Integrating Clustering and Multi-Document Summarization by Bi-Mixture Probabilistic Latent Semantic Analysis (PLSA) with Sentence Bases

Chao Shen and **Tao Li**
School of Computing and Information Sciences
Florida International University
Miami, Florida 33199
{*cshen001,taoli*}@*cs.fiu.edu*

Chris H. Q. Ding
Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX 76019
chqding@uta.edu

Abstract

Probabilistic Latent Semantic Analysis (PLSA) has been popularly used in document analysis. However, as it is currently formulated, PLSA strictly requires the number of word latent classes to be equal to the number of document latent classes. In this paper, we propose Bi-mixture PLSA, a new formulation of PLSA that allows the number of latent word classes to be different from the number of latent document classes. We further extend Bi-mixture PLSA to incorporate the sentence information, and propose Bi-mixture PLSA with sentence bases (Bi-PLSAS) to simultaneously cluster and summarize the documents utilizing the mutual influence of the document clustering and summarization procedures. Experiments on real-world datasets demonstrate the effectiveness of our proposed methods.

Introduction

Document clustering and multi-document summarization are two fundamental tools for understanding document data. Probabilistic Latent Semantic Analysis is a widely used method for document clustering due to the simplicity of the formulation, and efficiency of its EM-style computational algorithm. The simplicity makes it easy to incorporate PLSA into other machine learning formulations. There are many further developments of PLSA, such as Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) and other topic models see review articles (Steyvers and Griffiths 2007; Blei and Lafferty 2009). The essential formulation of PLSA is the expansion of the co-occurrence probability $P(word, doc)$ into a latent class variable z that separates word distributions from the document distributions given latent class. However, as it is currently formulated, PLSA strictly requires *the number of word latent classes to be equal to the number of document latent classes* (i.e., there is a one-to-one correspondence between word clusters and document clusters). In practical applications, however, this strict requirement may not be satisfied since if we consider documents and words as two different types of objects, they may have their own cluster structures, which are not necessarily same, though related.

Recently, an extension of PLSA, called "Factorization by Given Bases"(FGB), is proposed to simultaneously cluster and summarize documents by making use of both the document-term and sentence-term matrices (Wang et al. 2008b). By formulating the clustering-summarization problem as a problem of minimizing the Kullback-Leibler divergence between the given documents and the model reconstructed terms, the model essentially performs co-clustering on document and sentences. However, one limitation in the model is that *the number of document clusters is equal to the number of sentence clusters*.

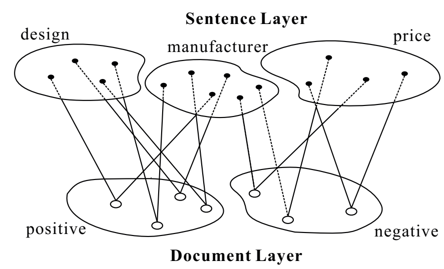


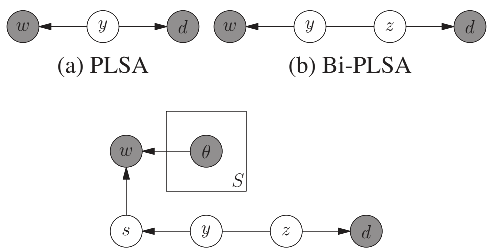
Figure 1: An example showing different cluster structures of documents and sentences.

In many applications, the sentences in the documents may have their own cluster structures, which may be different from the document cluster structures. An example is shown in Figure 1 where a set of product reviews are divided into two clusters: *positive* reviews and *negative* reviews, while the sentences are grouped into three clusters: *design*, *price* and *manufacturer information*. However, these two layers of cluster structures are related since each sentence cluster has its own distribution w.r.t document clusters, and vice versa. Hence, there exists mutual influence between these two layers of clustering.

Motivated by the above analysis, in this paper, we first propose a new formulation of PLSA that allows *the number of latent word classes to be different from the number of latent document classes*. Because our formulation resembles mixtures of different type classes, we call it "Bi-mixture PLSA" (Bi-PLSA). Then based on Bi-PLSA, we incorporate sentence information and propose a new model, *Bi*-mixture *PLSA* with *S*entence bases (Bi-PLSAS), extending from co-clustering of documents and words to co-clustering of documents and sentences. The new model simultaneously clusters the documents and sentences, and utilizes the mutual influence to improve the clustering of both layers. Meanwhile, an extractive summary composed of representative sentences for each sentence cluster can be easily produced. As a result, Bi-PLSAS leads to 1) a better document cluster method utilizing sentence information, and 2) an effective document summarization method taking the document context information into consideration.

In the following, we first describe the details of the proposed Bi-PLSA and Bi-PLSAS model. Then an illustrative example is given to demonstrate our proposed models, followed by the theoretical analysis. Finally, experimental results on document clustering and multi-document summarization are presented to evaluate the effectiveness of these models.

