Improving Relevance Feedback in Language Modeling Approach: Maximum a Posteriori Probability Criterion and Three-Component Mixture Model

Seung-Hoon Na, In-Su Kang, and Jong-Hyeok Lee

Div. of Electrical and Computer Engineering,
Pohang University of Science and Technology (POSTECH),
Advanced Information Technology Research Center (AITrc)
{nsh1979,dbaisk,jhlee}@postech.ac.kr

Abstract. Recently, researchers have tried to extend a language modeling approach to apply relevance feedback. Their approaches can be classified into two categories. One typical approach is the expansion-based feedback that sequentially performs 'term selection' and 'term re-weighting' separately. Another approach is the model-based feedback that focuses on estimating 'query language model', which predicts well users' information need. This paper improves these two approaches of relevance feedback by using *a maximum a posteriori probability criterion*, and *a three-component mixture model. A maximum a posteriori probability criterion* is a criterion for selection of good expansion terms from feedback documents. *A three-component mixture model* is the method that eliminates the noise of the query language model by adding a 'document specific topic model'. The experimental results show that our methods increase the precision of relevance feedback for a short length query. In addition, we make some comparative study between several relevance feedbacks in three document collections.

1 Introduction

The basic idea of the language modeling approach to information retrieval, first introduced by [5], is not to explicitly assume relevance information, but to assume individual document models for each document and estimate them. With these document models, documents are ranked by query likelihood where the document models will generate a given query. In spite of its mathematical simplicity, language modeling approaches have shown to perform well empirically showing comparative performance to classical probabilistic models.

The language modeling approach has had difficulty with handling relevance feedback within a well-founded framework. Some researchers have tried to incorporate relevance feedback into the language modeling approach in a principled fashion. Their approaches can be classified into two categories – expansion-based and model-based.

Expansion-based feedback is similar to the typical classical approach that sequentially performs 'term selection' and 'term re-weighting', [1], [5]. At the 'term selection', new query terms are selected from feedback documents, and at the 'term re-

K.-Y. Su et al. (Eds.): IJCNLP 2004, LNAI 3248, pp. 130–138, 2005.
© Springer-Verlag Berlin Heidelberg 2005

weighting', such selected terms are re-weighted according to its significance[1]. In this approach term selection criterion is very important, but it has only been dealt with heuristically so it is not naturally applied to term dependent or more general situations. To this end, we propose maximum a posteriori probability criterion that is more intuitively motivated with a principle fashion and provide tractable methods in more generalized situations.

Another alternative approach is model-based feedback that deals with the problem of estimating a 'query language model' that represents the user's information need [3], [4], [8]. The new expansion query is sampled from this estimated the query language model. Thus, the problem of estimation of the query language model is very important in this approach.

To estimate the query language model, Zhai and Jefferty [8], who organize model-based feedback, assumed that all terms in feedback documents are generated from the two-component mixture model that consists of the query language model and a background collection model. Unfortunately, this assumption have a problem when the feedback documents contain other topics as well as query-relevant topics. If we agree that most documents have multiple topics as well as query-relevant, two-component mixture model may cause the estimated query language model to have some noise. Thus, we use the 'three-component mixture model' that consists of a query language model and a background language model and a 'document specific topic model'. The three-component mixture model will estimate the query language model more correctly.

The remainder of the paper is organized as follows. In Section 2 we review background and previous relevance feedback methods in language modeling. Section 3 and Section 4 describes the maximum a posteriori probability criterion and the three-component mixture model, respectively. Section 5 shows experimental results of the new methods and previous relevance feedback methods. Finally, we offer conclusion and present our research direction.

2 Relevance Feedback Approaches in Language Modeling

The basic idea of language modeling ranks documents in the collection with the query-likelihood that a given query **q** would be observed during repeated random sampling from each document model [1], [5][2].

$$P(\mathbf{q} \mid \theta_D) = \prod_w P(w \mid \theta_D)^{c(w;\mathbf{q})} \tag{1}$$

where $c(w;\mathbf{q})$ is the number of term in given query, D is a given document.

Next, the retrieval problem is reduced to the problem of estimating a unigram language model $P(w|\theta_D)$.

---

[1] In this paper, we assume the viewpoint that term re-weighting in expansion-based approach, more elaborate document models are re-constructed by optimizing unknown smoothing parameters [1].

