International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

Language Model Adaptation for Relevance Feedback in Information Retrieval

Ying-Lang Chang and Jen-Tzung Chien

Department of Computer Science and Information Engineering
Cheng Kung University, Tainan
{ylchang,chien}@chien.csie.ncku.edu.tw

Abstract

Language model is a popular method of exploiting linguistic regularities for document retrieval. To improve retrieval performance, the scheme of *relevance feedback* is adopted by adjusting the query language model using the information feedback from the retrieved documents. This study presents a new Bayesian learning approach to *instantaneous* and *unsupervised adaptation* of language model for adaptive information retrieval. We aim to compensate the domain mismatch between query and documents by adapting the query language model to meet the domains of collected documents. The maximum *a posteriori* adaptation is executed solely by using the input query without additional collection of adaptation data. The retrieved top *N* documents are utilized as relevant documents and referred as feedback to estimate mixture of language models for Bayesian document retrieval. The experiments on using TREC datasets show that the proposed method significantly outperforms the other relevance feedback methods.

Index Terms: language model, Bayesian learning, relevance feedback, document retrieval

1. Introduction

As the internet grows prosperously, the information retrieval plays an important role to allow people speedily searching useful information and relevant documents they need. There are many algorithms proposed for information retrieval such as Boolean method, vector space model and probabilistic language model. The language model approach was used to explore the linguistic regularities to achieve the desirable performance [6]. In order to tackle the sparseness of words in collected documents, the Dirichlet smoothing method [12] was proposed to improve document retrieval accuracy. The query likelihood was calculated using the smoothed document language model. Due to the short keywords or queries, the recall and precision of relevance documents were limited. Accordingly, the relevance feedback methods were presented to reinforce the system performance. The traditional relevance information required manual tagging of relevant documents. To alleviate expensive manual operations, the pseudo or blind relevance feedback was addressed. Blind feedback was developed for vector space model (VSM) based information retrieval without any relevance judgments from the users [7]. Using VSM, the query and the document are represented by the vectors consisting of the products of term frequency and inverse document frequency for individual words in query and document, respectively. The vector dimension is given by the size of word vocabulary. The matching score is determined by calculating the cosine measure between query and document vectors. By performing relevance feedback, the top *N* retrieved documents were pretended to be relevant and used to augment the query words or remove the irrelevant query words. This scheme provides useful feedback information for query such as relevant words, concurrent words and synonyms. Also, several pseudo relevance feedback methods were proposed such as query reweighing, query expansion and model-based approaches [7]. This study presents the relevance feedback method by adapting the *query language model* to the topics of documents by using top *N* retrieved documents. The model adaptation is performed for document retrieval in an *online* and *unsupervised* manner. In the literature, language model adaptation methods have been investigated by using linear interpolation, minimum discrimination information [4], dynamic marginal [10] and many others [1][8]. Also, language model adaptation was performed by combining the methods of latent Dirichlet allocation [5] and probabilistic latent semantic analysis [2]. In this paper, we focus on query language model adaptation and propose a Bayesian learning method for the mixture of language models. The mixture weights, query language model and document language models are adapted through maximizing *a posteriori* probability given query and top *N* documents as observation data. The proposed method is derived through the expectation-maximization (EM) algorithm and evaluated by document retrieval task. Several state-of-art algorithms are included for comparison.

2. Background survey

2.1. Language model for document retrieval

Ponte and Croft [6] presented the language model approach to document retrieval. The multinomial distribution was used to represent a document. The counts of each term or word in a document were observed in calculation of unigram based document model. Zhai and Laffery [12] proposed a Bayesian smoothing scheme of document language model θ_D using the Dirichlet prior

$$\widetilde{P}(w \mid d) = \frac{c(w_t, d) + \mu \cdot P(w \mid C)}{|d| + \mu}, \qquad (1)$$

where the $c(w,d)$ and $|d|$ are the counts of word w and the number of words in document d, respectively, and μ and $P(w|C)$ are Dirichlet prior parameter and background language model, respectively. This smoothing method is also referred as the maximum *a posteriori* estimation by using Dirichlet density as prior distribution with background model and hyperparameter μ. The log likelihood of a query $Q = [q_1, \cdots, q_{|Q|}]$ matching with a document d is calculated by

$$\log P(Q \mid d) = \sum_{w \in Q} \log P(w \mid d), \qquad (2)$$

and used as the decision function for document ranking among document collection $C = [d_1, \cdots, d_{|C|}]$. The language

978-1-4244-2942-4/08/$25.00 (c) 2008 IEEE

model based document retrieval model [6][12] is acted as the baseline system and viewed as the initialization stage for starting the relevance feedback process.

