overfitting problem completely.

We propose a Bayesian logistic regression model[6] to predict the probability of relevance of the retrieved documents. Bayesian logistic regression extends the logistic regression model to a Bayesian framework by adding a prior distribution of the parameters to the model. *To address the overfitting problem, we reduce the term space to three features: retrieval score, distance to relevant documents and distance to non-relevant documents. The new algorithm re-estimates the probability of relevance for the initial retrieved documents by considering their retrieval score, distance to the relevant feedback documents and distance to the non-relevant documents.* One notable distinction between the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'08, July 20–24, 2008, Singapore.
Copyright 2008 ACM 978-1-60558-164-4/08/07 ...$5.00.

logistic regression relevance feedback model and the traditional relevance feedback model is the abstraction in the term feature space. Query expansion based on term space expands the original query with noisy terms from feedback documents, which could then impair the retrieval performance. The noisy terms come either from general English words or some document specific words. Reducing the feature space in the logistic regression model will significantly reduce the overfitting problem. The distance features in the logistic regression model update as more documents are selected for evaluation. If relevance feedback is processed in a batch setting, because we do not have labeled documents initially, the distance measure for the feedback documents cannot be calculated until all the feedback documents in the batch have been evaluated. Therefore, we propose to use online Bayesian logistic regression. In the online setting, the user evaluates feedback document one at a time, and then the algorithm updates the distance measures, as well as the regression parameters, upon the user's feedback. We also impose a strong Bayesian prior, which captures our prior belief of the model parameters, to further mitigate the overfitting problem.

How to actively choose good documents to present to the user for evaluation is another challenging problem, whose resolution further improves the performance of the relevance feedback process. The active choice of unlabeled data belongs to the broad area of active learning problems in supervised learning. Choosing uncertain data close to the decision boundary has been the primary active learning strategy in most kinds of machine learning tasks [3, 5, 18, 21]. On the other hand, this problem has not been well studied in the information retrieval community. Several efficient and effective heuristics [22, 19] have been proposed to focus on increasing the diversity of the chosen document set. In this paper, we propose a new active learning algorithm for the Bayesian logistic regression model. In order to reduce the overall prediction error, the new active learning algorithm tends to select documents which minimize the variance of the parameter posterior distribution. Our variance reduction approach for the active learning algorithm leads to capturing the elements which have been addressed in previous literature, such as relevance[22], diversity [22, 19] and uncertainty [3, 5, 18, 21] of the unlabeled documents, in a more principled and integrated manner. In [22], an effective active learning algorithm, which chooses documents based on their relevance score, diversity measure and density, achieves good performance.

The remainder of this paper is organized as following. In section 2, we review the related literature on the Bayesian logistic regression and active learning. In section 3, we first introduce the Bayesian logistic regression model for relevance feedback. In section 4, we present the new active learning approach. In section 5, we discuss the experimental setting and the experimental results. In Section 6, we conclude with a description of our current research, and present several future research directions.

2. RELATED WORK

Logistic regression [8] is one of the most widely used discriminative models in data mining. Regularization methods express our prior belief in the parameters, and penalize the estimate for deviating from the prior belief. In practice, we need to regularize the logistic regression model to avoid overfitting caused by limited number of training data with large feature size. For the Bayesian approach, regularization is achieved by specifying a prior distribution over the parameters and subsequently averaging over the posterior distribution. Genkin et al. [6] proposed Bayesian logistic regression to perform large scale text categorization and demonstrated good predictive capabilities. Dayanik et al. [4] incorporated domain knowledge by constructing informative prior distribution for the Bayesian logistic regression model. Bayesian logistic regression has also been applied in adaptive filtering [25] to learn the filtering threshold.

Active feedback is essentially an application of active learning in the Ad hoc information retrieval area. Active learning has been extensively studied in the supervised learning scenarios. Active learning algorithms can be categorized into two classes: those which choose unlabeled data based on the uncertainty of the data and those which choose unlabeled data based on the expected utility of the data.

Lewis and Gale [3] proposed an uncertainty sampling method for active learning. The Query by Committee (QBC) algorithm [5, 18] measures the uncertainty of a test example by employing the disagreement among different classification models as an effective score. Tong's support vector machine active learning approach [21] aims to select unlabeled documents to reduce the version space as much as possible. To summarize, these three approaches select documents close to the decision boundary, and belong to the first category.

Cohn et al. [2] proposed one of the first statistical analysis of active learning, demonstrating how to construct queries that maximize error reduction by minimizing the learner's variance. Roy and McCallum [15] proposed an active learning algorithm which reduces the expected log loss. Both of the above two algorithms calculate the expected loss, and belong to the second category of utility based approaches. The ideal loss function is the difference between the true model and the learned model. Because we do not know the true model in advance, we typically develop a surrogate objective function that approximates the model quality.

