http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (2): 236-246, 2010
ISSN 1812-5638

Semantic Clustering Based Relevance Language Model

[1]Qiang Pu and [2]Daqing He
[1]School of Computer Science and Engineering,
University of Electronic Science and Technology of China,
610054, Chengdu, Sichuan, People's Republic of China
[2]School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

Abstract: How to effectively generate clusters and use the information in clusters to improve information retrieval performance are still open research questions. By viewing a document as an interaction of a set of independent hidden topics, we propose a novel semantic clustering technique using independent component analysis. Then within language modeling framework, we apply the obtained semantic topic clusters into the estimation process of relevance model. We expect that semantic clustering will filter out those noisy documents so that the estimation of relevance model is only based on relevant documents and some useful semantic information. A semantic cluster is activated to be the most similar to a user's information need by user's query, the documents in the activated semantic cluster and the keywords of representing the activated semantic cluster are used for the estimation of relevance model. Therefore, we obtain a semantic cluster based relevance language model that uses pseudo relevance feedback technique without requiring any relevance training information. We applied the model in experiments on five TREC data sets. The experiment results show that our model can significantly improve retrieval performance over previous language models including relevance-based language models. We think that the main contribution of the improved performance comes from the estimation of relevance model on the semantic cluster that is closely related to a user's information need.

Key words: Semantic clustering, relevance language model, query expansion, pseudo relevance feedback, independent component analysis, information retrieval

INTRODUCTION

Pseudo Relevance Feedback (PRF), which assumes that top-ranked retrieved documents are relevant to the user's information need, is an effective method of enhancing user's initial query to improve the retrieval performance.

However, not all top-ranked retrieved documents are truly relevant and the noise introduced by non-relevant documents could cause the expansion of the query drifting away from the original query topic and may hurt performance for about one-third of a given set of topics (Sakai *et al*., 2005). If the relevant documents in the top-ranked retrieved set can be carefully identified, the query expansion (Efthimiadis, 2000) or query model update (Zhai and Lafferty, 2001) based on the relevant documents will be closer to user's information need.

Clustered-based retrieval method may group content similar documents into a cluster in which the documents match to the users' need if that cluster is relevant.

Within the language modeling framework (Ponte and Croft, 1998). Liu and Croft (2004) demonstrated that cluster-based retrieval can generate significantly better effectiveness over non-cluster retrieval if good clusters can be identified and used. They used K-means algorithm for clustering documents and selected certain clusters for smoothing the document language model.

Lee *et al*. (2008) improved PRF by using document clusters to find dominant documents and iteratively feeding the documents to emphasize the core topics of a query. They found that this resampling approach contributed to higher relevance density for feedback documents and resulted in more accurate retrieval.

Wang *et al*. (2008) studied negative examples in relevance feedback. Although, clustering was not explicitly employed, their way of handling negative examples could be easily extended into explicitly clustering according to the dissimilarity of returned documents to the query.

Corresponding Author: Qiang Pu, School of Computer Science and Engineering,
University of Electronic Science and Technology of China,
610054, Chengdu, Sichuan, People's Republic of China Tel: +1-412-624-6858 Fax: +1-412-648-7001

Sakai *et al*. (2005) proposed a selective sampling method for PRF, where the number of selected documents and that of the expansion terms for each topic were adjustable. They used memory resetting algorithm to select documents.

Wei and Croft (2006) proposed to use LDA-based language model for clustering and also allowed a document to be in multiple topics. They reported that LDA-based retrieval is a promising method for IR.

This study proposes a semantic clustering approach to organizing the top-ranked documents with similar semantic topics into a cluster, then uses documents in a semantic cluster close to the user's query, not the whole top-ranked documents, as feedback documents for a better PRF within the language modeling framework.

Our clustering approach is based on a high order statistical method called Independent Component Analysis (ICA) (Hyvärinen, 1999), which groups documents in an ICA latent semantic space instead of the traditional semantic space (Deerwester *et al*., 1990; Hofmann, 1999) which defines latent semantic topics as the dimensions that capture the maximal variance, but maximal variance is not typically considered as a real topic of the data (Hyvärinen, 1999).

The ICA is a method of representing a set of multivariate observations as a linear combination of unknown latent variables that are statistically independent (Hyvärinen, 1999). If we view documents as the interactions of a set of independent latent topics, ICA can reveal the semantic structure in original text data and the latent variables can be thought as the topics in the text data. As a result, documents can be grouped according to their probabilities of belonging to a topic in the latent semantic space constructed by ICA, so we call this clustering method semantic based clustering method. When applied for PRF, we hypothesize that this clustering method will bring some helpful semantic information in generating better clusters on the top-ranked retrieved documents so that relevant and non-relevant documents can be easily separated.

We propose an activation process to select one of semantic clusters identified by ICA, which is most similar to user's query. This activation performs two roles: first, it activates the documents under the topic, which filters out more noise than that of using all top-tanked retrieved documents as relevance feedback documents. Second, it also activates highly relevant keywords in the topic, on which we can estimate a semantic keywords cluster model that can enhance topic content part of the estimation of a document model.

The key difference of our semantic clustering approach from previous works above is that we cluster documents in a latent semantic space so that the selected clusters are semantically related to the user's query. The semantic information of clusters can be integrated into the estimation of a language model. For example, the probability of a document belonging to a semantic cluster can be used as a prior, which can differentiate the contribution of each feedback document to the estimation of a document model, but the traditional estimation usually uses a uniform distribution on feedback documents. Another example is that terms from semantic keywords cluster can represent and enhance topic content part when estimating a document model.

We assume that a relevance model (Lavrenko and Croft, 2001) exists between a user's query and the documents in semantic clusters, so both relevant documents and query can be considered as samples from the relevance model. Relevance model comes from the difficulty in estimating model parameters in the classical probabilistic model when we do not have relevance judgments (Zhai, 2008). We will estimate the relevance model based only on a user's query, the documents in semantic clusters without any relevance judgments.

SEMANTIC CLUSTERING-BASED RELEVANCE MODEL

Here, we describes clustering method using ICA framework and our semantic clustering-based relevance model.