Bi-mixture PLSA



(a) PLSA (b) Bi-PLSA

(c) Bi-PLSAS

Figure 2: The Graphical Models

In PLSA, the joint probability distribution of a word and a document, $p(w,d)$ can be decomposed as

$$p(w,d) = \sum_z p(w,d|z)p(z) = \sum_z p(w|z)p(z)p(d|z), \tag{1}$$

assuming that given latent class z, the word distribution and document distribution are independent. Its graphical model is shown in Figure 2a.

We generalize PLSA by introducing two latent class variables z_w, z_d where z_w indicates word class and z_d indicates document class. Under the similar assumption as PLSA that the word distribution and document distribution are independent given the corresponding class variant, the joint probability distribution is decomposed as

$$\begin{aligned} p(w,d) &= \sum_{z_w,z_d} p(w,d|z_w,z_d)p(z_w,z_d) & (2)\\ &= \sum_{z_w,z_d} p(w|z_w)p(d|z_d)p(z_w,z_d). & (3)\end{aligned}$$

The graphical model is shown in Figure 2b.

Relation to Nonnegative Matrix Factorization (NMF)

The bi-mixture model is motivated by our earlier work in proving PLSA is equivalent to NMF with I-divergence (Ding, Li, and Peng 2006) where we model word-document matrix $X = (X_{wd})$, (X_{wd} indicates the frequency of word w in document d), as $X = FG^T$. (F, G) are obtained by minimizing

$$I_{div} = \sum_{wd} \left[X_{wd} \log \frac{X_{wd}}{(FG^T)_{wd}} - X_{wd} + (FG^T)_{wd} \right] \tag{4}$$

and FG^T can always be expressed as $FG^T = \widetilde{\mathbf{F}} S \widetilde{\mathbf{G}}^T$ where $\sum_w \widetilde{\mathbf{F}}_{wk} = 1, \sum_d \widetilde{\mathbf{G}}_{dk} = 1$ and S is a diagonal matrix satisfying $\sum_k S_k = 1$. We have the following correspondence between NMF and PLSA of Eq.(1):

$$F_{wk} = p(w|z_k),\ G_{dk} = p(d|z_k),\ S_k = p(z_k). \tag{5}$$

The tri-factorization model (tri-NMF) (Ding et al. 2006) models the data $X = FSG^T$, where S is a $K \times L$ matrix and has been widely used for co-clustering. The tri-NMF model motivates us to generalize PLSA to Bi-PLSA. The correspondence between tri-NMF and the bi-mixture model of Eq.(2) is

$$\begin{aligned} &F_{wk} = p(w|z_w = k),\ G_{dl} = p(d|z_d = l),\\ &S_{kl} = p(z_w = k, z_d = l). \end{aligned} \tag{6}$$

Bi-mixture PLSA with Sentence Bases

In this section, we extend Bi-PLSA to incorporate sentence information. The advantage of sentences over words is that sentences are more readable, e.g. in extractive summarization methods they are directly used as a summary, while nontrivial extra work is needed to interpret the word clusters, particularly in the form of unigram distributions (Mei, Shen, and Zhai 2007).

Since we are more interested in the sentence clustering and hope to utilize the sentence information to help document clustering, we replace the z_w with z_s, a latent class variable indicating sentence class. To generate a word, instead of generating it directly from the class variable as in PLSA, we assume it be generated from a hidden summary sentence. Specifically, first a sentence class is generated, then based on the sentence class a sentence s is selected, which can be taken as a summary of the class, and finally a word is generated from the summary sentence selected. Note that here value of s is not necessarily the sentence in which the word actually belongs to, but can be any index of a sentence in the document set, so s is a hidden variable, indicating a representative sentence (summary) of the sentence class to generate the target word. To generate a word from a sentence, for each sentence s, a language model θ_s is trained on it beforehand, and all these language models are called as sentence bases, where words are generated from. The graphical model in Figure 2c illustrates this procedure. The joint probability distribution of a word and a document, $p(w,d)$, is then decomposed as

$$p(w,d) = \sum_{z_s,z_d} \sum_s p(w|\theta_s)p(s|z_s)p(d|z_d)p(z_s,z_d) \tag{7}$$

Bi-PLSAS Algorithm

For notation simplicity, we set

$$\begin{aligned} F_{ik} &= p(s = i|z_s = k), & G_{jk} &= p(d = j|z_d = l),\\ S_{kl} &= p(z_s = k, z_d = l), & B_{hi} &= p(w = h|\theta_i). \end{aligned} \tag{8}$$

Given a document collection, we have the document-word matrix $X = (X_{wd})$, where X_{wd} indicates the frequency of word w in document d. B is a set of sentence bases, each of which corresponds to a column, indicating the word generating distribution from a sentence. B is estimated on the sentences with Dilichlet smoothing beforehand. With the input X and B, the parameters of the Bi-PLSAS model, (F, S, G) are computed by the following iterative algorithm.