Active learning strategies have also been studied for the logistic regression model. Zhang and Oles [24] analyzed the value of the unlabeled data and presented a framework of active learning based on the maximization of the Fisher information matrix, given that the Fisher information represents the overall uncertainty of the classification model. Hoi et al.[9, 10] extended the framework of [24] to a batch learning setting. The first algorithm [9] solves the combinatorial optimization problem with an efficient greedy algorithm that approximates the objective function by a submodular function. The second algorithm [10] relaxes the integer value constraints to continuous value constraints, and thus the NP hard combinatorial optimization problem becomes a semidefinite programming (SDP) problem.

Active learning has not been well studied in the context of relevance feedback. Shen et al. [19] proposed several heuristic algorithms to stress the diversity of the feedback document set. Xu et al. [22] proposed an active learning algorithm which comprehensively considers the relevance, diversity and density of the feedback document set. Both of these two active learning approaches are based on the existing language model based relevance feedback algorithms [23], while the new active learning approach proposed in this paper is based on the Bayesian logistic regression relevance feedback model which addresses the overfitting problem. Be-

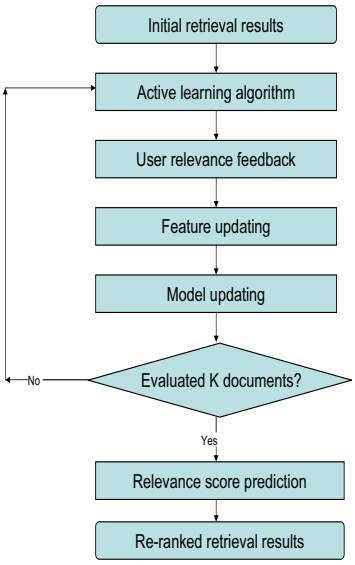

Figure 1: Procedure of Active Bayesian Logistic Regression Relevance Feedback Model

cause the overfitting problem is caused by large variance in the prediction achieved by the model, our new active learning approach chooses the feedback document which reduces the expected variance of the classification model the most. The variance reduction approach captures relevance, diversity and uncertainty of the unlabeled documents in a principled manner. Those factors have been identified as critical factors of active learning in the previous literature[19, 22, 3, 5, 18, 21].

3. BAYESIAN LOGISTIC REGRESSION FOR RELEVANCE FEEDBACK

The new relevance feedback algorithm is setup in an online setting, where the feedback documents are evaluated one at a time. We model the active relevance feedback in a Bayesian logistic regression framework, where the original term feature space is reduced to three features: retrieval score, distance to relevant feedback documents and distance to non-relevant feedback documents. We first use an active learning algorithm to select a feedback document for user evaluation. Based on the user's evaluation, we update the parameters of the logistic regression model, as well as the distance features since the feedback document sets are increasing. Based on the current model parameters and feature values, we select the second feedback document using the active learning algorithm. The above procedure is iterated until we have evaluated K documents. The procedure is illustrated in Figure 1. In the following sections, we will present the Bayesian logistic regression model, its application to relevance feedback, and the active learning algorithm in detail.

3.1 Bayesian logistic Regression Model

The goal of the logistic regression model is to predict the probability of relevance of the unevaluated documents. Suppose we have a set of training examples

$$D = \{(\boldsymbol{d_1}, y_1), \ldots, (\boldsymbol{d_i}, y_i), \ldots (\boldsymbol{d_n}, y_n)\}$$

The vectors $\boldsymbol{d_i}$ are the features of the training examples, and the values $y_i \in \{+1, -1\}$ are the class labels encoding relevant $(+1)$ or non-relevant (-1) of the vector in the category. Thus, the probability of relevance has the form

$$p(y_i|\boldsymbol{\beta}, \boldsymbol{d_i}) = \pi(y_i, \boldsymbol{d_i}, \boldsymbol{\beta}) = \frac{1}{1 + exp(-\boldsymbol{\beta}^T \boldsymbol{d_i} y_i)} \quad (1)$$

The key of the logistic regression is to estimate parameters $\boldsymbol{\beta}$. The training documents are limited in relevance feedback, and thus the logistic regression model is likely to overfit the training data. Regularization is an effective way to reduce overfitting. Taking a Bayesian point of view, we apply the Bayesian regularization approach [6] which constructs prior distributions on $\boldsymbol{\beta}$. The Gaussian distribution is a commonly used prior distribution. We impose a N dimensional multivariate Gaussian prior distribution for parameters $\boldsymbol{\beta}$ with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$.

$$p\left(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \propto \frac{1}{|\boldsymbol{\Sigma}|^{N/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\mu}\right)\right) \quad (2)$$

Ideally, we integrate over the posterior distribution of $\boldsymbol{\beta}$ to predict the probability of being relevant for the unlabeled documents. The ideal approach faces the following computational problem. The Gaussian distribution is not the conjugate prior for the logistic regression model, and multidimensional integration over the posterior distribution is intractable. Therefore, we are not able to derive a closed-form posterior distribution. Instead, we propose to apply the Laplace approach [11] to approximate the posterior distribution, and use the posterior mean to estimate class probability.