Maximum likelihood approach for ICA: The ICA model in text data analysis can be described as follows:

$$X = AS \tag{1}$$

where, X is a term-document matrix that holds m observed terms in each row with n documents in each column. A is an unknown m×k mixing matrix with non-orthogonal transformation basis and S is another unknown matrix that holds k latent topics in each row and n document samples in each column. In this paper, we use Bayes Information Criterion (BIC) (Hansen *et al*., 2001) to determine the k value which indicates how many latent topics exist in a set of documents. Each latent topic s_i (i = 1, 2,…, k) is mutually independent and non-Gaussian. Therefore:

$$p_S(S) = \prod_{i=1}^{k} p(s_i) \tag{2}$$

The term-document matrix X in ICA should be viewed as linear mixtures of latent topics. If the number of topics is assumed to be equal to the number of observed signals, the mixing matrix is square, the goal of ICA is then to find

the unmixing matrix $W = A^{-1}$ based on the observed matrix X. Therefore, we can rewrite Eq. 1 as:

$$X = W^{-1}S \tag{3}$$

In probabilistic framework, according to Eq. 2 and 3 the probability density function of X is then:

$$\hat{p}_X(X) = \left|\det W\right| p_S(S)\big|_{S=WX} = \left|\det W\right| \prod_{i=1}^{k} p(s_i)\big|_{S=WX} \tag{4}$$

where, $|\bullet|$ represents an absolute value. Because $\hat{p}_X(X)$ is also a function of W, we can denote it as $\hat{p}_X(X, W)$. According to Eq. 4, then the likelihood of X is:

$$l(X,W) = \ln \hat{p}_X(X,W) = \ln\left|\det W\right| + \sum_{i=1}^{k} \ln p_i(s_i)\big|_{S=WX} \tag{5}$$

Based on the principle of maximum likelihood, if we can estimate a $\hat{W}$ such that it makes the maximum of the mathematical expectation of E [l(X,W)], denote it as $\tilde{L}_{ML}(W)$, then $\hat{W}$ is the solution to the unmixing matrix. If finite samples of n documents are available, $\tilde{L}_{ML}(W)$ can be estimated by these finite samples as:

$$\tilde{L}_{ML}(W) = \frac{1}{n}\sum_{i=1}^{n} l(X(i),W) = \frac{1}{n}\sum_{i=1}^{n} \ln \hat{p}_X(X(i),W) \tag{6}$$

We can calculate the gradient of equation for updating the unmixing matrix W in an iterative optimization method.

$$\nabla_W \tilde{L}(W) = \frac{\partial}{\partial W}\ln\det W + \sum_{i=1}^{k} \frac{\partial \ln p(s_i)}{\partial s_i}\frac{\partial s_i^T}{\partial W} = (W^T)^{-1} + F(S)X^T \tag{7}$$

where, $\tilde{N}_W \tilde{L}(W)$ is the gradient of l(X,W) on W. We denote:

$$F(s_i) = \frac{\partial}{\partial s_i}\ln p(s_i)$$

where, Φ = -tanh, because it is suitable for the infomax solution to separate super-Gaussian signals, e.g. the text data (Kolenda, 2002). This also implies the source distribution $p(S) = 1/\pi \exp(-\ln \cosh S)$.

A natural gradient algorithm is used to iteratively estimate the optimal $\hat{W}$:

$$\tilde{L}(W + \Delta W) \approx \tilde{L}(W) + \left\langle \nabla_w \tilde{L}(W) \middle| \Delta W \right\rangle \tag{8}$$

where, ΔW is a micro-variant matrix neighbor to W, $\langle \nabla_w \tilde{L}(W) | \Delta W \rangle$ is the inner product. According to Amari (1998):

$$\Delta W = -\alpha \nabla_w \tilde{L}(W) W^T W \tag{9}$$

where, α is a learning ratio. When the iterative process stops, we get the solution to W, then the latent topics in a set of documents will be obtained by S = WX.

Preprocessing for ICA: To obtain fast-converging ICA decomposition, to make the mixing matrix square and to cure the ill-posed problem (Lautrup *et al*., 1995), we can use Principal Component Analysis (PCA) as a preprocessing step to form the latent semantic space before ICA decomposition. The corresponding technique for finding this semantic space is Singular Value Decomposition (SVD) by which the term-document matrix X is decomposed as:

$$X_{m\times n} = T_{m\times k} L_{k\times k} D^T_{n\times k} \tag{10}$$

where, L is a diagonal matrix containing k (k<=r) largest singular values of X, r is the rank of X. T and D are the term and document matrix in semantic space, of which columns correspond to the k largest singular values in L. The term matrix T is used as an input for ICA decomposition.

$$X_{m\times n} = T_{m\times k} A_{k\times k} S_{k\times n} \tag{11}$$

where, A is the ICA mixing matrix and S holds the k latent topics in a set of n documents.

Semantic clustering: According to the solution to unmixing matrix W and Eq. 10 and 11, the independent components can be calculated as follows:

$$S = A^{-1} L D^T = W L D^T \tag{12}$$

By using softmax normalization (Kolenda, 2002), the value in matrix S can be converted into a probability that describes the degree that a document belongs to a latent topic. For example, see the following matrix in Fig. 1, p_{ij} represents document doc_j belongs to topic $topic_i$ with a probability p_{ij}.

With the help of such a probability, a document doc_j could be assigned to a latent topic $topic_i$ according to the highest probability p_{ij}, as shown formally as follows:

$$p\left(doc_j \mid topic_i\right) = \arg\max_i p_{ij} \tag{13}$$

Therefore, according to Eq. 13, the clustering of documents can therefore converted to a cluster label assignment to each document based on the obtained maximum probability in semantic cluster matrix shown in Fig. 1.

	doc_1	doc_2	...	doc_n
$topic_1$	p_{11}	p_{12}	...	p_{1n}
$topic_2$	p_{21}	p_{22}	...	p_{2n}
⋮	⋮	⋮	...	⋮
$topic_k$	p_{k1}	p_{k2}	...	p_{kn}

Fig. 1: Semantic cluster matrix

We can use a set of keywords as representation of a semantic cluster. The probability of a keyword $term_i$ belonging to a cluster $topic_j$ can be obtained by the back projection technique (Kolenda, 2002) shown as follows:

$$p(term_i \mid topic_j) = e_{i,j} \in (TA)_{m,k} \quad (14)$$

where, i = (1,…,m), j = (1,…, k). The transformation TA holds the mixing proportions coming from the term space to the lower dimensional semantic topic space (Kolenda, 2002). If the probability $p(term_i|topic_j)$ is greater than a threshold, we view the $term_i$ as the keyword of this semantic topic $topic_j$.

