Relevance Feedback using Latent Information

Jun Harashima
Graduate School of Informatics,
Kyoto University,
Yoshida-honmachi, Sakyo-ku,
Kyoto, 606-8501, Japan
harashima@nlp.kuee.kyoto-u.ac.jp

Sadao Kurohashi
Graduate School of Informatics,
Kyoto University,
Yoshida-honmachi, Sakyo-ku,
Kyoto, 606-8501, Japan
kuro@i.kyoto-u.ac.jp

Abstract

We present a novel relevance feedback (RF) method that uses not only the surface information in texts, but also the latent information contained therein. In the proposed method, we infer the latent topic distribution in user feedback and in each document in the search results using latent Dirichlet allocation, and then we modify the search results so that documents with a similar topic distribution to that of the feedback are re-ranked higher. Evaluation results show that our method is effective for both explicit and pseudo RF, and that it has the advantage of performing well even when only a small amount of user feedback is available.

1 Introduction

The main purpose of information retrieval (IR) is to provide the user with documents that are relevant to his/her information needs. However, it is difficult to achieve this by one-off retrieval, since user queries are typically short and often ambiguous (Jansen et al., 2000).

Relevance feedback (RF) is a technique to solve this problem. The basic procedure of RF is as follows. First, a system obtains initial search results for a given query, and presents them to the user. The user then annotates some of the documents in the search results as being relevant or not, and the system modifies the search results using this feedback.

There are a variety of RF methods that depend on different retrieval models. Rocchio's algorithm (Rocchio, 1971) and the Ide dec-hi method (Ide, 1971) are well-known RF methods for the vector space model (Salton et al., 1975). In the probabilistic model (Spärck Jones et al., 2000), the weight of terms can be modified by feedback. For language modeling approaches (Ponte and Croft, 1998), Zhai and Lafferty (2001) proposed a fundamental RF method.

As described above, many methods have been proposed for RF. However, most of the previous methods use only the surface information in texts. That is, they ignore the latent information in texts, which could assist in improving IR performance. For example, they do not and cannot use the information of words for RF that do not appear in user feedback even if these words are highly probable from the latent topics of the feedback.

In this paper, we explore a novel RF method for language modeling approaches. In the proposed method, we use not only the surface information in texts, but also the latent information contained therein. More specifically, we infer the latent topic distribution in user feedback and in each document in the search results using latent Dirichlet allocation (LDA), and then we modify the search results so that documents with a similar topic distribution to that of the feedback are re-ranked higher. Evaluation results show that our method is effective for both explicit and pseudo RF, and that it has the advantage of performing well even when only a small amount of user feedback is available.

2 Language Modeling Approaches to IR

In this section, we describe the language modeling approaches to IR that form the basis of our method.

2.1 Overview

Language modeling approaches can be classified into three types: the query likelihood model

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 1037–1045,
Chiang Mai, Thailand, November 8 – 13, 2011.

(Ponte and Croft, 1998), the document likelihood model (Lavrenko and Croft, 2001), and the Kullback-Leibler (KL) divergence retrieval model (Lafferty and Zhai, 2001). In the query likelihood model, a document language model is constructed for each document in the collection. When a query is submitted by a user, the query likelihood is computed using the document model for each document. Then, the documents in the collection are ranked according to their likelihoods. In the document likelihood model, a query language model is constructed for a given query, and this is then used to compute the document likelihood for each document in the collection. The documents are then ranked by their likelihoods. In the KL-divergence retrieval model, both a query model and a document model are constructed, and the documents in the collection are ranked according to the KL-divergence between these models.

2.2 Language Model Construction

There are several ways of constructing a query model and a document model. One method is maximum likelihood estimation (MLE). The MLE of a word w_j with respect to a text $\boldsymbol{t}$ (e.g., query, document) is computed as

$$P_{\boldsymbol{t}}^{MLE}(w_j) = \frac{tf(w_j, \boldsymbol{t})}{|\boldsymbol{t}|}, \tag{1}$$

where $tf(w_j, \boldsymbol{t})$ represents the frequency of w_j in $\boldsymbol{t}$.

Dirichlet smoothed estimation (DIR) (Zhai and Lafferty, 2004) is also a well-known construction method. The DIR of w_j with respect to $\boldsymbol{t}$ is computed as follows.

$$P_{\boldsymbol{t}}^{DIR}(w_j) = \frac{tf(w_j, \boldsymbol{t}) + \mu P_{\boldsymbol{D}_{all}}^{MLE}(w_j)}{|\boldsymbol{t}| + \mu} \tag{2}$$

where $\boldsymbol{D}_{all}$ represents a collection, and μ represents the smoothing parameter that controls the degree of confidence in the frequency in $\boldsymbol{D}_{all}$ rather than in the frequency in $\boldsymbol{t}$.