The Laplace method [11] approximates the posterior distribution by a scaled Gaussian distribution. The mean of the Gaussian distribution is the Maximum a posteriori (MAP) estimate of the posterior mean, and the variance matrix is the Hessian of the log posterior distribution.

The likelihood function $\ell(\boldsymbol{\beta})$, which is also the posterior density $p(\boldsymbol{\beta}|D)$ with the logistic regression model, is shown in Equation (3).

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= p(\boldsymbol{\beta}|D) \propto p(D|\boldsymbol{\beta})p(\boldsymbol{\beta}) \quad (3) \\ &\propto \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta} - \boldsymbol{\mu}\right)\right)}{1 + \exp(-\boldsymbol{\beta}^T \boldsymbol{d_i} y_i)} \end{aligned}$$

where $p(\boldsymbol{\beta})$ is the prior distribution as in Equation(2). The Maximum a posteriori estimate (MAP) is a point estimate which maximizes the log of the posterior likelihood function (3). Thus, the MAP estimate is the maximum of the following likelihood function.

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\boldsymbol{map}} &= \arg\max_{\boldsymbol{\beta}} \ln \ell(\boldsymbol{\beta}) \quad (4) \\ &= \arg\max_{\boldsymbol{\beta}} -\ln(1 + \exp(-\boldsymbol{\beta}^T \boldsymbol{d_i} y_i)) - \\ &\quad -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}) \end{aligned}$$

We cannot derive a closed-form solution for the above optimization problem. it can be computed by any gradient descent method. The first derivative and second derivative

of the log-likelihood function can be derived as

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\boldsymbol{d_i} y_i}{1+\exp(\boldsymbol{\beta}^T \boldsymbol{d_i} y_i)} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}) \quad (5)$$

$$\frac{\partial l^2(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \frac{-\boldsymbol{d_i}\boldsymbol{d_i}^T}{(1+\exp(\boldsymbol{\beta}^T \boldsymbol{d_i} y_i))(1+\exp(-\boldsymbol{\beta}^T \boldsymbol{d_i} y_i))} - \boldsymbol{\Sigma}^{-1}$$

The covariance matrix of the posterior distribution can be approximated by the Hessian matrix of the log-likelihood at $\hat{\boldsymbol{\beta}}_{map}$ using the Laplace approach. Thus, we plug β_{map} into Equation (5).

$$\boldsymbol{\Sigma_{new}}^{-1} = \boldsymbol{\Sigma_{old}}^{-1} + \frac{\boldsymbol{d_i}\boldsymbol{d_i}^T}{(1+\exp(\hat{\boldsymbol{\beta}}_{\boldsymbol{map}}{}^T \boldsymbol{d_i} y_i))(1+\exp(-\hat{\boldsymbol{\beta}}_{\boldsymbol{map}}{}^T \boldsymbol{d_i} y_i))} \quad (6)$$

We update the posterior mean and variance after evaluating a feedback document, then the posterior distribution becomes the prior distribution for the next feedback document. After we have evaluated all the feedback documents, we can simply predict the probability of relevance for the remaining unlabeled documents using Equation (1), where the parameter $\boldsymbol{\beta}$ is the posterior mean $\hat{\boldsymbol{\beta}}_{map}$.

3.2 Applying Bayesian Logistic Regression to Relevance Feedback

We usually do not have enough training data in the relevance feedback application. With limited training data, traditional relevance feedback algorithms which use the whole term space as feature space, face the overfitting problem. For example, the overfitting problem can be caused by expanding the query with some off-topic terms occurring in the relevant feedback documents. Consequently, non-relevant documents containing these terms will be ranked highly. Various techniques have been proposed to reduce overfitting, such as deleting general English terms occurring in the collection[23]. To alleviate the overfitting problem, we project the whole term space into three features: retrieval score, distance to the relevant feedback documents and distance to the non-relevant feedback documents. Intuitively, a relevant document should have high retrieval score, small distance to the relevant documents, and large distance to the non-relevant documents. Thus, these three features constitute a new space, which captures the significant elements influencing the relevance ranking.