As a result, ICA first helps to cluster documents in the latent semantic space based on its topic separation ability. It then helps to identify a set of representative terms for each latent topic. We expect that the two helps from ICA will enable a semantic model that facilitates a better estimation of a relevance model as in relevance based language models.

Relevance model: Given a set of documents $D = (d_1,…,d_n)$ and a query $Q = (q_1,…,q_k)$, to what degree document D is relevant to query Q can be represented by the conditional probability p (R = 1|Q, D). Here, $R \in \{0, 1\}$ is a binary relevance random variable. 0 represents that the document and the query are not relevant, whereas 1 represents that the document is relevant to the query. According to Bayes' rule, p (R = 1|Q, D)∝p (Q, D|R = 1) where, ∝ represents is proportional to. We can further rewrite it by query likelihood as p (Q, D|R = 1)∝ p (D|Q, R = 1). This is the relevance model demonstrated by Lavrenko and Croft (2001).

Relevance model proves to work well even when there is no relevance training data available. If the top-ranked retrieved documents are used to approximate relevant documents, the estimate of relevance model is the estimate the probability of

$$p(D \mid Q,R) = \prod_{i=1}^{n} p(d_i \mid Q,R)$$

where, each probability $p(d_i \mid Q,R)$ captures term occurrences in relevant documents (Lavrenko and Croft, 2001; Zhai, 2008). Consider a term w could be generated in a relevant document d_j, the estimation of the relevance model is then converted to the calculation of probability: p(w|Q, R).

Estimation of a semantic clustering-based relevance model: Our approach is inspired by relevance based language models, but we employs a different approach to estimating the conditional probability p(w|Q, R) within the language model framework. We think that it is sub-optimal in the estimation of the relevance model to assume that all top-ranked retrieved documents are relevant documents because not all of them are really relevant. The following gives our estimation of the relevance model.

As stated, we view a document as being generated based on an interaction of a set of independent hidden topics and a set of documents can be grouped into a semantic cluster because they are all related to a hidden topic. If a hidden topic is identified to be relevant to a user's query, all documents in the semantic cluster associated with the hidden topic could be activated as the relevance documents to the query. Therefore, it is reasonable and optimal to model the relevance based on the semantic clusters.

Based on topical relevance criterion, we select the semantic cluster whose associated hidden topic is the most similar to the user's query. Therefore, for a query model θ_Q which is estimated from a user's query and a semantic model θ_s which can be estimated from a semantic cluster, we can utilize Kullback-Leibler (KL) divergence to measure the closeness between the two models (Lafferty and Zhai, 2001). The semantic cluster that has the smallest KL divergence value to the query would be activated for estimating the relevance model. This calculation can be written as follows:

$$\arg\min_{\theta_s} D(\theta_Q \| \theta_s) = \sum_{\substack{w \in V \\ \theta_s \in \Theta_s}} p(w \mid \theta_Q) \log \frac{p(w \mid \theta_Q)}{p(w \mid \theta_s)} \quad (15)$$

where, Θ_s represents all the semantic clusters derived from ICA algorithm. Term w is from the vocabulary V. Equation 15 clearly indicates that, instead of using all top-ranked retrieved documents, our approach only uses the documents that are semantically most similar to the query for estimating the relevance model.

We use semantic model θ_s as a bridge to compute the association between each term and the query. Similar to the Model 2 in (Lavrenko and Croft, 2001), we assume that the query terms $q_1,…, q_k$ are independent to each other, whereas we keep their dependence to term w. That is, it gives the probability of co-occurrence between w and the query in the semantic cluster. We can formally derive the probability p (w|Q, R) formula as follows:

$$p(w|Q,R) \propto p(Q|w,R)p(w) = p(w)\prod_{i=1}^{k} p(q_i|w) = p(w)\prod_{i=1}^{k} p(\theta_S|w)p(q_i|\theta_S) \quad (16)$$

where, q_i is the query term in a query Q. Using Bayes' theorem, $p(\theta_s|w)$ can be transformed to:

$$p(\theta_S|w) = \frac{p(w|\theta_S)p(\theta_S)}{\sum_{\theta_{S'} \in \Theta_S} p(w|\theta_{S'})p(\theta_{S'})} \quad (17)$$

A document model θ_d can be estimated from a document d in the activated semantic cluster model θ_s, denote as $\theta_d \overset{d}{\leftarrow} \theta_S$, then $p(w|\theta_s)$ can be computed as follows:

$$p(w|\theta_S) = \sum_{\theta_d \overset{d}{\leftarrow} \theta_S} p(d|\theta_S)p(w|\theta_d) \quad (18)$$

Note that in Eq. 18, we use the activated semantic model as well as document models to estimate the relevance model. Probability $p(d|\theta_s)$ uses the semantic models to establish different prior probabilities for each document model θ_d, which implies that document model contains query-related semantic information. This is consistent to our intuition that query-related documents contribute more to the estimation of relevance model than those documents that are dissimilar to the query.

We use both the semantic keywords cluster model θ_{sw} and the background model θ_c to smooth the estimation of a document model θ_d, the probability of $p(w|\theta_d)$ is computed as follows:

$$p(w|\theta_d) = \lambda_1 \frac{c(w,d)}{\sum_{w' \in d} c(w',d)} + \lambda_2 P(w|\theta_{SW}) + \lambda_3 P(w|\theta_C) \quad (19)$$

where, c (w,d) is the frequency of a term w in a document d. The semantic keywords cluster model θ_{sw} is directly obtained from the semantic keywords cluster generated by ICA using back projection technique (Eq. 14). θ_c is the collection language model. The parameters of linear interpolation λ_1, λ_2 and λ_3 ($\lambda_1+\lambda_2+\lambda_3 = 1$) are the coefficients of document topic model, the semantic keywords cluster model and the collection model respectively, which denote the corresponding contribution to the estimation of the document model. The term prior probability p (w) is defined as follows:

$$p(w) = \sum_{\theta_S \in \Theta_S} p(w|\theta_S)p(\theta_S) \quad (20)$$

where, $p(\theta_s)$ is the prior probability of the semantic models, which can be represented either by the entropy of each semantic cluster or by assuming a uniform distribution over all semantic clusters. In this study, we adopt the uniform distribution approach over the universe of semantic cluster models Θ_s.