(A0) Initialize F, S, G to a proper initial solution (F^0, S^0, G^0).

Iteratively update the solution using Steps (A1) and (A2) until convergence.

(A1) Compute the posterior probability $Q_{hj}^{ikl} \equiv P(z_s = k, z_d = l, s = i | w = h, d = j)$ as

$$Q_{hj}^{ikl} = \frac{B_{hi} F_{ik} S_{kl} G_{jl}}{(BFSG^T)_{hj}}. \tag{9}$$

(A2) Compute new F, G, S as:

$$\begin{aligned} F_{ik} &= \frac{\sum_{hjl} X_{hj} Q_{hj}^{ikl}}{\sum_{hijl} X_{hj} Q_{hj}^{ikl}}, G_{jl} = \frac{\sum_{hik} X_{hj} Q_{hj}^{ikl}}{\sum_{hijk} X_{hj} Q_{hj}^{ikl}}, \\ S_{kl} &= \frac{\sum_{hij} X_{hj} Q_{hj}^{ikl}}{\sum_{hijkl} X_{hj} Q_{hj}^{ikl}}. \end{aligned} \tag{10}$$

This algorithm is essentially an EM-type algorithm. We derive the algorithm below.

Derivation Of the Algorithm

The log-likelihood of the model on the document collection can be written as

$$\ell(F, S, G) = \sum_j \sum_h X_{hj} \log(BFSG^T)_{hj}. \tag{11}$$

Introducing the variables Q_{hj}^{ikl}, the objective function becomes

$$\ell(F, S, G) = \sum_j \sum_h X_{hj} \log \left(\sum_l \sum_k \sum_i Q_{hj}^{ikl} \frac{B_{hi} F_{ik} S_{kl} G_{jl}}{Q_{hj}^{ikl}} \right). \tag{12}$$

Using Jensen's Inequality we obtain a lower bound b as

$$\ell(F, S, G) \geq b(Q, F, S, G), \tag{13}$$

where

$$b(Q, F, S, G) \equiv \sum_j \sum_h X_{hj} \sum_l \sum_k Q_{hj}^{ikl} \log \left(\frac{F_{ik} S_{kl} G_{jl}}{Q_{hj}^{ikl}} \right). \tag{14}$$

The posterior probability Q and model parameters (F, S, G) are obtained as maximizing $b(Q, F, S, G)$ for one variable while fixing others.

Learning Q:

Q is obtained via fixing F, S, G,

$$P_Q : \max_Q b(Q, F, S, G) \quad s.t. \quad \sum_{i=1}^M \sum_{k=1}^K \sum_{l=1}^L Q_{hj}^{ikl} = 1. \tag{15}$$

Learning F: F is obtained via fix Q, S, G,

$$P_F : \max_F b(Q, F, S, G) \quad s.t. \quad \sum_{i=1}^M F_{ik} = 1. \tag{16}$$

Learning S: S is obtained via fix Q, F, G,

$$P_S : \max_S b(Q, F, S, G) \quad s.t. \quad \sum_{l=1}^L \sum_{k=1}^K S_{kl} = 1. \tag{17}$$

Learning G:

G is obtained via fixing Q, F, S

$$P_G : \max_G b(Q, F, S, G) \quad s.t. \quad \sum_{d=1}^N G_{dl} = 1. \tag{18}$$

Theorem 1 *Optimization of Eq.(15)-Eq.(18) has the optimal solutions as shown in Eq.(9) and Eq.().*

Theorem 2 *Convergence of the Algorithm. Starting with an initial solution (F^0, S^0, G^0), if we iteratively update the solution using Steps (A1) and (A2), obtaining*

$$\begin{aligned} &(F^0, S^0, G^0), (F^1, S^1, G^1), \cdots (F^t, S^t, G^t), \cdots \\ &\ell(F^t, S^t, G^t). \leq \ell(F^{t+1}, S^{t+1}, G^{t+1}). \end{aligned} \tag{19}$$

The proofs of the Theorem 1 and Theorem 2 are omitted due to the space limit.

Clustering and Summarization via Bi-PLSAS

Once we obtain the parameters $p(s|z_s)$, $p(d|z_d)$ and $p(z_d, z_s)$ in the Bi-PLSAS model , we can easily cluster the documents and sentences, and generate the summary.