To normalize all the features in the overall metric to be of comparable values, we normalize the retrieval score and use cosine distance measure. Thus, the values of all the three features range from 0 to 1. We apply standard normalization to normalize the retrieval score of document $\boldsymbol{d_i}$.

$$(RScore_i - RScore_{min}) \,/\, (RScore_{max} - RScore_{min}) \quad (7)$$

There are several ways to measure the distance of a document to a set of documents. Single linkage defines this distance as the minimum distance between the document and any document in the set; Complete linkage is the opposite of single linkage in that it defines this distance as the maximum distance between them; Average linkage takes the mean distance between the document and all the documents in the set. Because the distance measured by the single linkage method decreases with the number of documents that have been selected and the distance measured by the complete linkage method increases with the number of documents, we use the average linkage approach, which offers a smooth distance measure.

Initially, we set both the distance measures to some initial values, and dynamically update the distance measures as each feedback document is evaluated sequentially. We cannot obtain valid distance measures unless we have labeled documents in both classes. Thus, we do not start to train the logistic regression model until we have at least one relevant and one non-relevant feedback document.

Intuitively, the probability of relevance is correlated with the retrieval score and the distance to the non-relevant documents, and inversely correlated with the distance to the relevant documents. Thus, we set the prior mean for the parameters $\boldsymbol{\beta}$ as $\boldsymbol{\mu} = \{\mu_0, \mu_1, \mu_2\}$, where μ_0 and μ_2 are positive, and μ_1 is negative. We set the variance of the prior distribution as $\boldsymbol{\Sigma} = \text{diag}\{\sigma, \sigma, \sigma\}$. In the experiments, we will discuss how to initialize these values.

4. ACTIVE LEARNING FOR BAYESIAN LOGISTIC REGRESSION

We now derive an objective function for active learning in the context of the Bayesian logistic regression. The mean square error of the logistic regression model can be decomposed to three terms.

$$\begin{aligned} err &= \sum_i E\left[y_i - \pi(y_i, \boldsymbol{d_i}, \hat{\boldsymbol{\beta}})\right]^2 \quad (8) \\ &= \sigma_\epsilon^2 && \text{“Noise”} \\ &\quad +[E(\pi(y_i, \boldsymbol{d_i}, \hat{\boldsymbol{\beta}})) - \pi(y_i, \boldsymbol{d_i}, \boldsymbol{\beta})]^2 && \text{“Bias”} \\ &\quad +E[\pi(y_i, \boldsymbol{d_i}, \hat{\boldsymbol{\beta}}) - E(\pi(y_i, \boldsymbol{d_i}, \hat{\boldsymbol{\beta}}))]^2 && \text{“Variance”} \end{aligned}$$

A similar error decomposition is discussed in [8]. The goal of the active learning algorithm is to select unlabeled data which minimize the expected mean square error. The first term σ_ϵ^2 is the variance of the target around its true mean, and cannot be avoided no matter how accurately we estimate parameter $\boldsymbol{\beta}$; the second term is the squared bias, the amount by which the average of our estimate deviates from the true mean; the last term is the variance. Since we do not know the true parameter $\boldsymbol{\beta}$, calculating the bias term is infeasible. Thus, our active learning algorithm only focuses on minimizing the variance term.

We approximate the parameters of the posterior distribution by a Gaussian distribution using the Laplace method. The inverse of the covariance matrix can be approximated by $-\partial l^2(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^2$. Consequently, we shall favor an unlabeled document that decreases the variance significantly. Recall from Equation (6) that we will choose document $\boldsymbol{d_i}$ which maximizes

$$Score_i = \frac{1}{(1+\exp(\boldsymbol{\beta}^T \boldsymbol{d_i}))(1+\exp(-\boldsymbol{\beta}^T \boldsymbol{d_i}))} \boldsymbol{d_i}\boldsymbol{d_i}^T \quad (9)$$

In the above score function, the first term is maximized when $\boldsymbol{\beta}^T \boldsymbol{d_i} = 0$, which indicates that the data are on the decision boundary. Uncertainty is a commonly used unlabeled document quality measure in active learning. The underlying principle of uncertainty sampling [3], query by committee [5, 18] and version space reduction [21] is to choose uncertain data which are close to the decision boundary. The second term increases with the norm of the feature vector. The intuition is that a document with large retrieval score

and large distance to the previously selected documents is preferable. Relevant documents are more useful than non-relevant documents in relevance feedback, and retrieval score is the only indicator of the relevance before user evaluation, so documents with high retrieval score is more favorable. A large distance to the previously selected documents implies the diversity of the feedback document set, which helps to avoid similar and duplicate documents. An effective heuristic active learning selection approach that explicitly focuses on relevance, diversity and density was proposed in [22]. The new active learning algorithm in this paper, however, follows from principled analytics. We formally derive our active learning approach to minimize the variance of posterior distribution. The new active learning approach implicitly captures several widely applied heuristics in the previous work in active learning.