EXPERIMENT SETUP

To examine the performance of our model, we conducted experiments on five data sets taken from previous TREC campaigns: the Associated Press Newswire (AP) 1988-90 with query topics 51-200, Wall Street Journal (WSJ) 1987-92 with query topics 51-200, Financial Times (FT) 1991-94 with query topics 301-400, San Jose Mercury News (SJMN) 1991 with query topics 51-150, Los Angeles Times (LA) with query topics 301-400. For all collections, title field of TREC topics was used as the query. Queries that have no relevant documents in the judged pool for a specific collection had been removed from the query set for that collection. A summary of the collections and query sets is shown in Table 1.

The Indri 2.9 system was used for indexing and retrieval. All collections and queries were stemmed using the Porter stemmer and stop-words were removed as well.

We used queries 51-150 and the AP collection for parameters training and queries 151-200 on AP collection and all other queries and WSJ, SJMN, FT and LA collections for testing. The initial query results were generated using basic query likelihood language model. The implementation of ICA algorithm was from DTU: Toolbox (Kolenda *et al*., 2002). To design our study in the form of comparative experiments, the basic query likelihood language model and Indri's implementation of Lavrenko's relevance based language model were used as the two baselines. We also compared our results to other cluster-based retrieval models. The evaluation measure used Mean Average Precision (MAP).

As stated, our model for PRF involves semantic clustering and the estimation of relevance model. After obtaining initial retrieval results, 50 top-ranked documents were selected for semantic clustering by ICA. During the semantic clustering, the optimal number k of the topic was estimated by Bayes information criterion. When choosing terms to be the cluster representative, we examined the

Table 1: TREC collections used for experiments

Collection	Description	No. of docs	Queries (Topics title only)	No. of queries with relevant docs
AP	Association Press 88-90	242,918	51-200	149
WSJ	Wall Street Journal 87-92	173,252	151-200	50
SJMN	San Jose Mercury News	90,257	51-150	94
FT	Financial Times 91-94	210,158	301-400	95
LA	Los Angeles Times	131,896	301-400	98

probability of $p(term_i|topic_j)$, if the probability was greater than an ad hoc threshold 0.3, the term was selected as the representative of the cluster, refer to Eq. 14 for calculation detail.

We chose those semantic clusters whose KL divergence was closest to the query during the estimation of the relevance model, see Eq. 15. Based on the selected cluster, we selected n terms $w_1, \ldots, w_n$ to do the query expansion. The expansion terms were combined with the original query using linear interpolation with a parameter β to tune the relative importance between the expansion terms and the original query. The expanded query in the indri query form is:

weight(β #combine($q_1 \ldots q_k$) (1 - β) #weight(p_1 $w_1 \ldots p_n$ w_n))

Equation 19 contains three parameters λ_1, λ_2 and λ_3 for tuning the estimation of a document model. Intuitively, we assume that $\lambda_1 > \lambda_2 > \lambda_3$. The parameters tuning conducted on the training topics demonstrated that the retrieval performance was the best when λ_1 was set between 0.4 and 0.7. Therefore, we set these three parameters as: $\lambda_1 = 0.45$, $\lambda_2 = 035$, $\lambda_3 = 0.2$.

EXPERIMENT RESULTS

Semantic clustering-based relevance model in improving retrieval: To control PRF, there are three parameters to be tuned: the number of feedback documents d, the number of feedback terms n and the coefficient β for integrating original query terms with the expanded terms. We performed an exhaustive search to look for the optimal parameter values. The set of parameter values tested in our training were: d$\in$ {5, 10, 25, 50}, n$\in$ {10, 25, 50, 75, 100}. For decreasing computation, we fixed $\beta = 0.5$ which can result in better and safe results according to Zhai and Lafferty (2001).

Under the X-axes in Fig. 2, there were two lines of numbers. The numbers at the bottom line, that were the numbers of feedback documents d, d = (5, 10, 25, 50), organized the X-axes into four groups. Each group means that the corresponding number of documents were used as feedback documents. The numbers at the top line represented that the numbers of feedback terms n, n = (5, 10, 25, 50, 75, 100) were used as feedback terms in each of the four groups. The Y-axes showed the MAP values corresponding to d and n.

During the training, our semantic clustering based relevance model (SRM in Fig. 2) achieved better performance over Lavrenko and Croft's relevance model (RM in Fig. 2) in almost all cases. But we did not perform statistical testing on these results, all statistical testing