[2] There is some difference between authors about interpretation of a query. [5] treats a query as a set, while [1] interpreted a query as a sequence of words. In this paper, we adopt the sequence interpretation.

As mentioned in the section 1, previous relevance feedback methods in the language modeling approach have been explored by two distinct approaches: expansion-based, model-based. In the remainder of the section, we will give details on each approach.

2.1 Expansion-Based Feedback

Ponte's method [6], the first heuristic work in expansion-based feedback, used the ratio method that select terms having a high generative probability on top retrieved document models, but a low generative probability on the collection language model.

$$LR(w) = \sum_{D \in \mathcal{R}} \log \frac{P(w \mid \theta_D)}{P(w \mid \theta_C)} \tag{2}$$

Where $\mathcal{R}$ is set of feedback documents, $P(w|\theta_D)$ is the probability of term w given the document model θ_D for D. The ratio method performs in practice well, but is not based on the well-founded framework. The development of a well-designed framework for the term selection (expansion) is one an important issues in this approach.

Zhai and Lafferty [7] mentioned on the 'query modeling role' of smoothing that distinguish the common and non-informative terms in a query different from the 'estimation role'. Hiemstra [1] proposed the term specific smoothing that conceptually connects the query modeling role of smoothing into re-weighting of classical probability model. In term specific smoothing, the importance of query term can be quantified with a separate smoothing parameter λ_i for each term.

2.2 Model-Based Feedback

Zhai and Lafferty [8] introduced a model-based feedback within KL divergence, motivated by Ng's work [4] that generalizes the most language modeling approaches and makes the feedback problem more tractable. Model-based feedback does not re-estimate smoothing parameters, but instead estimates the 'query language model' that is the probability distribution for the number of each terms in a new query to obtain high retrieval performance. The documents are ranked with inverse proportion to KL-divergence between the query language model and the given document language model.

$$D(\theta_Q \,||\, \theta_D) = -\sum_{w} p(w \mid \theta_Q) \log p(w \mid \theta_D) + C(\theta_Q) \tag{3}$$

where $P(w|\theta_Q)$ is the query language model.

3 Maximum Posterior Probability Criterion

One term selection problem is to select 'good' terms for constructing a new query among terms occurring in feedback documents. In the language modeling approach, a query is a sample generated by a specific document model. Thus, the term selection problem is to find a query sample to predict the top document models in an entire

query sample space. Here, the query sample space is a set of all possible sequences of vocabulary[3].

To make this problem tractable, imagine that a single specific document model θ_D is given and we want to search the best query sample (with fixed length) to predict well this document model. A reasonable strategy to this problem is to maximize the a posteriori probability of the query sample of the document model.

$$\mathbf{q}^* = \arg\max_{\mathbf{q} \in Q} P(\theta_D \mid \mathbf{q}) \tag{4}$$

where Q is query sample space. **q*** is the best sample to predict this document model. We call this strategy by *maximum a posteriori criterion*.

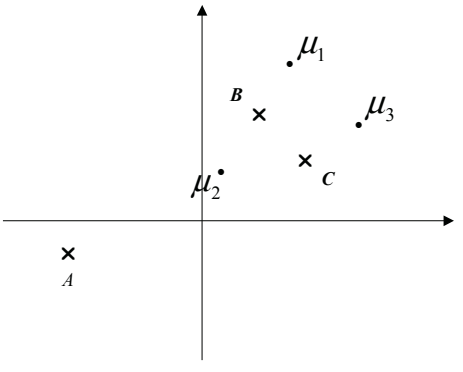

Fig. 1. Imaginary situation for Maximum posterior probability criterion

Figure 1 illustrates this maximum a posteriori criterion, where the problem is simplified to a problem in the two-dimensional real number space. Here, three Gaussian distribution $N(\mu_i, \sigma)$ with the mean μ_i and the variance σ respectively and a sample space { A, B, C } are given. Among the sample space, if we find a sample predicting well of the second Gaussian distribution $N(\mu_2, \sigma)$, we will select sample A, because it maximizes posterior probability on the second Gaussian model. Sample B and C give larger likelihoods than A on the second Gaussian, but they never good samples because their posterior probability for $N(\mu_2, \sigma)$ is not the maximum in this sample space. However, if we use the maximum a posteriori probability criterion, then sample A will be selected.