2.3 RF for Language Modeling Approaches

Zhai and Lafferty proposed a fundamental RF method for the language modeling approaches (Zhai and Lafferty, 2001). When user feedback is given, they construct a language model for the feedback. Then, a new query model is constructed by interpolating the feedback model with the original query model, which is used to obtain the initial search results. Finally, they modify the search results using the new query model. They show the effectiveness of their method through their experiments, and report that the performance is better than that of Rocchio's algorithm.

3 LDA

In this section, we explain LDA, which is employed in the proposed method.

3.1 Overview

LDA (Blei et al., 2003) is one of the most popular topic models, and is viewed as an Bayesian extension of Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999). In PLSI, it is assumed that each document has a unique topic proportion $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. In contrast, LDA posits that $\boldsymbol{\theta}$ can take any values in the $(K-1)$ simplex, a topic proportion that means a point of the simplex is drawn from a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$. Note that the parameter $\boldsymbol{\alpha}$ is a K-vector with components $\alpha_k > 0$. In LDA, the probability of a document $\boldsymbol{d}_i$ in the training data is calculated as follows.

$$\begin{aligned} &P(\boldsymbol{d}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= \int P(\boldsymbol{\theta}|\boldsymbol{\alpha}) \left(\prod_{j=1}^{J} \left(\sum_{k=1}^{K} P(w_j|z_k, \boldsymbol{\beta})\, P(z_k|\boldsymbol{\theta}) \right)^{tf(w_j, \boldsymbol{d}_i)} \right) d\boldsymbol{\theta} \end{aligned}$$

where $z_k (k = 1, \ldots, K)$ represents a topic, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ represents the distributions over words for each topic z_k.

3.2 Parameter Estimation

In LDA, the expectation-maximization algorithm cannot be used to estimate the parameters, since the computation of the posterior distribution of latent variables is intractable. Thus, a wide variety of techniques using the variational method and Gibbs sampling, have been proposed to estimate the parameters (Blei et al., 2003; Griffiths and Steyvers, 2004). Here, we explain the technique using the variational method, as this is employed in the proposed method.

First, variational parameters $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{iK})$ and $\boldsymbol{\phi}_i = (\boldsymbol{\phi}_{i1}, \ldots, \boldsymbol{\phi}_{iJ})$ are introduced for each document $\boldsymbol{d}_i$ in the training data. Then, the optimal values of these are found by repeatedly computing the following pair of

update equations:

$$\phi_{ijk} \propto \beta_{kj} \exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{k'=1}^{K} \gamma_{k'}\right)\right) \quad (3)$$

$$\gamma_{ik} = \alpha_k + \sum_{j=1}^{J} \phi_{ijk}\, tf(w_j, \boldsymbol{d}_i) \quad (4)$$

where Ψ is the first derivative of the log Γ function.

Next, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are updated using $\boldsymbol{\gamma}_i$ and $\boldsymbol{\phi}_i$ for each $\boldsymbol{d}_i$. In the original paper, a Newton-Raphson method was used to estimate $\boldsymbol{\alpha}$ (Blei et al., 2003). However, estimation with the Newton-Raphson method has the disadvantages that it takes a long time and each estimated α_k can be a negative value under certain circumstances. It is known that the fixed-point iteration method (Minka, 2000) is a better estimation technique, and hence, we present the update equations based on this method. The update equations for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are given below.

$$\beta_{kj} \propto \sum_{i=1}^{I} \phi_{ijk}\, tf(w_j, \boldsymbol{d}_i)$$

$$\alpha_k = \frac{\sum_{i=1}^{I}\{\Psi(\alpha_k + n_{ik}) - \Psi(\alpha_k)\}}{\sum_{i=1}^{I}\{\Psi(\alpha_0 + |\boldsymbol{d}_i|) - \Psi(\alpha_0)\}} \alpha_k^{old}$$

where $n_{ik} = \sum_{j=1}^{J} \phi_{ijk}\, tf(w_j, \boldsymbol{d}_i)$, $\alpha_0 = \sum_{k'=1}^{K} \alpha_{k'}$, and α_k^{old} represents α_k before the update.