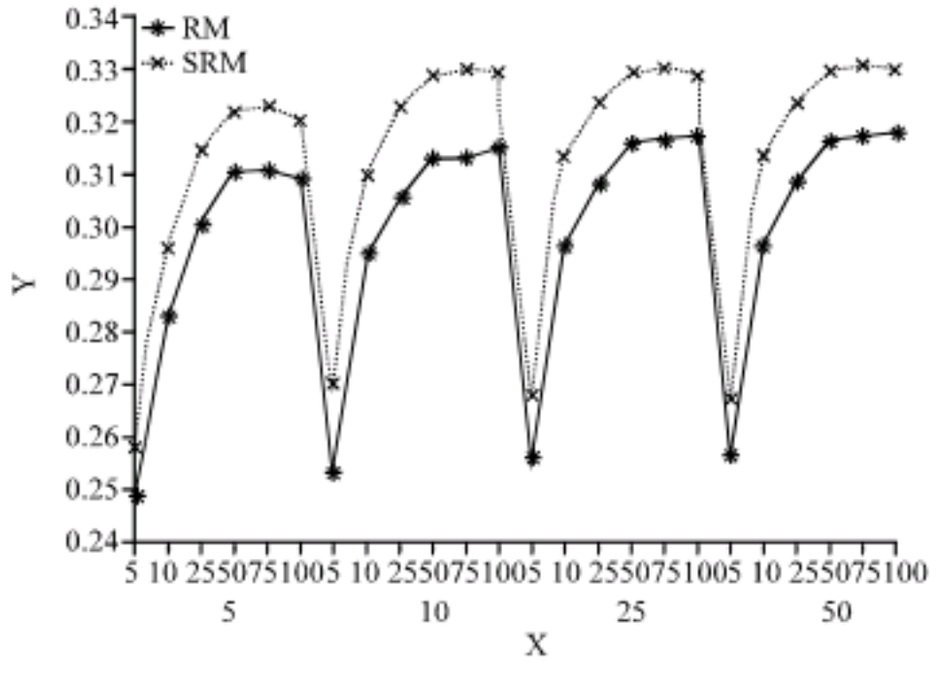

Fig. 2: Training of relevance feedback parameters for RM and SRM: the number of feedback documents d and the number of feedback terms n with fixed interpolation coefficient parameter $\beta = 0.5$

Table 2: Performance comparisons on training data sets among three retrieval algorithms

Performance	LM	%chg	RM	%chg	SRM
Rel	16768		16768		16768
Rret	8514	20.37	9000^{α}	+13.87	$10248^{\alpha\beta}$
MAP	0.2106	56.32	0.2307	+42.7	$0.3292^{\alpha\beta}$
RP	0.2519	43.11	0.2641	+36.5	$0.3605^{\alpha\beta}$
11Avg	0.2285	50.9	0.244	+41.31	$0.3448^{\alpha\beta}$

Training: TREC AP collection, queries 51-150

were done on the testing results. Based on the tuning results and a fixed β value, we selected the following combination of parameters: (1) with SRM, the number of feedback documents d = 50 and the number of feedback terms n = 75; (2) with RM, d = 25, n = 100. Note that, because we only chose the top 50 initial retrieved documents to perform the semantic clustering, which means that the number of documents in a cluster will at most be equal to 50, thus the training parameter d = 50 means that we used all the documents in the selected semantic cluster for relevance feedback.

Table 2 gave the retrieval performance of SRM against that of RM and basic query likelihood language model (LM) on training topics. The superscript α indicated statistically significant improvements over LM and β indicated statistically significant improvements over RM respectively with a 95% confidence by the Wilcoxon test.

In Table 2, comparing SRM against LM and RM, MAPs were improved by 56.32% (0.3292 vs. 0.2106) and 12.2% (0.3292 vs. 0.2307), R-precisions were improved by 43.11% (0.3605 vs. 0.2519) and 36.5% (0. 3605 vs. 0.2641), 11-point averages were improved by 50.9% (0.3448 vs. 0.2285) and 41.31% (0.3448 vs. 0.2440). Meanwhile, the number of relevant retrieved documents of SRM showed 20.37 and 13.87% improvements over LM and RM,

Table 3: Performance comparisons on testing data sets among three retrieval algorithms

Collection	LM	%chg	RM	%chg	SRM	Upper
AP	0.1956	+71.01	0.2500	+33.80	$0.3345^{\alpha\beta}$	$0.4451^{\alpha\beta\gamma}$
WSJ	0.3118	+39.16	0.3289	+31.92	$0.4339^{\alpha\beta}$	$0.5138^{\alpha\beta\gamma}$
SJMN	0.2112	+69.03	0.2497	+42.97	$0.3570^{\alpha\beta}$	$0.5593^{\alpha\beta\gamma}$
FT	0.2514	+32.74	0.2600	+28.35	$0.3337^{\alpha\beta}$	$0.6212^{\alpha\beta\gamma}$
LA	0.2164	+36.69	0.2307	+28.22	$0.2958^{\alpha\beta}$	$0.6398^{\alpha\beta\gamma}$

The superscript α indicates statistically significant improvements over LM, β indicates statistically significant improvements over RM, and γ indicates statistically significant improvements over SRM respectively with a 95% confidence by the Wilcoxon test

Table 4: Performance comparisons on testing data sets among four retrieval algorithms

Collection	CBDM	%chg	LBDM	%chg	Resampling	%chg	SRM
AP	0.2775	+20.54	0.2869	+16.59	0.2906	+15.11	0.3345
WSJ	0.3445	+25.95	0.3606	+20.33	0.4033	+7.59	0.4339
SJMN	0.2673	+33.56	0.2603	+37.15			0.3570
FT	0.2845	+17.29	0.2907	+14.79			0.3337
LA	0.2621	+12.86	0.2715	+8.99			0.2958

respectively. SRM had statistically significant improvements over LM and RM on the all measures. However, there was no significant difference between LM and RM on all measures except the number of relevant retrieved documents.

Table 3 showed the performance comparisons on the testing data sets among LM, RM and SRM retrieval algorithms. The parameter setting for obtaining these results were: (1) with SRM, the number of feedback documents d = 50, the number of feedback terms n = 75 and coefficient $\beta = 0.5$, (2) with RM, d = 25, n = 100 and $\beta = 0.5$. The upper in the last column referred to the upper bound performance when using SRM. We selected the top 50 true relevant documents as feedback documents to obtain the upper bound. This upper bound will help us to establish the best performance that our semantic clustering-based relevance model could produce when all feedback documents are truly relevant. Looking at other columns, our SRM method had significant improvements over LM and RM on all collections whereas RM had no significant difference over LM on all collections.

In Table 4, we compared the retrieval results on the testing data sets with another three important related works: cluster-based method (CBDM) from Liu and Croft (2004), LDA-based method (LBDM) from Wei and Croft (2006) and Resampling method from Lee *et al*. (2008). We neither implemented CBDM, LBDM and Resampling methods nor conduct statistical tests between our results and that of the three methods in our experiments. The data presented in Table 4 were directly copied from their original publications. From the relative improvement percentages, however, we believe that our method has clear advantage in choosing better relevance documents over the three methods.