At this point, we can treat the query selection problem in relevance feedback. To this end, we extend the above maximum a posterior criterion into more generalized criterion that can be applied to the situation that multiple document models $\theta_{D_1}, \theta_{D_2}, \ldots, \theta_{D_m}$ are given. The term expansion problem is to find the best query sample that 'simultaneously' maximizes a posteriori probability for each document model. Clearly, it forces us to combine all posteriori probabilities on each document model with '*and event*'. Thus,

[3] In our term selection problem, the query sample space Q is restricted with query size and uniqueness of each query term.

$$\mathbf{q}^* = \arg\max_{\mathbf{q}\in Q} \prod_{i=1}^{m} P(\theta_{D_i} \mid \mathbf{q}) \tag{5}$$

Also, if we assume term independence, then

$$\begin{aligned} \prod_i P(\theta_{D_i} \mid \mathbf{q}) &= \prod_i \frac{P(\mathbf{q} \mid \theta_{D_i}) P(\theta_{D_i})}{P(\mathbf{q})} \\ &= \prod_i \prod_{w\in \mathbf{q}} \frac{P(w \mid \theta_{D_i}) P(\theta_{D_i})}{P(w)} \end{aligned} \tag{6}$$

where $P(w) = \sum_{D\in\mathcal{M}} P(w \mid \theta_D) P(\theta_D)$, $P(\theta_D) = \frac{1}{|\mathcal{M}|}$, and $\mathcal{M}$ is the set of all existing document model in a given collection.

To maximize (6), we must select terms with rank ordered by the following individual term score.

$$score(w) = \prod_i \frac{P(w \mid \theta_{D_i}) P(\theta_D)}{P(w)} \tag{7}$$

If we assume uniform prior probability $P(\theta_D)$,. then,

$$score(w) \propto \sum_i \log \frac{P(w \mid \theta_{D_i})}{\frac{1}{|\mathcal{M}|} \sum_{D\in\mathcal{M}} P(w \mid \theta_D)} \tag{8}$$

Now, this formula (8) will be used in term sorting for term selection.

4 Three-Component Mixture Model

In model-based feedback, query language model estimation is important, because it play the two roles of re-weighting and term expansion. To estimate a query language model, Zhai and Lafferty [8] suggested the two-component mixture model with the unknown query language model and a background collection language model. For $D\in\mathcal{R}$,

$$P(w \mid D) = \lambda\ P(w \mid \theta_Q) + (1-\lambda)\ P(w \mid \theta_C) \tag{9}$$

However, this two-component mixture model can make the query language model include some irrelevant portions, because the feedback documents have multiple topics. It is difficult to catch this portion of a query language model by using only a collection background model. To build a more accurate query language model, it is necessary to revise this model to eliminate these irrelevant portions.

To this end, we add a single document specific model into the original two-component mixture model. As a result, we obtain the following three-component mixture model.

$$P(w \mid D) = \lambda_Q P(w \mid \theta_Q) + \lambda_S p(w \mid \theta_D^s) + \lambda_C p(w \mid \theta_C) \tag{10}$$

where $\lambda_Q+\lambda_S+\lambda_C=1$, and $p(w \mid \theta_D^s)$ is the non-relevant topic model of the document D.

4.1 Approximation to Document Specific Model

One ad-hoc method is a naive approximation that almost all terms in feedback documents are irrelevant to the query topic.

$$p(w \mid \theta_D^s) \approx p(w \mid \hat{\theta}_D) \tag{11}$$

Although this approximation may not inconsistent to the assumption of the relevant feedback, query language modeling can be estimated more carefully. The naïve approximation will bring only highly shared portions in feedback documents into the query language model.

Another alternative method is the mixture approximation using a topic collection language model.

$$p(w \mid \theta_D^s) \approx \pi p(w \mid \hat{\theta}_D) + (1-\pi) p(w \mid \theta_{C_w}) \tag{12}$$

where $p(w \mid \theta_{C_w})$ is a topic collection language model that is estimated from all documents which include term *w*.