Finally, the updates of $\boldsymbol{\gamma}_i$ and $\boldsymbol{\phi}_i$ for each $\boldsymbol{d}_i$ and those of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are iterated until convergence. Once all the parameters have been estimated, we can obtain the probability of a word w_j given a document $\boldsymbol{d}_i$ as

$$P_{\boldsymbol{d}_i}^{LDA}(w_j) \simeq \frac{\sum_{k=1}^{K} \beta_{kj}\gamma_{ik}}{\sum_{k=1}^{K} \gamma_{ik}}. \quad (5)$$

3.3 Inference of Unseen Texts

One major advantage of LDA over PLSI is that it has a natural way of inferring the probabilities of unseen texts, which are not included in the training data. When we compute the probabilities of an unseen text $\boldsymbol{t}$, the variational parameters $\boldsymbol{\gamma_t}$ and $\boldsymbol{\phi_t}$ are estimated using Eqs.(3) and (4). Then, for example, the probabilities of words given $\boldsymbol{t}$ can be obtained using Eq.(5).

3.4 LDA in IR

Certain works using LDA for IR are closely related to our work. Wei and Croft (2006) incorporate LDA into a query likelihood model, while Zhou and Wade's work can be viewed as a study that incorporates LDA into a KL-divergence retrieval model (Zhou and Wade, 2009). Such works successfully utilize the latent information in texts through LDA, and report that the latent information is effective for ad-hoc retrieval. Although there are many differences between our work and those mentioned above, one of the biggest differences is that whereas the other works explored the effectiveness of the latent information for ad-hoc retrieval, we explore it for RF beyond ad-hoc retrieval.

4 Proposed Method

4.1 Overview

An overview of the proposed method is illustrated in Figure 1. First, when a query is submitted by a user, we obtain the initial search results (Step 1). Next, for each document in the search results, we construct a hybrid language model that contains not only the surface information, but also the latent information in the document (Step 2). Then, when user feedback is given, we also construct a hybrid language model for it (Step 3). Finally, we construct a new query model by interpolating the original query model with the feedback model. We also re-rank the initial search results using this new model so that documents with a similar topic distribution to that of the user feedback are re-ranked higher (Step 4). In the following subsections, we describe each step in detail.

4.2 Acquisition of Initial Search Results

In the proposed method, we employ a KL-divergence retrieval model (Lafferty and Zhai, 2001) to obtain the initial search results for a given query. First, we construct the MLE-based query model $P_{\boldsymbol{q}}^{MLE}(\cdot)$ for a query $\boldsymbol{q}$ using Eq.(1). Then, for each document containing $\boldsymbol{q}$ in the collection, the KL-divergence between the DIR-based document model and the MLE-based query model is computed. That is, the score of a document $\boldsymbol{d}$ for a query $\boldsymbol{q}$ is defined as follows.

$$initial\_score(\boldsymbol{d}, \boldsymbol{q}) = -KL(P_{\boldsymbol{q}}^{MLE}(\cdot)||P_{\boldsymbol{d}}^{DIR}(\cdot))$$

Finally, the initial search results $\boldsymbol{D_q} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{|\boldsymbol{D_q}|})$ are obtained by ranking the documents according to their scores.

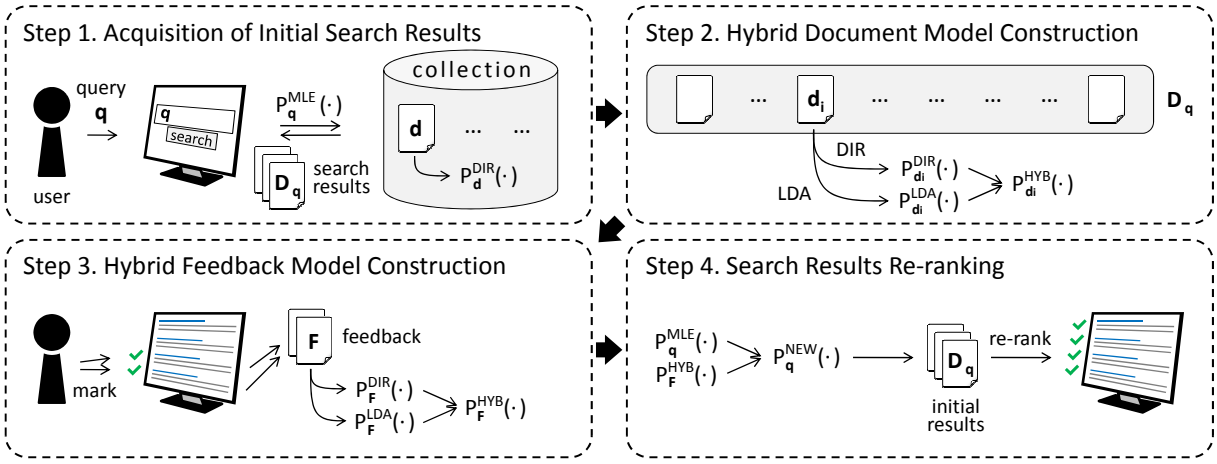

Figure 1: Overview of the proposed method.