Clustering The cluster membership of a document d can be obtained by

$$\begin{aligned} z(d)^* &= \arg\max_{z_d} p(z_d|d) \\ &= \arg\max_{z_d} p(z_d, d) \\ &= \arg\max_{z_d} p(d|z_d) \textstyle\sum_{z_s} p(z_d, z_s). \end{aligned} \tag{20}$$

Similarly, the cluster membership of a sentence s can be derived using

$$z(s)^* = \arg\max_{z_s} p(s|z_s) \sum_{z_d} p(z_d, z_s) \tag{21}$$

Summarization To generate a summary for the document collection, first, the marginal probability of every sentence cluster z_s is calculated as $p(z_s) = \sum_{z_d} p(z_s, z_d)$, and those clusters with small marginal probability values are removed. Then, the sentences are extracted from the remaining sentence clusters based on $p(s|z_s)$.

An Illustrative Example

Table 1 presents an example dataset of Apple product reviews. The dataset contains four documents, each of which is composed of two sentences. The first two documents are positive reviews about Apple's revolutionary design, while the last two documents are negative reviews about the price. Note that all four documents contain generic background information about Apple.

Figure 3 shows the typical experimental results using Bi-PLSA, FGB, and Bi-PLSAS. We can observe that Bi-PLSA actually wrongly clusters D1, D4 together and D2,D3 together, since D2, as a whole document, has more words

$$
\begin{array}{cc}
 & \begin{array}{cc} topic1 & topic2 \end{array} \\
\begin{array}{c} D1 \\ D2 \\ D3 \\ D4 \end{array} & \begin{pmatrix} 0.0 & 0.47 \\ 0.50 & 0.0 \\ 0.50 & 0.0 \\ 0.0 & 0.53 \end{pmatrix}
\end{array}
$$

(a) Bi-PLSA

$$
\begin{array}{cc}
 & \begin{array}{cc} topic1 & topic2 \end{array} \\
\begin{array}{c} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \end{array} & \begin{pmatrix} 0.1345 & 0.1521 \\ 0.1683 & 0 \\ \mathbf{0.2721} & 0.0220 \\ 0.2597 & 0 \\ 0.0384 & 0.2371 \\ 0 & \mathbf{0.2941} \\ 0.1270 & 0.0979 \\ 0 & 0.2023 \end{pmatrix}
\end{array}
\qquad
\begin{array}{cc}
 & \begin{array}{cc} topic1 & topic2 \end{array} \\
\begin{array}{c} D1 \\ D2 \\ D3 \\ D4 \end{array} & \begin{pmatrix} 1.0 & 0.0 \\ 1.0 & 0.0 \\ 0.0 & 1.0 \\ 0.0 & 1.0 \end{pmatrix}
\end{array}
$$

(b) FGB

$$
\begin{array}{cc}
 & \begin{array}{ccc} topic1_S & topic2_S & topic3_S \end{array} \\
\begin{array}{c} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \end{array} & \begin{pmatrix} 0.1348 & 0.1362 & 0.2144 \\ 0.0279 & 0.2565 & 0.0001 \\ \mathbf{0.4980} & 0.0749 & 0.0144 \\ 0.0160 & \mathbf{0.3654} & 0.0001 \\ 0.2326 & 0.0210 & 0.2282 \\ 0.0003 & 0.0002 & \mathbf{0.3011} \\ 0.0902 & 0.1458 & 0.0320 \\ 0.0002 & 0.0001 & 0.2097 \end{pmatrix}
\end{array}
\qquad
\begin{array}{cc}
 & \begin{array}{cc} topic1_D & topic2_D \end{array} \\
\begin{array}{c} D1 \\ D2 \\ D3 \\ D4 \end{array} & \begin{pmatrix} 0.4192 & 0.0 \\ 0.5766 & 0.0 \\ 0.0 & 0.5517 \\ 0.0039 & 0.4479 \end{pmatrix}
\end{array}
$$

(c) Bi-PLSAS

Figure 3: Results of Bi-PLAS, FGB and Bi-PLSAS on the example dataset. Bold numbers indicate the corresponding sentences are selected as the representatives for the associated clusters. In (a), word clustering result of Bi-PLSA is omitted due to the space limit. In (b), document clusters and sentence clusters are the same, referred as $topic1$ and $topic2$. In (c), there are three sentence clusters: $topic1_S$, $topic2_S$ and $topic3_S$, and two document clusters: $topic1_D$, $topic2_D$.

D1	S1	Apple is a corporation manufacturing consumer electronics.
	S2	Apple is a lot more revolutionary to most American.
D2	S3	Apple is an American company focusing consumer electronics.
	S4	The design of Apple products is more revolutionary than others.
D3	S5	Apple is an company focusing consumer electronics.
	S6	The price of Apple products are high even to American.
D4	S7	Apple is a corporation manufacturing consumer electronics.
	S8	With the performance, Apple price is high.