4.2 Estimation of Query Language Model

To estimate query language model, we use the EM algorithm, which iteratively updates query language model θ_Q to maximize (locally) generative likelihood of feedback documents. Initially, $P(w \mid \theta_Q^{(0)})$ are set to

$$P(w \mid \theta_Q^{(0)}) = \frac{\sum_{D \in \mathcal{R}} c(w; D)}{\sum_{D \in \mathcal{R}} \sum_{w} c(w; D)} \tag{13}$$

where *c(w;D)* is the count of term *w* in document *D*.

Next, we perform the E-step and M-step iteratively.

E-step:

$$P(w \text{ is Rel} \mid D)^{(k)} = \frac{\lambda_Q p(w \mid \theta_Q)^{(k)}}{\lambda_Q p(w \mid \theta_Q)^{(k)} + \lambda_S p(w \mid \theta_D^s) + \lambda_C p(w \mid \theta_C)} \tag{14}$$

M-step:

$$p(w \mid \theta_Q)^{(k+1)} = \frac{\sum_{D \in \mathcal{R}} c(w; D) P(w \text{ is Rel} \mid D)^{(k)}}{\sum_{D \in \mathcal{R}} \sum_{w} c(w; D) P(w \text{ is Rel} \mid D)^{(k)}} \tag{15}$$

where λ_Q and λ_S and λ_C are constants.

5 Experimentation

All feedback methods described in this paper are evaluated in NTCIR3 test collections and topics.

1. KR: Korea Economic Daily (1994)
   66147 number of documents, 30 Topics
2. JA: Mainichi Newspaper (1998-1999)
   220078 number of documents, 42 Topics
3. CH: CIRB010, United Daily News (1998-1999)
   381682 number of documents, 42 Topics

In all experiments, we used four types of query provided by NTCIR 3 task: Title, Description, Concept, All (All: consists of all topic fields). Table 1 describes the average number of query terms for each collection. For indexing, we performed preliminary experimentations on NTCIR-3 test collections using various indexing methods (Morphlogy and word, bi-character). As a result, we found that bi-character indexing units are highly reliable for Korean or other Asian Languages. Our all experiments in this paper are performed using the bi-character indexing unit.

Table 1. The average length of query for each topic and collection

	Title	Desc	Conc	All
KR	5.5	19.4	13.8	109.9
JA	8.5	33.1	18.9	212.3
CH	6.3	20.3	14.3	143.3

Table 2. Relevance feedback methods evaluated in experimentation

Approach	Symbol	Method
Expansion-based	lr	Likelihood ratio method
	lr+	Likelihood ratio method and term specific smoothing
	mpp	Maximum a posteriori probability criterion
	mpp+	Maximum a posteriori probability criterion and term specific smoothing
Model-based	qc	Two-component mixture model
	qdc	Three-component mixture model using naïve approximation

5.1 Relevance Feedback Performance

Table 2 lists the relevance feedback methods evaluated in this experimentation. Three previous methods are marked with 'lr', 'lr+', 'qc'. Here, '+' indicates that the method used the term specific smoothing [1] for term reweighting. And, our methods are marked with 'mpp', 'mpp+','mpp*', 'qdc'.

In expansion-based approaches (lr, lr+, mpp, mpp+), top 15 retrieved documents used as feedback documents, and 50 new query terms are added into the original query. In model-based approaches (qc, qdc), the top 10 retrieved documents are used to estimate the query language model. Table 3 shows the final experimental results of relevance feedback methods in three collections. Free parameters of two-component mixture model and three-component mixture model are set by λ=0.25 and, λ_D=0.175, λ_S=0.075, λ_C=0.75, respectively. We selected parameters empirically that performed well in our preliminary experimentation.

As shown in Table 3, all relevance feedback methods improve significantly the performance of our initial retrieval results. 'mpp' and 'mpp+' showed a slight im-

provement over 'lr' and 'lr+' of [6] for title and description and concept, but they sometimes show a lower performance than 'lr' and 'lr+', although the ratio method is almost equal to maximum a posteriori probability criterion when using Jelinek smoothing. Also, the 'qdc' shows a notable improvement over the original approach 'qc', although the approximation method for 'qdc' is *adhoc* naïve approximation that may be dangerous. In KR, CH collection, it does not seem to significantly increase performance, but in JA collection, it improves average performance over the original method more than 1 percent. If we use better model to approximate the document specific model, then the performance can be improved.