4.3 Hybrid Document Model Construction

For each document in the search results, we construct a hybrid language model, which we call the HYB-based language model. In this model, we take into account the latent information in documents as well as the surface information.

First, an LDA-based document model that contains the latent information in the document is constructed for each document. We perform LDA on $\boldsymbol{D_q}$ to infer the topic distribution in each document $\boldsymbol{d}_i$, and estimate the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_i$ for each $\boldsymbol{d}_i$ as described in Section 3.2. Then, the LDA-based document model is constructed by computing the probabilities of words given $\boldsymbol{d}_i$ using Eq.(5). In this model, we can allocate high probability to words that are highly probable from the latent topic distribution of a document.

Next, for each $\boldsymbol{d}_i$, we construct an HYB-based document model $P^{HYB}_{\boldsymbol{d}_i}(\cdot)$ by interpolating the DIR-based document model with the LDA-based document model as follows.

$$P^{HYB}_{\boldsymbol{d}_i}(w_j) = (1-a)P^{DIR}_{\boldsymbol{d}_i}(w_j) + aP^{LDA}_{\boldsymbol{d}_i}(w_j)$$

where a is a parameter that controls the reliability of the LDA-based document model.

This interpolation is motivated by significant improvements reported in (Wei and Croft, 2006). They also interpolate a DIR-based document model with an LDA-based document model, and perform LDA on the whole collection to construct the LDA-based document model. However, executing LDA throughout the whole collection requires high computational cost. In the proposed method, we can avoid this problem by performing LDA only on the set of search results, which is much smaller than the whole collection.

4.4 Hybrid Feedback Model Construction

When feedback is given, we also construct an HYB-based language model for it. First, we obtain feedback $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{|\boldsymbol{F}|})$ that is relevant to the user's information need. Note that, in this study, we are not concerned with whether $\boldsymbol{F}$ is explicit, implicit, or pseudo feedback. Moreover, we have no preference of whether each $\boldsymbol{f}_i$ is a whole document or part of a document (e.g., title, snippet).

Next, we perform LDA on $\boldsymbol{F}$ to infer the topic distribution in $\boldsymbol{F}$, and construct the LDA-based feedback model $P^{LDA}_{\boldsymbol{F}}(\cdot)$. To be more precise, we generate a virtual relevant document $\boldsymbol{f}$ by combining each $\boldsymbol{f}_i$, and infer the variational parameter $\boldsymbol{\gamma}_{\boldsymbol{f}}$ as described in Section 3.3. Then, $P^{LDA}_{\boldsymbol{F}}(\cdot)$ is constructed using Eq.(5).

Finally, we construct the HYB-based feedback model $P^{HYB}_{\boldsymbol{F}}(\cdot)$, which contains the surface and latent information in $\boldsymbol{F}$. $P^{HYB}_{\boldsymbol{F}}(\cdot)$ is constructed in the same manner as $P^{HYB}_{\boldsymbol{d}_i}(\cdot)$. That is,

$$P^{HYB}_{\boldsymbol{F}}(w_j) = (1-a)P^{DIR}_{\boldsymbol{F}}(w_j) + aP^{LDA}_{\boldsymbol{F}}(w_j)$$

where $P^{DIR}_{\boldsymbol{F}}(\cdot)$ is constructed using Eq.(2).

4.5 Search Results Re-ranking

We construct a new query model, which is used to re-rank the initial search results. The new query model $P^{NEW}_{\boldsymbol{q}}(\cdot)$ is constructed by interpolating the original query model $P^{MLE}_{\boldsymbol{q}}(\cdot)$ with the hybrid feedback model $P^{HYB}_{\boldsymbol{F}}(\cdot)$ as follows.

$$P^{NEW}_{\boldsymbol{q}}(w_j) = (1-b)P^{MLE}_{\boldsymbol{q}}(w_j) + bP^{HYB}_{\boldsymbol{F}}(w_j)$$

where b is a parameter that controls the reliability of the feedback model. This interpolation is based

on Zhai and Lafferty's linear combination method (see Section 2.3).