The Fisher information for logistic regression is defined by the same expression as Equation(9). It is well-known from the standard Cramer-Rao lower bound [12] that the covariance of any unbiased estimator of $\boldsymbol{\beta}$ is $\geq \frac{1}{n} I(\boldsymbol{\beta})^{-1}$. Here $I(\boldsymbol{\beta})$ is the Fisher information of parameter $\boldsymbol{\beta}$. Zhang and Oles [24] proposed an active learning scheme which maximizes the Fisher information. Our variance reduction active learning approach is coincident with the fisher information active learning approach, but is derived from a new perspective. We summarize the computational procedure for active Bayesian regularization in Table 1.

Furthermore, instead of directly selecting documents with highest score, we propose an sampling scheme. Earlier work [16] has demonstrated that sampling from a distribution of effectiveness scores is preferable to direct selection. It reduces the chance of selecting outliers and allows the algorithm to balance exploration and exploitation. We propose two methods to convert the absolute score to a distribution. For the first approach, the sampling distribution weight for document $\boldsymbol{d_i}$ is given by

$$P(\boldsymbol{d_i}) = score_i / \sum_i score_i \tag{10}$$

For the second approach, we applied the softmax action selection rules. The most common method uses a Gibbs, or Boltzmann distribution. It chooses document x_i with probability

$$P(\boldsymbol{d_i}) = \exp(score_i/T) / \sum_i \exp(score_i/T) \tag{11}$$

where T is a positive parameter called the temperature. High temperatures equalize the actions, while low temperatures differentiate the actions. Randomized algorithms are effective and popular to balance between exploration and exploitation[20].

Table 1: Pseudo code for active Bayesian logistic regression algorithm

FUNCTION: Predict probability of relevance **INPUT:** $D = d_1, d_2, \ldots, d_N$ **OUTPUT:** probability of relevance

SET relDoc to 0
SET nonRelDoc to 0
FOR $k = 0$ to $K - 1$
Score a document $\boldsymbol{d_i}$ using Equation (9).
IF $(y_i = 1)$ **THEN**
relDoc ++
Update distRel of the unselected documents.
ELSE
nonRelDoc ++
Update distNonRel of the unselected documents.
END IF
IF(relDoc $>$ 0 and nonRelDoc $>$0) **THEN**
Update posterior mean $\hat{\beta}_{map}$ using Equation (4).
Update posterior variance using Equation (6).
END IF
k++
END FOR
Predict probability of relevance for the remaining documents using Equation (1) with $\hat{\beta}_{map}$.

5. EXPERIMENTAL METHODS AND RESULTS

5.1 Experimental Dataset and Procedure

To evaluate our active Bayesian logistic regression algorithm described in the previous section, we experimented with three TREC datasets. The first one is the TREC 2003 HARD track, which use part of the AQUAINT dataset plus two additional datasets (Congressional Record (CR) and Federal Register (FR)). We do not have the additional datasets in the TREC 2003 HARD track. Our results are still comparable to other published TREC 2003 HARD results, although the data are a little different. The second one is the TREC 7 dataset, which contains data from the TREC Disk 4 and 5 (excludes Congressional Record). The last one is the TREC 8 dataset, which contains the same document set as TREC 7 dataset. Because the topic titles are most similar to the user's practical search behavior, we use only the topic titles as queries on all the 50 topics. Data pre-processing is standard: terms were stemmed using the Porter Stemming and stop words were removed by using standard stop word list.

To measure the performance of the logistic regression relevance feedback algorithms, we use two standard ad hoc retrieval measures: (1) Mean Average Precision (MAP), which is calculated as the average of the precision after each relevant document is retrieved, reflects the overall retrieval accuracy. (2) Precision at 10 documents (Pr@10): this measure does not average well and only gives us the precision for the first 10 documents. It reflects the utility perceived by a user who may only read up to the top 10 documents on the first page. In the experiments, we include all the feedback documents for evaluation, and this evaluation scheme is also applied in [19].