Table 3. Average precision of relevance feedback methods for KR and JA collections (above) and CH collection (bottom). The evaluation measure is non-interpolate average precision. Underlined number and bold number indicate that the method improves against the previous method and that the method has best performance in the column of the table, respectively

	KR				JA			
	Title	Desc	Conc	All	Title	Desc	Conc	All
init	0.3090	0.2496	0.3277	0.4203	0.2975	0.2937	0.3169	0.4186
lr	0.3491	0.3524	0.3965	0.4797	0.3506	0.3719	0.3778	0.4615
mpp	0.3508	0.3538	0.4001	**0.4799**	0.3569	0.3746	0.3834	0.4583
lr+	0.3703	0.3603	0.4068	0.4671	0.3729	0.3909	0.3891	**0.4625**
mpp+	0.3721	0.3608	**0.4073**	0.4698	0.3797	0.3922	**0.3920**	0.4561
qc	0.3874	0.3716	0.3795	0.4561	0.3847	0.3835	0.3684	0.4374
qdc	**0.3914**	**0.3779**	0.3859	0.4610	**0.3882**	**0.3956**	0.3816	0.4251

	CH			
	Title	Desc	Conc	All
init	0.2392	0.1997	0.2516	0.3266
lr	0.2953	0.2882	0.3190	**0.3851**
mpp	0.2920	0.2913	0.3223	0.3846
lr+	0.2961	0.2796	0.3124	0.3662
mpp+	0.2914	0.2845	0.3149	0.3641
qc	0.3136	0.3224	0.3220	0.3679
qdc	**0.3155**	**0.3306**	**0.3256**	0.3701

One interesting result is the effect of re-weighting (marked with '+') according to query length. For a short length query like title or concept, re-weighting is highly effective, but for a long length query re-weighting degrades the performance. One possible explanation is that, as the given query is longer, the number of matching terms in the document gives more large impact on performance, while the effect of the term weighting is much weaker. Thus, to perform well for a long query, we may need a method that incorporates the coordination level matching into the current re-weighting scheme.

Another interesting effect is that model-based feedbacks show a superior performance than the expansion-based feedbacks when the initial retrieval performance is relatively low (less than 0.31). This means that the model-based feedback is more robust than the expansion based approach over the initial retrieval performance. This can be explained by the duplication of term weighting in the query language model. The default weight is the likelihood odd between the document language model and

the collection language model, and another additional weight is the term distribution from query language model. However, the default weight disappears in term specific re-weighting of the expansion-based approach.

6 Conclusion

In this paper, we proposed two methods to improve relevance feedback in the language modeling approach, and performed comparative experimentations between several relevance feedback methods, including our new methods.

Experimental results are summarized as following. 1) Relevance feedback significantly improved the performance of baseline retrieval results. 2) The new proposed methods in this paper increased the performance of previous methods a bit, but sometimes the improvement is significant. 3) For a short length query, the term specific smoothing had shown to improve significantly retrieval feedback. However, for a long query, it seems to decrease performance. 4) Model-based feedback is more robust over the initial retrieval performances against the expansion-based feedback. By contrast, when the initial retrieval performance had good, expansion-based feedback showed a better performance over than model-based feedback.

Acknowledgements

This work was supported by the KOSEF through the Advanced Information Technology Research Center(AITrc) and by the BK21 Project.

References

1. Hiemstra, D.: Term Specific Smoothing for Language Modeling Approach to Information Retrieval: The Importance of a Query Term. In Proceedings of 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2002)
2. Lafferty, J. and Zhai, C.: Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2001)
3. Lavrenko, V. and Croft, B.: Relevance-based language models. In Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2001)
4. Ng, K.: A Maximum Likelihood Ratio Information Retrieval Model. In TREC-8 Workshop Notebook (1999)
5. Ponte, A. and Croft, J.: A language modeling approach to information retrieval. In Proceedings of 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1998)
6. Ponte, A.: A language modeling approach to information retrieval. In PhD thesis, Dept. of Computer Science, University of Massachusetts (1998)
7. Zhai, C. and Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In Proceedings of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2001)
8. Zhai, C. and Lafferty, J.: Model-based Feedback in the Language Modeling Approach to Information Retrieval. In Proceedings of the 10th Annual International ACM Conference on Information and Knowledge Management (2002)