Then, for each document $\boldsymbol{d}_i$ in the search results $\boldsymbol{D_q}$, we compute the KL-divergence between $P^{HYB}_{\boldsymbol{d}_i}(\cdot)$ and $P^{NEW}_{\boldsymbol{q}}(\cdot)$. That is, the score of document $\boldsymbol{d}_i$ for query $\boldsymbol{q}$ and feedback $\boldsymbol{F}$ is defined as

$$re\text{-}ranking\_score(\boldsymbol{d}_i, \boldsymbol{q}, \boldsymbol{F}) = -KL(P^{NEW}_{\boldsymbol{q}}(\cdot)||P^{HYB}_{\boldsymbol{d}_i}(\cdot)).$$

Finally, we obtain the revised search results by re-ranking the documents in $\boldsymbol{D_q}$ according to their re-ranking scores.

5 Experiments

5.1 Overview

We conducted three experiments to evaluate the performance of our method. An overview of each experiment is given below.

Experiment 1. Effectiveness with Respect to Explicit and Pseudo RF

We examined how well our method performed in re-ranking the initial search results using explicit and pseudo feedback. In the experiment with explicit RF, we obtained the top 100 documents with the highest initial scores as the initial search results for a given query, and re-ranked them using our method with two relevant documents that were given explicitly. In our experiments, we employed the queries and the relevant documents provided by NTCIR (see Section 5.3). Then, we compared the results with the following three (re-)ranking results.

INIT This is the ranking of the initial search results obtained based on the KL-divergence retrieval model.

WORD This is the ranking of the search results obtained after simple RF, where we used only the surface information (i.e., words) in the feedback and the documents in the search results. This ranking is equivalent to the ranking obtained using our method with $a = 0$.

REPR This is the re-ranking result obtained using Zhai and Lafferty's RF method (Zhai and Lafferty, 2001). The process of the method is almost the same to that of WORD. The main difference is that it modifies the probabilities of words in feedback by a background word distribution (see their paper). We chose their method as it is a representative RF method for language modeling approaches.

Our method can also be applied to pseudo RF. Hence, we also explored the effectiveness of our method in this regard. With pseudo RF, the top n documents in the initial search results are assumed to be relevant, and the search results are re-ranked based on this assumption. We implemented pseudo RF using our method for $n = 10$, and compared the results with the three (re-)ranking results described above.

Experiment 2. Effect of the Amount of Feedback

It is important to know how well our method performs when only a small amount of feedback is obtained, because in practice users generally cannot be bothered to provide feedback, and thus sufficient feedback is rarely obtained. We investigated how the amount of explicit feedback affected the performance of our method. To be more precise, we reduced the amount of available explicit feedback little by little, and observed the change in precision at 10 top re-ranked documents (P@10). For this experiment, we used seven different amounts of explicit feedback: 2^1, 2^0, 2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, and 2^{-5} relevant documents. Note that, for example, 2^{-1} documents means that we used half a document's worth of words in the relevant documents given as explicit feedback. In this case, half the words were sampled randomly from the feedback, and only these words were used for RF.

Experiment 3. Sensitivity to Parameters

It is also important to know how the reliability of the LDA-based document model and the HYB-based feedback model affect the performance of our method. Hence, we investigated how sensitive our method is to parameters a and b. We re-ranked the initial search results using different values for these parameters ranging from 0 to 1 in steps of 0.1, and measured how the performance of our method changed according to these values.

5.2 Configuration of Our Method

The configuration of our method is given below. For the DIR estimation, we set the smoothing parameter $\mu = 1,000$. This setting was also employed in other works (Zhai and Lafferty, 2001; Wei and Croft, 2006). The number of topics K for LDA was set to 20, since with this setting we obtained better results in the preliminary experiments, in which we performed LDA with K rang-

ing from 10 to 100 in steps of 10. We set the initial values of $\alpha_k (k = 1, \ldots, K)$ to 1, and the initial values of $P(w_j|z_k, \boldsymbol{\beta})$ to random values. The number of iterations for the variational parameters and that for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ were set to 10. Additionally, we limited the size of the vocabulary in LDA, designated as J in Section 3, to $1,000$. We selected $1,000$ words based on their importance to the search results. Note that the importance of a word w_j to the search results $\boldsymbol{D_q}$ is defined as $df(w_j, \boldsymbol{D_q}) * \log(|\boldsymbol{D}_{all}|/df(w_j, \boldsymbol{D}_{all}))$, where $df(w_j, \boldsymbol{D})$ represents the document frequency of w_j in documents $\boldsymbol{D}$.