We employed the Lemur Toolkit [1] as our retrieval system and the KL-Divergence language retrieval model as our baseline retrieval model. We first compared the Bayesian logistic regression algorithm with other relevance feedback algorithms such as the mixture model algorithm [23] and the divergence minimization algorithm [23]. The mixture model algorithm models the feedback documents as a mixture of feedback topic model and background collection model. It uses EM algorithm to estimate the feedback topic model and interpolates with the original query model. The divergence minimization algorithm models the relevance feedback in an optimization framework, and tends to minimize the diver-

gence between the feedback topic model and the feedback documents, and in the same time maximize the divergence between the feedback topic model and the background collection model. It also interpolates the original query model with the feedback topic model. We then applied the variance minimization active learning algorithm to the logistic regression model, and compared the result with other active learning algorithms, including the TOP K [19], cluster Centroid [19], Active RDD algorithm [22], and random selection.

5.2 Comparison of Relevance Feedback Algorithms

In the experiments, we used TREC 2003 HARD as training data and optimized the MAP performance by tuning parameters μ_1, μ_2, μ_3 and σ. We set the Bayesian regression prior parameters $\mu_1 = 2$, $\mu_2 = -4$, $\mu_3 = 2$ and $\sigma = 1$. Thus, the logistic regression model (LogReg) put a larger weight on the relevant documents than the non-relevant documents since the relevant documents are more indicative of relevance than the non-relevant documents. We also choose the parameters for the mixture models relevance feedback algorithm (Mixture) and the divergence minimization relevance feedback algorithm (DivMin) [23] in the same way. We set the weighting parameter $\lambda = 0.8$ and the interpolation parameter $\alpha = 0.8$ for the mixture model algorithms, and $\lambda = 0.8$ and $\alpha = 0.4$ for the divergence minimization algorithm. We applied the same parameter setting for the other two datasets. In the experiments, we set the number of feedback documents $K = 6$. To reduce the computation, we only selected feedback documents from the top 100 documents and re-ranked the top 1000 retrieved documents. To validate the effectiveness of non-relevant feedback documents, we reduce the feature space in the logistic regression model by eliminating the feature of distance to non-relevant documents. We also include this approach (LogReg-2) in the comparison.

Table 2 shows the performance comparison. When compared with the model based feedback results, which themselves are actually very strong when compared with the published official TREC results, the Bayesian logistic regression algorithm performs significantly better than the divergence minimization model in terms of both MAP and PR@10 with a $8\% - 28\%$ margin. The logistic regression algorithm also performs significantly better than the mixture model approach with a upto 20% margin. When compared with the logistic regression algorithm without non-relevant feedback, the logistic regression algorithm with both relevant and non-relevant feedback performs better with upto 7% improvement. This shows that the non-relevant feedback documents help to improve the retrieval accuray by lowering the rank of the documents similar to the non-relevant feedback documents.

5.3 Comparison of Active Learning Algorithms

We used the Bayesian logistic regression as our relevance feedback baseline, and compared our variance reduction active learning algorithm (Variance) and its random sampling extensions (Variance-sampling and Variance-softmax) with the existing active learning algorithms, such as top K, cluster centroid[19], random selection and Active-RDD [22]. We tuned parameter values for these active learning algorithms based on the TREC 2003 HARD dataset. We set the relevance parameter equal to 0.3 and the diversity parameter equal to 0.3 for the Active-RDD algorithm [22]. We set the temperature parameter $T = 1$ in the softmax algorithm.

Table 3 shows the comparison results. From the results, we conclude that the variance reduction algorithm performs consistently well among all the active learning approaches. The Variance-sampling algorithm and Variance-softmax algorithm perform better than the basic variance reduction method on TREC 2003 HARD dataset, and perform worse on the other two datasets. The TREC 2003 HARD dataset is considered as a dataset with easy queries, because it has a better baseline retrieval performance than the TREC 7 and TREC 8 datasets without feedback. Since the probability of selecting similar relevant documents from the top ranked list is high for easy queries, selecting a diversified feedback document set helps to improve the retrieval performance for the dataset with easy queries. Randomized approaches bring more exploration to the document selection scheme and tend to choose more diversified documents. Therefore, the randomized approaches benefit the datasets with easy queries (TREC 2003 HARD). By contrast, datasets with difficult queries (TREC 7 and TREC 8) do not benefit from the randomness in the Variance-sampling and Variance-softmax algorithms.

All the results shown so far were obtained by fixing the feedback document size $K = 6$. We also examined how the value of K may affect our conclusions. We compared the top K, random selection and variance reduction algorithm by varying K in Figure 2. From Figure 2, we conclude that our variance reduction active learning algorithm consistently performs better than all the other active learning algorithms for the TREC 7 and TREC 8 datasets, and performs better than Top K and random selection algorithm on the TREC 2003 HARD dataset.