Table 1: An example dataset of four documents and eight sentences.

overlapping with D3 than with D1. Both FGB and Bi-PLSAS, which utilize the sentence information, can cluster the documents correctly. However, with the restriction that the number of sentence clusters should be the number of document clusters, FGB can only cluster the sentences into two groups, one of which has an incorrect representative sentence S3. On the contrary, Bi-PLSAS can group the sentences into three clusters: *company information*, *design* and *price*, each with the right representative sentence.

Theoretical Analysis of PLSA Algorithms

First, Bi-PLSAS model contains Bi-PLSA and the standard PLSA as special cases. By setting $B = I$, the Bi-PLSAS model reduces to Bi-PLSA model. Further restricting S to diagonal, Bi-PLSA becomes the standard PLSA. Therefore, the algorithm in Eqs.(9,10) is the generic algorithm for these PLSA models.

Now, we prove a fundamental theorem about these PLSA algorithms.

Theorem 3. In each iteration of the PLSA algorithm of Eqs.(9,10), the marginal distributions are preserved, i.e,

$$\sum_h (BFSG^T)_{hj} = \sum_h X_{hj} / \sum_{hj} X_{hj},\ \forall\ j, \tag{22}$$

$$\sum_j (BFSG^T)_{hj} = \sum_j X_{hj} / \sum_{hj} X_{hj},\ \forall\ h. \tag{23}$$

Proof. Due to the normalization $\sum_h B_{hi} = 1$, we have

$$\sum_h (BFSG^T)_{hj} = \sum_i (FSG^T)_{ij} = \sum_{ikl} F_{ik} S_{kl} G_{jl}. \tag{24}$$

Since $\sum_i F_{ik} = 1$, we evaluate $\sum_{kl} S_{kl} G_{jl}$, which is

$$\sum_{kl} \frac{\sum_{hij} X_{hj} Q_{hj}^{ikl}}{\sum_{hijkl} X_{hj} Q_{hj}^{ikl}} \frac{\sum_{hik} X_{hj} Q_{hj}^{ikl}}{\sum_{hijk} X_{hj} Q_{hj}^{ikl}} = \sum_l \frac{\sum_{hik} X_{hj} Q_{hj}^{ikl}}{\sum_{hijkl} X_{hj} Q_{hj}^{ikl}}.$$

Because $\sum_{ikl} Q_{hj}^{ikl} = 1$, thus we recover Eq.(22). Eq.(23) can be similarly proved. QED.

Equivalence between Bi-mixture PLSA and tri-NMF

Here we provide important properties of tri-NMF model using I-divergence and show it is equivalent to the bi-mixture PLSA. The tri-NMF model parameters F, S, G are obtained by minimizing the I-divergence between X_{wd} and $(FSG^T)_{wd}$:

$$I_{div} = \sum_{wd} \Big[X_{wd} \log \frac{X_{wd}}{(FSG^T)_{wd}} - X_{wd} + (FSG^T)_{wd} \Big]. \tag{25}$$

The relation between this tri-NMF model and the bi-mixture model is characterized by the correspondence of Eq.(6). Let $F_{w=i,k} = F_{ik}$ and $G_{d=j,l} = G_{jl}$, one can easily derive the following updating rules:

$$F_{ik} \longleftarrow F_{ik} \sum_j X_{ij} \frac{(SG^T)_{kj}}{(FSG^T)_{ij}} \Big/ \sum_{j'} (SG^T)_{kj'} \tag{26}$$

$$G_{jl} \longleftarrow G_{jl} \sum_i X_{ij} \frac{(FS)_{il}}{(FSG^T)_{ij}} \Big/ \sum_{i'} (FS)_{i'l}, \tag{27}$$

$$S_{kl} \longleftarrow S_{kl} \sum_{ij} X_{ij} \frac{F_{ik} G_{jl}}{(FSG^T)_{ij}} \Big/ \sum_{i'j'} F_{i'k} G_{j'l}. \tag{28}$$

Now we prove the following theorem:

Theorem 4. At convergence, the solution of tri-NMF preserves the marginal distributions:

$$\sum_i (FSG^T)_{ij} = \sum_i X_{ij},\ \forall\, j \tag{29}$$

$$\sum_j (FSG^T)_{ij} = \sum_j X_{ij}.\ \forall\, i \tag{30}$$

Proof. We have $\sum_i (FSG^T)_{ij} = \sum_{il} (FS)_{il} G_{jl}$. Now at convergence, Eq.(27) becomes equality. Substituting the RHS as G_{jl}, we have

$$\begin{aligned} & \sum_i (FSG^T)_{ij} \\ = \ & \sum_{i_1 l} (FS)_{i_1 l}\Big(G_{jl}\sum_i X_{ij}\frac{(FS)_{il}}{(FSG^T)_{ij}} \Big/ \sum_{i'}(FS)_{i'l}\Big) \\ = \ & \sum_l G_{jl}\sum_i X_{ij}\frac{(FS)_{il}}{(FSG^T)_{ij}} = \sum_i X_{ij}\frac{\sum_l G_{jl}(FS)_{il}}{(FSG^T)_{ij}}. \end{aligned}$$

The nominator and denominator cancel out. Thus we recover Eq.(29). Eq.(30) can be similarly proved. QED.