Figure 3 showed the comparison of retrieval performance in precision-recall curves. In both training and testing phases, SRM always showed a clearly higher performance than LM and RM. Only at high recall side on WSJ and FT collections, SRM showed a little bit worse performance than that of LM and RM. We also observed that RM and SRM contributed differently to different collections. When trained on the AP collection and tested on WSJ and FT collections, RM had a little bit improvements over LM. However, under the same condition, our SRM method contributed more improvements over LM, especially at low recall side. When tested on the AP collection and the SJMN collection, both RM and SRM had consistent improvement over LM. The LA collection was probably a specific collection for testing relevance based methods. RM could hardly improve the performance over LM and achieved even worse results than LM at the low recall side. Our SRM method also struggled on the LA collection. Its performance had the largest distance to the upper bound and its improvements over RM and LM on the LA collection were also not as salient as that of on other collections.

Semantic clusters in capturing topics: Present evaluation methods not only included extrinsic retrieval performance measures, but also considered the quality of the semantic clusters. By assuming that a set of relevant expansion terms would be topically closer to the true topic than non-relevant expansion terms, we can evaluate the quality of the semantic clusters. We defined four kinds of topics, two topics were estimated from the expansion terms, which were generated from either RM or SRM, written as $Topic_{RM}$ and $Topic_{SRM}$, a random topic was estimated from the background collection model, written as $Coll_{all}$ and a true topic was estimated from the relevant documents of a query, written as $Coll_{rel}$. The topic distance between Topic and Coll was measured by KL divergence:

$$\text{topic dist} = \sum_{w \in V} p(w \mid \text{Topic}) \log \frac{p(w \mid \text{Topic})}{p(w \mid \text{Coll})} \qquad (21)$$

where, V is the vocabulary, Topic is either $Topic_{SRM}$ or $Topic_{RM}$, Coll is either $Coll_{all}$ or $Coll_{rel}$, Coll$\in$ {AP, WSJ, SJMN, FT, LA}. We expect that a topic from better expansion terms would have smaller distance to $Coll_{rel}$ but larger distance to $Coll_{all}$.

Table 5 gave the results of the comparison of topic distance between $Topic_{SRM}$ (represented as SRM in Table 5), $Topic_{RM}$ (RM in Table 5) to $Coll_{all}$ (All in Table 5) and $Coll_{rel}$ (Rel in Table 5) on testing collections. Symbol L and S meant that $Topic_{SRM}$ had larger or smaller topic distance than that of $Topic_{RM}$ to $Coll_{all}$ and $Coll_{rel}$,

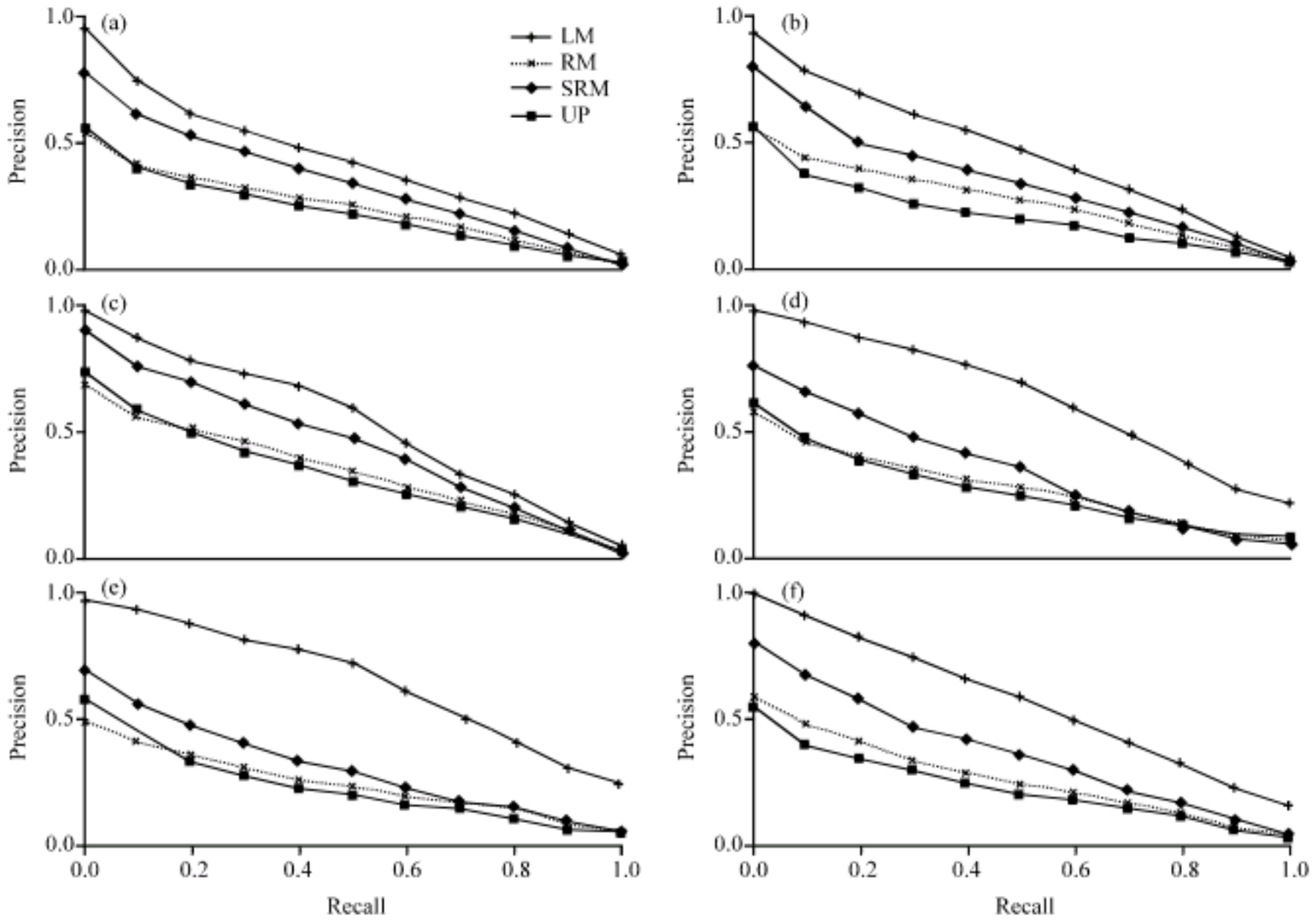

Fig. 3: Comparison of retrieval performance in training and testing. In each plot, the RM and SRM methods are compared with the basic language model. (a) Training on AP collection, queries 51-150, (b) test on AP collection, queries 151-200, (c) test on WSJ collection, queries 151-200, (d) test on FT collection, queries 301-400, (e) test on LA collection, queries 301-400 and (f) test on SJMN collection, queries 51-150