As we discussed before, selecting a diversified feedback document set is helpful to the dataset with easy queries, because the chance of selecting similar relevant documents from the top ranked list is high for easy queries. Therefore, the cluster centroid algorithm, which clusters the retrieved documents and selects the centroid documents from each cluster, emphasizes only diversity, and achieves a better performance than the variance reduction algorithm on the TREC 2003 HARD dataset. When the feedback document size becomes larger, the advantage of our active learning algorithm over the TOP K algorithm on all the three datasets becomes more obvious. The active learning problem in relevance feedback is a cold start problem, which is different from the traditional active learning problem in the supervised learning problems. For a cold start learning problem, positive training data are more valuable than the negative training data for the first a few documents. Consequently, when the feedback document size is small (fewer than 4 documents), the top K algorithm performs almost as well as the variance reduction active learning algorithm.

6. CONCLUSIONS

To address the overfitting problem in relevance feedback, we have proposed a principled active Bayesian logistic regression model for the relevance feedback in information retrieval. The new model reduces the original feature set to three features: retrieval score, distance to relevant documents and distance to non-relevant documents, and online

Table 2: Average performance of relevance feedback approaches. A single star(*) indicates that the performance of our Bayesian logistic regression relevance feedback algorithm is significantly better than the existing methods used in the corresponding column according to Wilcoxon signed rank test at the level of 0.1.

Method		Baseline	Mixture	DivMin	LogReg-2	LogReg	Improv. over Baseline	Improv. over Mixture	Improv. over DivMin	Improv. over LogReg-2
HARD 2003	MAP	0.3198*	0.3817*	0.3553*	0.3832	**0.3851**	20.42%	0.89%	8.39%	0.50%
	Pr@10	0.5000*	0.5420*	0.5200*	0.6140	**0.6140**	22.80%	11.33%	18.08%	0.00%
TREC 7	MAP	0.1868*	0.2476*	0.2253*	0.2401*	**0.2574**	37.79%	3.96%	14.24%	7.20%
	Pr@10	0.4220*	0.4980*	0.4660*	0.5560*	**0.5980**	41.71%	20.08%	28.33%	7.55%
TREC 8	MAP	0.2488*	0.3240	0.2892*	0.3018*	**0.3160**	27.01%	−2.4%	9.34%	4.71%
	Pr@10	0.4560*	0.5860*	0.5240*	0.5800*	**0.6180**	35.53%	5.46%	17.94%	6.55%

Table 3: Average performance of active learning approaches, A single star(*) indicates that the performance of our variance reduction active learning algorithm is significantly better than the existing methods used in the corresponding column according to Wilcoxon signed rank test at the level of 0.1.

		Top K	Random	Cluster	RDD	Variance	Var-sampling	Var-softmax
HARD 2003	MAP	0.3851	0.3792*	0.3932	**0.3951**	0.3891	0.3936	0.3946
	Pr@10	0.6140	0.6180	0.6300	0.6300	0.6100	0.6280	**0.6340**
TREC 7	MAP	0.2574*	0.2268*	0.2325*	0.2541*	**0.2608**	0.2263	0.2346
	Pr@10	0.5980	0.5060*	0.5120*	0.5900*	**0.6020**	0.5200	0.5400
TREC 8	MAP	0.3160*	0.2865*	0.3110*	0.3056*	**0.3268**	0.2805	0.2818
	Pr@10	**0.6180**	0.5340*	0.5660*	0.5860*	0.6060	0.5520	0.5560

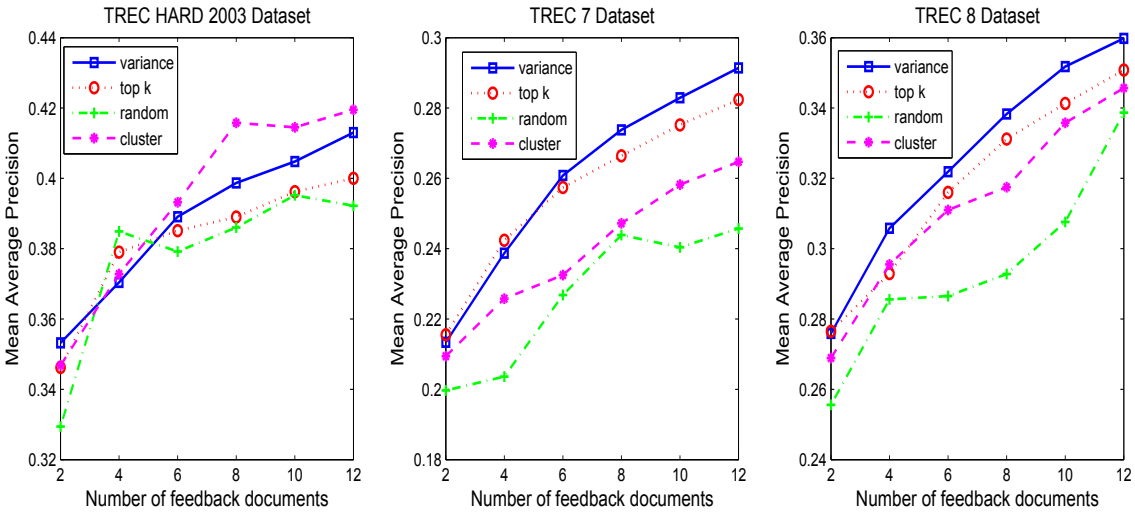

Figure 2: Sensitivity of average performance of different active learning approaches on K.