Theorem 4 ensures the preservation of marginal distribution of tri-NMF. This is useful for probability interpretation of the tri-NMF model. More importantly, this property ensures that the 2nd and 3rd terms in Eq.(25) are equal. Thus the I-divergence objective function is equivalent to the KL-divergence, indicating that tri-NMF has the same objective as bi-mixture PLSA. We note that, however, the detailed algorithms (Eqs.(9,10) for Bi-PLSA and Eqs.(26,27,28) for tri-NMF) are different: starting from the same initial solution, these two algorithms will converge to different final solutions.

Experiments on Document Clustering

Experimental Setup

Dataset	#docs	#sentences	#word	#Doc Cluster	#Word Cluster
DBLP	552	13916	3000	9	11
CSTR	550	3134	2000	4	11
NG20	500	2744	2000	10	-

Table 2: Summary of datasets used in document clustering experiments.

We use the three datasets in our experiments: DBLP dataset, CSTR dataset, and a subset of 20 Newgroup dataset (Lang 1995). The first two datasets are from (Li et al. 2008). The characteristics of all three datasets are summarized in Table 2. To measure the clustering performance, we use purity and normalized mutual information (NMI)(Manning, Raghavan, and Schütze 2008) as our performance measures.

Comparison for Different Numbers of Sentence/Word Clusters

We first compare the proposed Bi-PLSA and Bi-PLSAS with the baselines: PLSA and FGB. To show the effect of two hidden class variables, we evaluate the performance of Bi-PLSA and Bi-PLSAS with different numbers of word/sentence clusters [1].

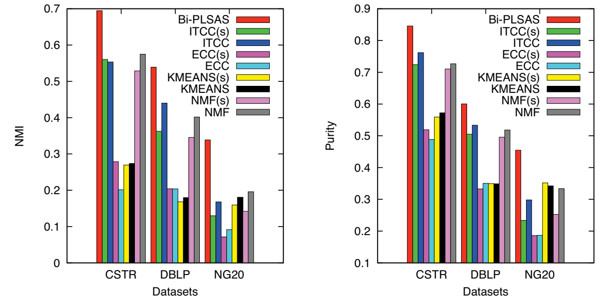


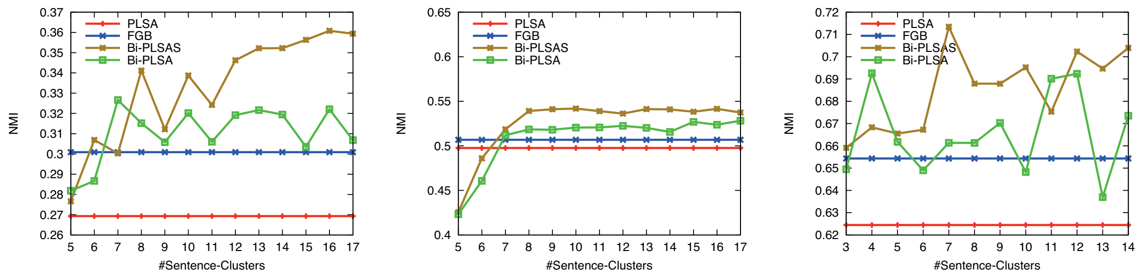
Figure 6: Comparison with (Co-)clustering methods. “(s)” at the end the method name indicates the method is conducted on the document-sentence matrix, instead of the document-word matrix.

From the Figure 4, we can see 1) Bi-PLSA outperforms PLSA for most numbers of word clusters, since the Bi-PLSA has freedom to set different value from the number of document clusters; 2) The better performance of FGB than PLSA demonstrates the effectiveness of sentence bases in document clustering; 3) Bi-PLSAS combines the advantage of FGB and Bi-PLSA, and performs the best among all the methods.