5.3 Data Set

In our experiments, we employed the test collection used in the Web Retrieval Task in the Third NTCIR Workshop (Eguchi et al., 2002). The NTCIR Workshops are a series of evaluation workshops designed to enhance research in information access technologies. The test collection consists of $11,038,720$ Japanese Web pages and 47 information needs. For each information need, about $2,000$ documents are rated as highly relevant, fairly relevant, partially relevant, or irrelevant. We used only 40 information needs in our experiments. The remaining 7 (with identification numbers: 0011, 0018, 0032, 0040, 0044, 0047, and 0061) were not used, because we could not retrieve 100 documents for each (see Section 5.1).

Figure 2 gives an example of an information need for the Web Retrieval Task in the Third NTCIR Workshop. The meaning of each element is given below.

⟨NUM⟩ gives the identification number of the information need.

⟨TITLE⟩ provides up to three terms that are similar to the actual query submitted to a real search engine.

⟨DESC⟩ describes the user's information need in a single sentence.

⟨RDOC⟩ provides up to three identification numbers of examples of relevant documents for the information need.

We employed the terms in the ⟨TITLE⟩ tag as the query, and the documents in the ⟨RDOC⟩ tag as explicit feedback. Note that since the numbers of terms and documents differed depending on the information need, we employed the first two terms in the ⟨TITLE⟩ tag and the first two documents in the ⟨RDOC⟩ tag for each information need.

5.4 Evaluation Method

We used P@10, mean average precision (MAP), normalized discounted cumulative gain at 10 top (re-)ranked documents (NDCG@10), and NDCG@100 (Järvelin and Kekäläinen, 2002) in the evaluation. In the calculation of P@10 and MAP, documents that were rated as highly relevant, fairly relevant and partially relevant were regarded as relevant, while documents rated as irrelevant and unrated documents were regarded as irrelevant. Note that MAP was calculated using all the (re-)ranked documents (i.e., 100 documents). In calculating NDCG, we assessed the relevance score of documents rated highly relevant, fairly relevant and partially relevant as 3, 2, and 1 respectively.

To evaluate the effectiveness of explicit RF, we decided in advance which documents would be used as explicit feedback as described in Section 5.3, and if these were included in the initial search results and the re-ranked results, we removed them from both sets of results. One common problem in the evaluation of the effectiveness of explicit RF is how to handle documents that users have marked as relevant (i.e., the input to RF methods) (Hull, 1993). If the initial search results and the re-ranked results are compared in a straightforward manner, the latter have an advantage. This is due to the fact that documents that are known to be relevant tend to be re-ranked higher. However, if we remove them from the re-ranked results, they have a disadvantage. This is especially true if there are few relevant documents. Therefore, we removed the documents used as explicit feedback from both the initial search results and the re-ranked results in the experiments with explicit RF.

In contrast, we did not apply extra care in Experiment 1 with pseudo RF, and measured the performance of each method using its raw (re-)ranked results.

5.5 Experimental Results

Experiment 1. Effectiveness with Respect to Explicit and Pseudo RF

Table 1 gives the results for explicit RF. Owing to space limitations, we only show the optimal results in terms of P@10 for each method. The optimal results for WORD were obtained using our

⟨NUM⟩ 0008 ⟨/NUM⟩
⟨TITLE⟩ Salsa, learn, methods ⟨/TITLE⟩
⟨DESC⟩ I want to find out about methods for learning how to dance the salsa ⟨/DESC⟩
⟨RDOC⟩ NW011992774, NW011992731, NW011992734 ⟨/RDOC⟩

Figure 2: Example of an information need for a Web Retrieval Task in the Third NTCIR Workshop.

Table 1: Effectiveness with respect to explicit RF.

	P@10	MAP	NDCG@10	NDCG@100
INIT	0.278	0.106	0.220	0.249
WORD	0.310	0.111	0.228	0.250
REPR	0.303	0.107	0.236	0.249
OURS	0.383	0.117	0.284	0.255

Table 2: Effectiveness with respect to pseudo RF.

	P@10	MAP	NDCG@10	NDCG@100
INIT	0.298	0.112	0.243	0.268
WORD	0.303	0.111	0.258	0.274
REPR	0.300	0.112	0.250	0.270
OURS	0.330	0.112	0.283	0.278

method with $a = 0$ and $b = 0.7$, while those for OURS (our method) were obtained with $a = 0.2$ and $b = 0.7$.