updates the distance measures as more documents are selected for feedback. The feature reduction approach effectively reduces overfitting. A Bayesian logistic regression approach is applied to learn the parameters of the model. The new active learning algorithm chooses unlabeled documents, so as to minimize the expected variance. Experimental results show significant performance improvement against the existing active relevance feedback algorithms.

There are several interesting research directions that may further improve the effectiveness of active Bayesian logistic regression model. First, an optimal stopping policy can stop the feedback evaluation process optimally, and we would term this as an adaptive algorithm. Second, the current active learning scheme is a greedy algorithm. Consequently, designing a new active learning algorithm which considers the trade-off between exploration and exploitation, will benefit the retrieval performance. Third, the average linkage used in distance measure loses information in higher moments, and thus designing richer feature set that contains more information will benefit the overall performance.

7. ACKNOWLEDGMENTS

We acknowledge support from Cisco, University of California's MICRO program. We also appreciate suggestions from anonymous reviewers.

8. REFERENCES

[1] The lemur toolkit. http://www.lemurproject.org.

[2] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. In *Advances in Neural Information Processing Systems*, volume 7, pages 705–712. The MIT Press, 1995.

[3] L. D. and G. W. Training text classifiers by uncertainty sampling. In *International ACM Conf. on Research and Development in Information Retrieval*, 1994.

[4] A. Dayanik, D. D. Lewis, D. Madigan, V. Menkov, and A. Genkin. Constructing information prior distributions from domain knowledge in text classfication. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006.

[5] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[6] A. Genkin, D. Lewis, and D. Madigan. large-scale bayesian logistic regression for text categorization. Technical report, DIMACS, 2004.

[7] D. Harman. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, 1992.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Spinger, 2001.

[9] S. Hoi, R. Jin, J. Zhu, and M. Lyu. Batch mode active learning and its application to medical image classification. In *proceedings of the 23rd international conference on machine learning*, 2006.

[10] S. C. H. Hoi, R. Jin, and M. R. Lyu. Large scale text categorization by batch mode active learning. In *Proceedings of International World Wide Web Conference*, 2006.

[11] R. Kass, L. Tierney, and J. Kadane. validity of posterior expansions based on laplace's method. *Bayesian and likelihood methods in statistics and econometrics*, 1990.

[12] S. Kay. *Fundamentals of statistical signal processing*. Prentice-Hall, 1993.

[13] R. Kohavi and D. Sommerfield. Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. In *The First International Conference on Knowledge Discovery and Data Mining (KDD)*, 1995.

[14] J. Rocchio. *Relevance feedback in information retrieval,In The Smart System - experiments in automatic document processing*. Prentice Hall, Englewood Cliffs, NJ, 1971.

[15] N. Roy and A. McCallum. Toward optimal active learning through sampling estimation of error reduction. In *Proc. 18th International Conf. on Machine Learning*, pages 441–448. Morgan Kaufmann, San Francisco, CA, 2001.

[16] M. Saar-Tsechansky and F. Provost. Active sampling for class probability estimation and ranking. *Machine learning*, pages 153–178, 2004.

[17] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):133–168, 1990.

[18] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, 1992.

[19] X. Shen and C. Zhai. Active feedback in ad hoc information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–66, March 2005.

[20] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[21] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Proceedings of 17th International Conference on Machine Learning*, pages 999–1006, 2000.

[22] Z. Xu, R. Akella, and Y. Zhang. Incorporating diversity and density in active learning for relevance feedback. In *29th European Conference on Information Retrieval (ECIR)*, 2007.

[23] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth ACM International Conference on Information and Knowledge Management*, pages 403–410, 2001.

[24] T. Zhang and F. J. Oles. A probability analysis on the value of unlabeled data for classification problems. In *International Conference on Machine Learning*, pages 1191–1198, 2000.

[25] Y. Zhang, W. Xu, and J. Callan. Exploration and exploitation in adaptive filtering based on bayesian active learning. In *Proceedings of 20th International Conf. on Machine Learning*, pages 896–903, 2003.