Comparison with (Co-)clustering Methods

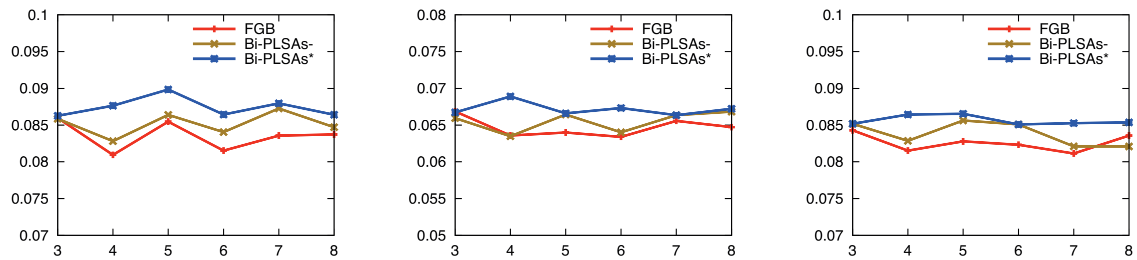
Here we compare Bi-PLSAS with (a) two co-clustering methods: ITCC (Dhillon, Mallela, and Modha 2003) (the Information-theoretic co-clustering algorithm) and ECC (Cho et al. 2004) (the Euclidean co-clustering algorithm) ; and (b) two classic document clustering methods: KM(the traditional K-means Algorithm) and NMF (Xu, Liu, and Gong 2003) (document clustering based on Nonnegative Matrix Factorization). Since Bi-PLSAS essentially conducts co-clustering on document and sentence sides, we also apply these competing methods on the document-sentence matrix, where the matrix entries indicate the similarities between documents and sentences. The number of word/sentence clusters is set to the true value of the number of word clusters for the DBLP and CSTR datasets, and the number of document clusters for the NG20 dataset. As shown in Figure 6, clustering on document-sentence matrix in most cases is worse than clustering directly on document-word matrix. This is because that directly co-clustering the document-sentence matrix is not effective in utilizing sentence information since the document-word and sentence-word relations are lost in the process. Our Bi-PLSAS model achieves best results by utilizing the mutual influence between document clustering and sentence clustering.

[1]Note that in PLSA or FGB, the number of word/sentence clusters is fixed to be the number of document clusters when the number of document clusters are given. Thus the performance of PLSA/FGB is just a constant. For comparison purpose, we plot them as two horizontal lines in Figure 4.



(a) NG20 (b) DBLP (c) CSTR

Figure 4: Comparison with baselines for different numbers of sentence/word clusters.



(a) DUC04 (b) DUC05 (c) DUC06

Figure 5: Comparison of Bi-PLSAS and FGB for variant numbers of document clusters.

Experiments on Document Summarization

Experiment Settings

	DUC04	DUC05	DUC06
Type of Summarization	Generic	Query-focused	Query-focused
#topics	NA	50	50
#documents per topic	10	25-50	25
Summary length	665 bytes	250 words	250 words

Table 3: Brief description of the datasets used in summarization.

In this section, experiments are conducted to demonstrate the effectiveness of the Bi-PLSAS on summarization tasks. One generic summarization dataset DUC04 and two query-focused summarization datasets, DUC05 and DUC06 are used in experiments[2]. The summary of the datasets is shown in Table 3. To conduct query-focused summarization, those sentences which do not contain any non-stopword term in the given query are first filtered out, then same as generic summarization, Bi-PLSAS model is first computed and sentences are then extracted based on the model to form the summary. For simplicity, we fix the number of document and sentence clusters. Number of document clusters is set to 4 for all datasets, and number of sentence clusters is set to 5 for DUC04 and 10 for DUC05 and DUC06. To automatically deciding the number of clusters, model selection criteria such as Akaike information criterion (AIC)(Akaike 1974) and Bayesian information criterion (BIC)(Schwarz 1978) can be used. ROUGE (Lin and Hovy 2003) toolkit (version 1.5.5) is used to measure the summarization performance.

[2]http://duc.nist.gov

Comparison with Different Methods

We compare the proposed method with following methods:

- LSA: conducts latent semantic analysis on terms by sentences matrix as proposed in (Gong and Liu 2001).
- KM: calculates sentence similarity matrix using cosine similarity and performs K-means algorithm to clustering the sentences and chooses the center sentences in each clusters.
- NMF: similar procedures as KM and uses NMF as the clustering method.
- DUCBest: the highest scores of the DUC participants.
- FGB: conducts document clustering and summarization simultaneously, using sentence language models as base language models (Wang et al. 2008b).

Several recent proposed systems in query-focused document summarization tasks are also included for performance comparison. They are:

- SingleMR: proposes a manifold-ranking based algorithm for sentence ranking (Wan, Yang, and Xiao 2007).
- MultiMR: uses multi-modality manifold-ranking method by utilizing within-document and cross-document sentence relationships as two separate modalities (Wan and Xiao 2009).