Based on this table, we can confirm that our method is effective with respect to explicit RF. Our method significantly improved the initial search results across all metrics. Additionally, it outperformed two other baseline RF methods, with the differences in all metrics being statistically significant (Wilcoxon signed-rank test, $p < 0.05$). There were no significant differences between the baseline methods, since they were similar in process to each other. These results suggest that the latent information in the user feedback and each document in the search results is useful for explicit RF.

As a result of the investigation, we found that our method made good use of the words that did not appear in the feedback but were highly probable from the latent topic distribution of the feedback. Consider the information need in Figure 2 as an example. The documents employed as user feedback did not contain the words "technique" or "level", which are related to the information need. As such, the baseline methods could not use these words. In contrast, our method allocated a certain degree of probability to these highly probable words using LDA, despite the words not appearing in the feedback, and hence raised the score of relevant documents in the search results containing these words.

Table 2 gives the results for pseudo RF. The values of the parameters for WORD and OURS were determined as: $a = 0, b = 0.7$, and $a = 0.1, b = 0.6$, respectively. Note that the results of INIT in Table 2 differ from those in Table 1. This is because although we removed the documents used as user feedback from the initial search results (and the re-ranked results) in the experiment with explicit RF, we did not remove them from any of the results in this experiment.

From this table, we can see that our method significantly improved the initial search results. Additionally, our method outperformed the baseline methods. From these results, we can conclude that our method is also effective with respect to pseudo RF.

Experiment 2. Effect of the Amount of Feedback

Figure 3 shows the effect of the amount of explicit feedback on the performance of our method. For comparison with baseline methods, we also present their results. The parameters for each method were identical to those used in Experiment 1 with explicit RF.

From this figure, we can see that our method achieved consistently high performance. For example, when 2^0 relevant documents (i.e., one relevant document) were given as user feedback, our method improved the initial search results by about 35% in P@10. Additionally, a notable feature is that although the improvements in the baseline methods almost disappeared, our method performed well when only a small amount of feedback was obtained. For example, improvement of about 18% was achieved even with only 2^{-5} documents, which constituted an average of 52 words in our experiment. The reason for this is, once again, that our method is able to use not only the surface words in the feedback, but also the highly probable words from its latent topic distribution.

As described above, our method can re-rank search results using a small amount of feedback. This suggests that our method is practically useful, and that it performs well even if only a part of a document (e.g., title, snippet), the relevance of which is easier to determine than that of the whole document, are given as user feedback.

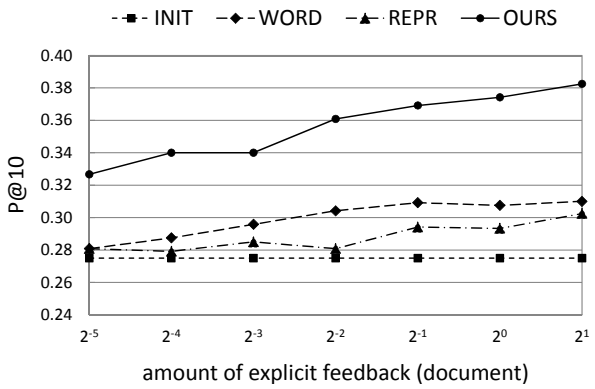

Figure 3: Effect of the amount of feedback.

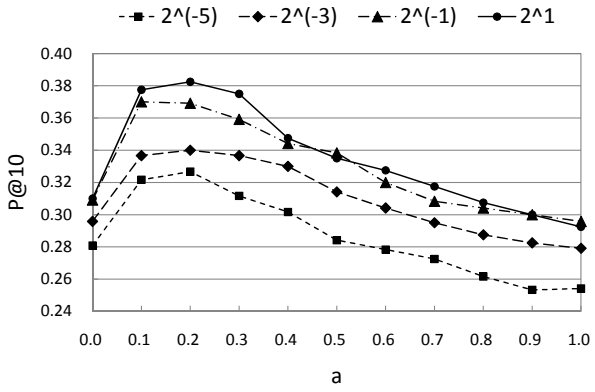

Figure 4: Sensitivity to parameter a.

Experiment 3. Sensitivity to Parameters

The sensitivity of the performance to a is illustrated in Figure 4. Each line in the figure represents the results with different amounts of explicit feedback: 2^{-5}, 2^{-3}, 2^{-1}, and 2^{1} relevant documents. The value of the other parameter b was fixed to 0.7. From this figure, we can see that the performance of our method is sensitive to the value of a, and that the optimal value is about 0.2. Despite the goal and setting being different to ours, this optimal value is similar to that reported in (Wei and Croft, 2006), where the DIR- and LDA-based document models were interpolated as in our work.