2.2. Relevance feedback methods

Importantly, we concern the issue of relevance feedback for improving document retrieval performance. Two model-based feedback methods [11] are surveyed. First, a mixture feedback (denoted by MIX) method is introduced. Using this method, it is assumed that top N retrieved documents are generated from the relevance model $\theta_F = P(w|F)$, which is mixed with the background model $\theta_C = P(w|C)$ and the feedback model using feedback documents. The maximum likelihood (ML) criterion is applied to find the solution by

$$\hat{\theta}_{F,\mathrm{MIX}} = \arg\max_{\theta_F} \sum_{w\in F} c(w,F) \log \begin{bmatrix} (1-\lambda)P(w|F) \\ +\lambda P(w|C) \end{bmatrix}, \quad (3)$$

where F is the observed top N documents. According to this ML criterion, the MIX model parameter $\hat{\theta}_{F,\mathrm{MIX}}$ was derived by EM algorithm. Second, a divergence feedback (denoted by DIV) method is described. This method estimates the feedback model by minimizing Kullback-Leibler (KL) divergence between the relevance model and feedback document model, and simultaneously pulling out KL-divergence between relevance model and background model

$$\hat{\theta}_{F,\mathrm{DIV}} = \arg\min_{\theta_F} \frac{1}{N} \sum_{k=1}^{N} D_{\mathrm{KL}}(\theta_F \| \theta_{d_{rk}}) - \gamma D_{\mathrm{KL}}(\theta_F \| \theta_C), \quad (4)$$

where d_{rk} denote the rank k document. The parameter γ is empirically determined. In general, both methods view top N documents as observation data and use them to estimate the parameters of relevance model $\hat{\theta}_F$. The relevance model is then interpolated with original query model θ_Q to achieve the reliable query model by

$$\hat{\theta}_Q = (1-\kappa)\cdot\theta_Q + \kappa\cdot\hat{\theta}_F, \quad (5)$$

using an interpolation parameter κ. In test session, KL divergence between the estimated query model $\hat{\theta}_Q$ and the document model θ_D is calculated by

$$D_{KL}(\hat{\theta}_Q \| \theta_D) = \sum_{w\in F} P(w|\hat{\theta}_Q) \log \frac{P(w|\hat{\theta}_Q)}{P(w|\theta_D)}, \quad (6)$$

for document retrieval. Typically, these methods correspond to the *indirect estimation* because the query model is indirectly calculated through estimating the relevance model and making interpolation with the original query model. The interpolation coefficient is empirically determined. This study presents the *direct estimation* of query model, where the interpolation weights are simultaneously estimated under the same criterion. Model robustness is assured without the need of sensitive parameters λ, γ and κ.

3. Unsupervised language model adaptation

3.1. System configuration

In what follows, we address the Bayesian learning of language model for document retrieval. Figure 1 shows the system procedure of unsupervised language model adaptation using relevance feedback. First, the query words are used to calculate the query language model. Through the first-pass retrieval process, the top N documents are retrieved in rank list. These documents are pretended to be relevant and act as the adaptation data or complementary information for unsupervised learning. The query model is adapted by maximizing the posterior density, which is a product of the likelihood of adaptation data and the prior density given the current query model. The second-pass retrieval process is performed to estimate the mixture weights and the query/document language models. The query model is accordingly re-estimated by checking convergence condition. The KL divergence in (6) is used for document ranking in retrieval process.

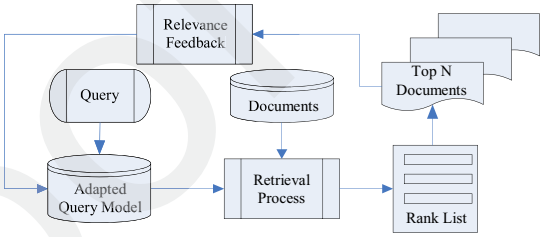

Figure 1 Language model adaptation via relevance feedback.