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LSA	0.34145	0.06538	0.12042
KM	0.34872	0.06937	0.12339
NMF	0.36747	0.07261	0.12961
FGB	0.38724	0.08115	0.13096
DUCBest	0.38224	0.09216	0.13325
Bi-PLSAS	0.38853	0.08764*	0.13112

(a) Generic summarization on DUC04

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LSA	0.30461	0.04079	0.10883
KM	0.31762	0.04938	0.10806
NMF	0.32026	0.05105	0.11278
FGB	0.34851	0.06243	0.12206
DUCBest	0.37978	0.07431	0.12979
SingleMR	0.36316	0.06603	0.12694
MultiMR	0.36909	0.06836	0.12877
SemanSNMF	0.35006	0.06043	0.12266
Bi-PLSAS	0.36028*	0.06769*	0.12587*

(b) Query focused summarization on DUC05

Systems	ROUGE-1	ROUGE-2	ROUGE-W
LSA	0.33078	0.05022	0.11220
KM	0.33605	0.05481	0.12450
NMF	0.33850	0.05851	0.12637
FGB	0.38712	0.08295	0.13371
DUCBest	0.41017	0.09513	0.14264
SingleMR	0.39534	0.08335	0.13766
MultiMR	0.40306	0.08508	0.13997
SemanSNMF	0.39551	0.08549	0.13943
Bi-PLSAS	0.39384*	0.08497	0.13852

(c) Query focused summarization on DUC06

Table 4: Comparison of the methods on Multi-Document Summarization (* indicates that the improvement of Bi-PLSAS model over the baseline FGB is statistically significant).

- SemanSNMF: uses semantic role analysis to constructs sentence similarity matrix, on which then symmetric non-negative matrix factorization is conducted to cluster sentences and finally selects the most important sentences in each cluster (Wang et al. 2008a).

Table 4 shows the ROUGE evaluation results on three datasets. From the results, we observe that: our method achieves high ROUGE scores and significantly improve the baseline FGB. Also the proposed method is comparable with newly developed summarizers which adopt various advanced techniques like semantic role analysis and manifold ranking.

Comparison Between Bi-PLSAS and FGB for Different Numbers of Document Clusters

In FGB, every sentence cluster corresponds to a document cluster, while in Bi-PLSAS, such restriction is removed and users can choose a different proper number of sentence clusters to generate the summary. Figure 5 shows the comparison between Bi-PLSAS and the baseline FGB with different numbers of document clusters. Bi-PLSAS* indicates using Bi-PLSAS with the best number of sentence clusters; Bi-PLSAS- indicates using Bi-PLSAS and the number of sentence clusters is set to be the number of document clusters. It can been seen that Bi-PLSAS together with a proper sentence cluster number can significantly outperform the baseline FGB, and even when the same sentence cluster number is used, Bi-PLSAS still outperforms FGB.

Conclusions

In this paper, we propose a new formulation of PLSA to incorporate the sentence information, allowing the number of latent sentence classes to be different from the number of latent document classes. We show that the new formulation with the modeling flexibility is useful for many applications such as document clustering and summarization. Experimental results on real-world datasets demonstrate the effectiveness of our proposal.

Acknowledgement

This work of C. Shen and T. Li is supported in part by NSF grants DMS-0915110, CCF-0830659, and HRD-0833093. The work of C. Ding is supported by NSF grants DMS-0915228 and CCF-0830780.

References

Akaike, H. 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* 19(6):716–723.

Blei, D., and Lafferty, J. 2009. Topic models. *Text mining: classification, clustering, and applications* 71.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*.

Cho, H.; Dhillon, I.; Guan, Y.; and Sra, S. 2004. Minimum sum-squared residue co-clustering of gene expression data. In *SDM*, 114–125.

Dhillon, I.; Mallela, S.; and Modha, D. 2003. Information-theortic co-clustering. In *SIGKDD*.

Ding, C.; Li, T.; Peng, W.; and Park, H. 2006. Orthogonal nonnegative matrix tri-factorizations for clustering. In *SIGKDD*.

Ding, C.; Li, T.; and Peng, W. 2006. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. In *AAAI*.

Gong, Y., and Liu, X. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR*, 19–25. ACM.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *ICML*.

Li, T.; Ding, C.; Zhang, Y.; and Shao, B. 2008. Knowledge transformation from word space to document space. In *SIGIR*.

Lin, C., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL-HLT*.

Manning, C.; Raghavan, P.; and Schütze, H. 2008. An introduction to information retrieval.

Mei, Q.; Shen, X.; and Zhai, C. 2007. Automatic labeling of multinomial topic models. In *SIGKDD*.

Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics* 6(2):461–464.

Steyvers, M., and Griffiths, T. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*.

Wan, X., and Xiao, J. 2009. Graph-based multi-modality learning for topic-focused multi-document summarization. *IJCAI*.

Wan, X.; Yang, J.; and Xiao, J. 2007. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI*, volume 7.

Wang, D.; Li, T.; Zhu, S.; and Ding, C. 2008a. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *SIGIR*, 307–314. ACM.

Wang, D.; Zhu, S.; Li, T.; Chi, Y.; and Gong, Y. 2008b. Integrating clustering and multi-document summarization to improve document understanding. In *CIKM*.

Xu, W.; Liu, X.; and Gong, Y. 2003. Document clustering based on non-negative matrix factorization. In *SIGIR*, 267–273. ACM.