The sensitivity to b is depicted in Figure 5. The value of a was fixed to 0.2. According to this figure, we can see that the setting of b also affected the performance of our method. Additionally, this figure shows that if we set the value appropriately, the interpolated new query model is more effective than both the original query model on its own (i.e., $b = 0.0$) and the feedback model on its own (i.e., $b = 1.0$). These findings concur approximately with the results presented in (Zhai and Lafferty, 2001).

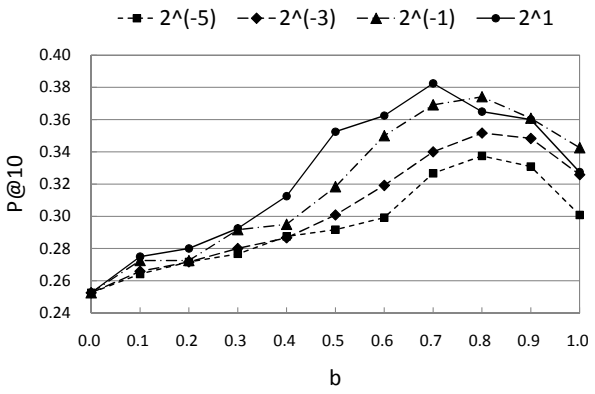

Figure 5: Sensitivity to parameter b.

5.6 Discussion

Although our method achieved good performance in our experiments, we also encountered a problem in that the method took a long time to execute. More specifically, our method required about one minute to estimate the parameters of LDA in Step 2. (On the other hand, the time required for Steps 1, 3, and 4 was only a few seconds.) Thus, we need to explore ways of reducing the time for parameter estimation so that our method can be used in real situations. One way of doing this is to choose a faster estimation technique. For example, collapsed variational methods may provide a viable solution (Teh et al., 2006; Asuncion et al., 2009). Another alternative is to decrease the size of the vocabulary, designated as J in Section 3. For example, we conducted an additional experiment, in which we set $J = 100$, and confirmed that the time for parameter estimation fell to about 10 seconds without any significant change in performance.

6 Conclusion

In this paper, we proposed a novel RF method using latent information, and discussed the effectiveness thereof. Using LDA, our method infers the distributions over latent topics in the feedback and in each document in the search results. Then, documents whose topic distribution resembles that of the feedback are regarded as being relevant to the user's information need, and are re-ranked higher. Through our experiments, we confirmed that our method achieves good performance for both explicit and pseudo RF, and that it provides the benefit of performing well even when only a small amount of feedback can be obtained. As future work, we aim to explore ways of reducing the execution time of our method so that it can be used in practical situations.

References

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. 2009. On smoothing and inference for topic models. In *Proceedings of UAI 2009*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Jounal of Machine Learning Research*, 3:993–1022.

Koji Eguchi, Keizo Oyama, Emi Ishida, Kazuko Kuriyama, and Noriko Kando. 2002. The web retrieval task and its evaluation in the third ntcir workshop. In *Proceedings of SIGIR 2002*, pages 375–376.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of NAS 2004*, pages 5228–5235.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR 1999*, pages 50–57.

David Hull. 1993. Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of SIGIR 1993*, pages 329–338.

Eleanor Ide. 1971. New experiments in relevance feedback. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 337–354. Prentice-Hall Inc.

Bernard James Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management*, 36(2):207–227.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

John Lafferty and Chengxiang Zhai. 2001. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR 2001*, pages 111–119.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *Proceedings of SIGIR 2001*, pages 120–127.

Thomas P. Minka. 2000. Estimating a dirichlet distribution. Technical report, Microsoft.

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proceedings of SIGIR 1998*, pages 275–281.

Joseph John Rocchio. 1971. Relevance feedback in information retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall Inc.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Karen Spärck Jones, S. Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments. *Information Processing and Management*, 36(6):779–808,809–840.

Yee Whye Teh, David Newman, and Max Welling. 2006. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Proceedings of NIPS 2006*.

Xing Wei and W.Bruce Croft. 2006. Lda-based document models for ad-hoc retrieval. In *Proceedings of SIGIR 2006*, pages 178–185.

Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM 2001*, pages 403–410.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214.

Dong Zhou and Vincent Wade. 2009. Latent document re-ranking. In *Proceedings of EMNLP 2009*, pages 1571–1580.