Table 5: Comparison of topic distance

	Avg topic distance		Count of topics					
Collection	SRM	RM	LB	SW	LW	SB	LE	SE
AP								
All	5.14	4.12	43	0	2	4	0	1
Rel	1.40	1.57	17	1	1	30	0	1
WSJ								
All	4.23	3.86	31	3	3	13	0	0
Rel	1.39	1.73	10	2	4	34	0	0
SJM								
All	3.84	3.16	67	2	16	7	2	0
Rel	1.26	1.28	35	9	9	39	0	2
FT								
All	4.55	3.32	60	1	24	3	6	1
Rel	1.58	1.52	27	13	12	36	2	5
LA								
All	5.16	4.02	63	7	16	7	4	1
Rel	1.60	1.42	39	11	12	31	1	4

respectively. B, W and E meant that $Topic_{SRM}$ has better, worse or equal MAP value than $Topic_{RM}$, respectively.

We observed that $Topic_{SRM}$ had larger average topic distance to $Coll_{all}$ than $Topic_{RM}$ (for example, 5.14 vs. 4.12 on AP collection), which indicated that our expansion terms were far from random topic than that of $Topic_{RM}$. Meanwhile, $Topic_{SRM}$ had smaller average topic distance to $Coll_{rel}$ than $Topic_{RM}$ on collection AP, WSJ and SJMN (1.40 vs. 1.57, 1.39 vs. 1.73 and 1.26 vs. 1.28), which indicated that our expansion terms were closer to the true topic than that of $Topic_{RM}$. This demonstrated that the feedback documents selected by our method were closer to the true relevance documents. This implied that $Topic_{SRM}$ should get better performance on AP, WSJ and SJMN collections than $Topic_{RM}$ and the results in Fig. 3 were consistent with this prediction. However, $Topic_{SRM}$ obtained larger average topic distance to $Coll_{rel}$ on FT and LA collections than $Topic_{RM}$ (1.58 vs. 1.52 and 1.60 vs. 1.42). Although, this was different from what we obtained from other collections, it was consistent with the results shown in Fig. 3. That is, the performance of SRM on the FT collection had no improvement when recall was above 0.6 and the performance of SRM on the LA collection showed less improvement over RM than those on other collections.

We also observed certain correlation relationship between the topic distance and MAP. For example, our results showed that when examining the distance to random topic $Coll_{all}$ on AP collection, the results from

43 topics (86% of total 50 topics, the summation of columns LB and SW) were consistent with the correlation between the topic distance and MAP, but that of 7 topics (14% of total 50 topics, summation of columns LW, SB and SE) were not. From other collections such as WSJ, SJMN, FT and LA collections, the percentages of correlated topics were 68, 73.4, 70.53 and 71.4%, respectively. If examining the distance to $Coll_{rel}$ on the five collections: AP, WSJ, SJMN, FT and LA, the percentages of the correlated topics (the summation of columns LW and SB) were: 62, 76, 51, 50.5, 33.7%, respectively. Here, we saw that the LA collection was a problematic collection with only 33.7% correlation rate.

RESULTS AND DISCUSSION

Retrieval performance: Present experiments showed positive retrieval results of our SRM method. Table 3 and Fig. 3 show that our SRM method works better than LM and RM at low recall and high precision side in all cases, which makes our semantic clustering based relevance model an attractive choice in high precision applications. This indicates that many of documents in our semantic cluster are relevant to the query. We analyzed the performance improvement may be contributed by the difference between the SRM method and the RM method: (1) RM method that using all the top-ranked initial retrieved documents as feedback documents will inevitably add non-relevant documents to the estimation process of the relevance model. Our SRM method benefited from the topic separation capability of ICA which tries to group relevance documents into same semantic cluster, (2) RM method estimated relevance model based only on document models whereas SRM method took the semantic cluster information into consideration by estimating semantic models that will indicate how a document was generated from a semantic cluster model, (3) RM method viewed each document equal role in relevance model estimation whereas SRM method differentiated the contribution of each document to the estimation of relevance model, referred to Eq. 13. (4) RM method just used a collection model to smooth the document model, but it's not optimal to estimate a probability distribution of an unseen term in a document by using the same collection model. SRM used the semantic keywords cluster along with the collection model to smooth the document model and the semantic keywords can enhance the modeling of a document that is related to the semantic topic.

However, Fig. 3 also shows that when high recall (e.g., larger than 0.7) is needed, our method, like other methods in comparison, brought much noise in the feedback so that the precision of the results dropped fast. Besides of the problem of parameter tuning, a better ICA algorithm that can better group the documents in latent semantic space may be required.

Comparing with other cluster-based retrieval method, Table 4 shows that SRM method has better retrieval performance over the other three methods. We think the main reason comes from the different clustering methods: (1) SRM method grouped documents semantically in a latent semantic space that will bring semantic information into estimation of relevance model, like the document prior probability and the semantic keywords for a cluster. But other cluster based methods can utilize such kinds of semantic information from cluster, (2) other cluster based methods smoothed the language model by counting an unseen term in a cluster and using a collection model for smoothing. Although comparing to the traditional smoothing methods, these methods used a cluster to smooth in addition, there is no topic-related semantic information to employ. SRM method used the probability of a term belonging to a semantic cluster, referred to Eq. 14 and 19 and (3) other cluster based methods treated a cluster as a large document model, SRM treated a cluster as a semantic model that is responsible for the documents and terms generation.