3.2. Query likelihood function

Given N high-ranking documents $F = [d_{r1}, \cdots, d_{rN}]$ and user query $Q = [q_1, \cdots, q_{|Q|}]$, we regard F as relevant dataset and use it for query language model adaptation. First of all, the initial query model $P(w|Q)$ and feedback document models $\{P(w|d_{rk})\}_{k=1}^{N}$ of query word w are estimated by maximum likelihood method. The likelihood function is generated by the mixture of query and document language models

$$\hat{P}(w|\theta) = \sum_{k=1}^{N} m_k P(w|d_{rk}) + m_0 P(w|Q) = \sum_{k=0}^{N} m_k P(w|d_{rk}), \quad (7)$$

where $\{m_k\}_{k=0}^{N}$ is the set of $N+1$ mixture weights. The original query model is viewed as a specific mixture component of the model, where the notation is simplified by $P(w|d_{r0}) = P(w|Q)$. The parameter set θ consists of the mixture weights $\{m_k\}_{k=0}^{N}$ and query/document language models $\{P(w|d_{rk})\}_{k=0}^{N}$. The mixture weights satisfy $\sum_{k=0}^{N} m_k = 1$. Figure 2 shows the mixture of language models, which is adapted by top N feedback documents. In conventional methods, the relevance model was estimated from top N documents and combined by original query model indirectly. In this study, we regard user query and top N documents as observation data and directly build the query model $\hat{P}(w|\theta)$ through a set of mixture weights and top N documents.

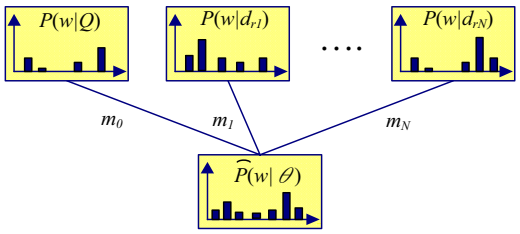

Figure 2 Query likelihood function using top N documents.

3.3. Unsupervised adaptation by relevance feedback

According to Bayesian learning theory, the model parameters are estimated by maximum *a posteriori* (MAP) criterion

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid Q) = \arg\max_{\theta} \log P(Q \mid \theta) + \log P(\theta), \quad (8)$$

where $P(\theta)$ denotes the prior density. In this study, top N documents act as the relevance feedback for language model adaptation. However, top N documents F are unknown and the sequence of mixture weight labels K is unseen beforehand. Due to the problem of incomplete data Q, the expectation-maximization (EM) algorithm [3] is applied for MAP parameter estimation subject to the constraint $\sum_{k=0}^{N} m_k = 1$. In E-step, the auxiliary function of complete data (Q, F, K) of new estimate $\hat{\theta}$ given current estimate θ is calculated by an expectation function

$$R(\hat{\theta} \mid \theta) = E[\log P(Q, F, K \mid \hat{\theta}) \mid Q, \theta] + \log P(\hat{\theta}). \quad (9)$$

In (9), the accumulated log likelihood function is expressed by

$$\log P(Q,F,K \mid \hat{\theta}) = \sum_{t=1}^{|Q|} \log \left[\sum_{k=0}^{N} \hat{m}_k \hat{P}(w_t \mid d_{rk}) \right]. \quad (10)$$

Since the parameters of mixture weights and query/document models are multinomial distributions, we accordingly choose the Dirichlet densities as the conjugate priors in MAP estimation. The prior distribution $P(\theta \mid \varphi)$ is expressed to be proportional to

$$\sum_{k=0}^{N} \left[(\nu_k - 1) \log m_k + \sum_{t=1}^{|Q|} (l_{k,w_t} - 1) \log P(w_t \mid d_{rk}) \right], \quad (11)$$

where ν_k and l_{k,w_t} are hyperparameters of parameters m_k and $P(w_t \mid d_{rk})$, respectively. In this study, the initial hyperparameters are specified by $l_{k,w_t}^{(0)} = 1 + \alpha P_{\mathrm{ML}}(w_t \mid C)$ and $\nu_k^{(0)} = 1 + \beta m_{\mathrm{ML},k}$ where the ML estimates of background language model and mixture weight are adopted. The values α and β are used to tune the balance between initial hyperparameters and likelihood functions. In M-step, we obtain new estimates of mixture weights and language models by

$$\hat{m}_k = \frac{\sum_{t=1}^{|Q|} \xi_{k,w_t} + (\nu_k^{(i)} - 1)}{\sum_{k'=0}^{N} \left[\sum_{t=1}^{|Q|} \xi_{k',w_t} + (\nu_{k'}^{(i)} - 1) \right]}, \quad (12)$$

$$\hat{P}(w_t \mid d_{rk}) = \frac{c(w_t, d_{rk}) \xi_{k,w_t} + (l_{k,w_t}^{(i)} - 1)}{\sum_{t'=1}^{|Q|} c(w_{t'}, d_{rk}) \xi_{k,w_{t'}} + (l_{k,w_{t'}}^{(i)} - 1)}, \quad (13)$$

where $\xi_{k,w_t} = P(k, d_{rk} \mid Q = \{w_t\}, \theta)$ denotes the posterior probability of query word w_t staying at mixture component k using current estimates $\theta = \{m_k, P(w_t \mid d_{rk})\}$ and feedback documents F

$$\xi_{k,w_t} = \frac{m_k P(w_t \mid d_{rk})}{\sum_{k'=0}^{N} m_{k'} P(w_t \mid d_{rk'})}. \quad (14)$$

After several iterations, EM procedure converges to MAP estimates $\theta_{\mathrm{MAP}} = \{m_{\mathrm{MAP},k}, P_{\mathrm{MAP}}(w_t \mid d_{rk})\}$. During EM iterations, the hyperparameters are updated according to the expectation function of posterior distribution of new estimates $\hat{\theta}$. The expectation function is expressed by a product of Dirichlet densities given the new hyperparameters

$$\nu_k^{(i+1)} = \nu_k^{(i)} + \sum_{t=1}^{|Q|} \xi_{k,w_t}, \quad (15)$$

$$l_{k,w_t}^{(i+1)} = l_{k,w_t}^{(i)} + c(w_t, d_{rk}) \xi_{k,w_t}. \quad (16)$$

Hyperparameters are updated from $\varphi^{(i)} = \{\nu_k^{(i)}, l_{k,w_t}^{(i)}\}$ to $\varphi^{(i+1)} = \{\nu_k^{(i+1)}, l_{k,w_t}^{(i+1)}\}$. EM procedure iteratively converges to MAP estimates $\theta_{\mathrm{MAP}} = \{m_{\mathrm{MAP},k}, P_{\mathrm{MAP}}(w_t \mid d_{rk})\}$. Finally, the adapted query language model is obtained by $\{P_{\mathrm{MAP}}(w_t \mid Q) = P_{\mathrm{MAP}}(w_t \mid d_{r0})\}$. The parameter size is obtained by V+N where V is the vocabulary size of query model and N is the number of mixture weights. This size is much less than $2V$ of the other methods in (3)-(4) where the parameters of feedback model and query model are needed. Also, the computation costs are comparable because EM iterations are performed in these methods.

4. Experiments

4.1. Experimental setup and implementation issue

We evaluated the document retrieval performance by using the baseline methods including vector space model (VSM) and language model (LM) and the relevance feedback methods including divergence feedback (DIV), mixture feedback (MIX) and the proposed method using adaptive Bayesian mixture model (denoted by ABM hereafter). The evaluation was performed on two TREC data sets (WSJ87 and AP89_01_06) [9]. There were 46,448 documents in WSJ87 and 42,682 documents in AP89_01_06. Topic 51-100 queries were used as test queries. The words in the fields of title and description were adopted as query words. Following the standard TREC evaluation, we retrieved top 1,000

documents for each query and calculated the mean average precision (mAP). Additionally, we calculated the precision at different numbers of retrieved documents. Different cases of top N documents were evaluated to see the sensitivity of feedback document size to retrieval performance. In the implementation, the interpolation weights of background language model in MIX and DIV methods were set to be 0.5 [11]. The parameters μ, λ, γ and κ were specified by default values in Lemur toolkit (http://www.lemurproject.org/). Also, the coefficients α and β were set to be 0.2 and 0.01, respectively. The number of EM iteration was fixed to be three.

4.2. Experimental results

First of all, Table 1 shows the effect of top N documents in relevance feedback methods using VSM, DIV, MIX and ABM. AP89_01_06 dataset is adopted for evaluation. We compare different numbers of top N documents (N=10, 20, 50, and 100). The mAP is decreased when higher N is considered. It is because that the selected documents contain more irrelevant documents. Merging too many top N documents does not help retrieval performance. We fix N=20 in subsequent evaluation. Table 2 reports the retrieval performance with/without relevance feedback in terms of mAP and precision at positions 5, 20, 100 and 500. The results of using WSJ87 and AP89_01_06 are provided. For most cases, the performance of relevance feedback methods is better than those of baseline system without feedback information. Notably, ABM consistently obtains the best performance compared to VSM, DIV and MIX. In addition, Figure 3 displays the precision-recall curves for four relevance feedback methods in using AP89_01_06 dataset. Compared to VSM, DIV and MIX, the proposed ABM obtains the highest precisions in case of low recalls and attains comparable precisions in case of high recalls.