Although, SRM had obtained significant improvements over other methods, there is still a large gap to its upper bound and the difference is still statistical significant (Table 3, 4 and Fig. 3). This tells us two important messages: (1), our semantic clustering method, although had demonstrated its effectiveness, still either missed true relevance documents or selected non-relevant documents during the feedback process; (2), the significant better performance of the upper bound over SRM demonstrates that there is still a large room for further improvement by SRM method if we concentrate on better selection of topic-related documents.

Topic distance: Topic distance experiment showed the topic estimated from our SRM method generally has closer distance to the true topic and larger distance to the collection random topic. That indicates that our SRM method captures true relevant documents for feedback, whereas RM method which directly used all top-ranked retrieved documents as relevance feedback documents brings much noise into feedback.

The SRM obtained larger average topic distances to the true topic on FT and LA collections than RM did. It explained that the performance of SRM on the FT and LA collections has less improvement over RM than that on the other collections, especially in the case of high recall. We think that may be caused by: (1) the relevant

documents per topic in FT and LA collections are the lowest among the five collections. Less relevant documents would make search topics difficult and thus make PRF less effective and (2) training feedback parameters cannot fit all different collections well.

Although, the topic distance of SRM to the true topic was larger than that of RM, SRM still showed the significant performance improvement over the RM method. We think that it is because the feedback documents of SRM method were chosen from the activation process of user's query, which could filter out noise document from top ranked initial retrieval documents by semantic cluster information. The semantic information actually helped to form document models that are closer to the query model. For example, the document prior probability can contribute differently to the document model and the keywords of semantic cluster can keep feedback topic from random non-relevant topics, but corresponding information lacks in the RM method.

CONCLUSIONS

By viewing a document as an interaction of a set of independent hidden topics, we proposed a semantic clustering technique using independent component analysis and applied the semantic topic clusters selected by a query activation process to estimate a relevance language model without requiring relevance training information.

Based on the positive experiment results on five different TREC collections, we can draw the following conclusions. (1) it is helpful to use semantic clustering to cure the topic drifting problem in PRF, (2) estimating a document language model from documents belonging to activated semantic clusters is more effective than estimating blindly from all top-ranked retrieved documents and (3) our semantic clustering based relevance language model is a valid and robust method for PRF and we think that the main contribution of the retrieval improvement comes from the estimation of relevance model on the semantic cluster that is most similar to the query.

There are further directions to improve our semantic clustering based relevance model, (1) a more stable number estimation of independent components in data, which can reveal the real topic structure in data, should be focused on, (2) intuitively, the prior probability of a semantic cluster could play a positive role in the relevance model estimation because it differentiates the semantic importance of each semantic cluster and (3) we want to design a model estimation process with less or no parameters tuning and expect such a model would come closer to the upper-bound performance.

ACKNOWLEDGMENTS

This study was partially supported by China Scholarship Council, the University of Pittsburgh and NSF under Grant NSF/IIS 0704628.

REFERENCES

Amari, S.I., 1998. Natural gradient works efficiently in learning. Neural Comput., 10: 251-276.

Deerwester, S., S.T. Umais, G.W. Furnas, T.K. Landauer and R. Harshman, 1990. Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci. Technol., 41: 391-407.

Efthimiadis, E.N., 2000. Interactive query expansion: a user-based evaluation in a relevance feedback environment. J. Am. Soc. Inform. Sci. Technol., 51: 989-1003.

Hansen, L.K., J. Larsen and T. Kolenda, 2001. Blind detection of independent dynamic components. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-11, Salt Lake City, Utah, USA., pp: 3197-3200.

Hofmann, T., 1999. Probabilistic latent semantic indexing. Proceedings of the 22th ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, Berkeley, LA, USA., pp: 10.1-1.33.

Hyvärinen, A., 1999. Survey on independent component analysis. Neural Comput. Surveys, 2: 94-128.

Kolenda, T., 2002. Adaptive Tools in Virtual Environments: Independent Component Analysis for Multimedia. Informatics and Mathematical Modelling, Technical University of Denmark, Kansas City, Missouri, USA.

Kolenda, T., L.K. Hansen, O. Winther and S. Sigurdsson, 2002. Dtu: Toolbox. Informatics and Mathematical Modeling, Technical University of Denmark, Kansas City, Missouri, USA.

Lafferty, J. and C. Zhai, 2001. Document language models, query models and risk minimization for information retrieval. Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, Sept. 9-13, New Orleans, LA, USA., pp: 111-119.

Lautrup, B., L.K. Hansen, I. Law, N. Morch, C. Svarer and S.C. Strother, 1995. Massive weight sharing: A cure for extremely ill-posed problems. Proceedings of the Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks, (WSBRFTNN'95), New York, USA., pp: 137-148.

Lavrenko, V. and W.B. Croft, 2001. Relevance-based language models. Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval, Sept. 9-13, New Orleans, LA, USA., pp: 120-127.

Lee, K.S., W.B. Croft and J. Allan, 2008. A cluster-based resampling method for pseudo-relevance feedback. Proceedings of the 31th ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 20-24, Singapore, pp: 235-242.

Liu, X. and W.B. Croft, 2004. Cluster-based retrieval using language models. Proceedings of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 25-29, Sheffield, UK., pp: 186-193.

Ponte, J.M. and W.B. Croft, 1998. A language modeling approach to information retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 24-28, Melbourne, Australia, pp: 275-281.

Sakai, T., T. Manabe and M. Koyama, 2005. Flexible pseudo relevance feedback via selective sampling. ACM Trans. Asian Language Inform. Proc., 4: 111-135.

Wang, X., H. Fang and C. Zhai, 2008. A study of methods for negative relevance feedback. Proceedings of the 31th ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 20-24, Singapore, pp: 219-226.

Wei, X. and W.B. Croft, 2006. LDA-based document models for ad-hoc retrieval. Proceedings of the 29th ACM SIGIR Conference on Research and Development in Information Retrieval, (CRDIR'06), Seattle, WA., pp: 178-185.

Zhai, C. and J. Lafferty, 2001. Model-based feedback in the language modeling approach to information retrieval. Proceedings of the 10th International Conference on Information and Knowledge Management, Oct. 05-10, ACM, New York, USA., pp: 403-410.

Zhai, C., 2008. Statistical language models for information retrieval: A critical review. Found. Trends Inform. Retrieval, 2: 137-213.