Table 1 Mean average precision by using different top N documents.

	N=10	N=20	N=50	N=100
VSM	0.1708	0.1837	**0.1706**	**0.1793**
DIV	0.1973	0.1946	**0.1918**	**0.1916**
MIX	0.2001	0.2004	0.1947	**0.1961**
ABM	0.2182	0.2263	0.2001	**0.1900**

Table 2 Precision at positions 5, 20, 100, 500 and mean average precision with/without relevance feedback methods.

	WSJ87					
	No Feedback		Feedback			
	VSM	LM	VSM	DIV	MIX	ABM
P@5	0.359	0.465	0.3796	0.4898	0.4776	0.5673
P@20	0.269	0.348	0.2939	0.3510	0.3622	0.3673
P@100	0.152	0.182	0.1586	0.1908	0.1959	0.1898
P@500	0.055	0.059	0.0553	0.0615	0.0630	0.0607
mAP	0.294	0.354	0.2747	0.3684	0.3761	0.3835
	AP89_01_06					
	No Feedback		Feedback			
	VSM	LM	VSM	DIV	MIX	ABM
P@5	0.319	0.443	0.4000	0.4681	0.4383	0.5447
P@20	0.278	0.348	0.3255	0.3596	0.3766	0.3830
P@100	0.142	0.166	0.1557	0.1715	0.1806	0.1781
P@500	0.047	0.049	0.0500	0.0500	0.0510	0.0504
mAP	0.160	0.187	0.1837	0.1946	0.2004	0.2263

5. Conclusions

This paper addressed the Bayesian learning of language models by using the information source from relevance feedback. The instantaneous and unsupervised adaptation was performed by adopting the high-ranking retrieved documents. The query likelihood function was established under the mixture of language models. An EM algorithm was developed to solve MAP estimation of language model. The updating formulas of language models, mixture weights and their hyperparameters were derived. From the experimental results, the proposed ABM method achieved the highest precisions compared to baseline system without relevance feedback and other state-of-art model-based relevance feedback methods. In the future, we are applying the proposed method for Chinese spoken document retrieval.

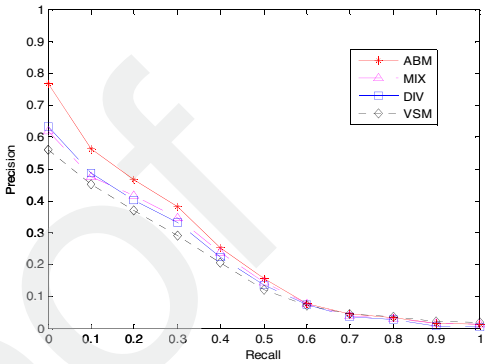

Figure 3 Precision-recall curves for different methods.

6. References

[1] M. Bacchiani and B. Roark, "Unsupervised language model adaptation", in *Proc. of ICASSP*, pp. 224-227, 2003.

[2] J.-T. Chien and M.-S. Wu, "Adaptive Bayesian latent semantic analysis", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 198-207, 2008.

[3] P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B, vol. 39, no. 1, pp. 1-38, 1977.

[4] M. Federico, "Efficient language model adaptation through MDI estimation", in *Proc. of EUROSPEECH*, vol. 4, pp. 1583-1586, 1999.

[5] Heidel, H. Chang and L. Lee, "Language model adaptation using latent Dirichlet allocation and an efficient topic inference algorithm", in *Proc. of INTERSPEECH*, pp. 2361-2364, 2007.

[6] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval", in *Proc. of ACM-SIGIR*, pp. 275-281, 1998.

[7] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.

[8] Y. C. Tam and T. Schultz, "Unsupervised language model adaptation using latent semantic marginals", in *Proc. of INTERSPEECH*, pp. 2206-2209, 2006.

[9] E. Voorhees and D. Harman, *Proc. of Text Retrieval Conference (TREC1-9)*, NIST special Publications, 2001.

[10] W. Wang and A. Stolcke, "Integrating MAP, marginals, and unsupervised language model adaptation", in *Proc. of INTERSPEECH*, pp. 618-621, 2007.

[11] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval", in *Proc. of CIKM*, pp. 403-410, 2001.

[12] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval", in *Proc. of ACM-SIGIR*, pp. 334-342, 2001.