Topic Modeling Ensembles

Zhiyong Shen, Ping Luo, Shengwen Yang, Xukun Shen

HP Laboratories
HPL-2010-158

Keyword(s):

Topic model, Ensemble

Abstract:

In this paper we propose a framework of topic modeling ensembles, a novel solution to combine the models learned by topic modeling over each partition of the whole corpus. It has the potentials for applications such as distributed topic modeling for large corpora, and incremental topic modeling for rapidly growing corpora. Since only the base models, not the original documents, are required in the ensemble, all these applications can be performed in a privacy preserving manner. We explore the theoretical foundation of the proposed framework, give its geometric interpretation, and implement it for both PLSA and LDA. The evaluation of the implementations over the synthetic and real-life data sets shows that the proposed framework is much more efficient than modeling the original corpus directly while achieves comparable effectiveness in terms of perplexity and classification accuracy.

External Posting Date: October 21, 2010 [Fulltext] Approved for External Publication
Internal Posting Date: October 21, 2010 [Fulltext]
To be published and presented at the Tenth IEEE International Conference on Data Mining, Sydney, Australia, December 14-17, 2010.

Topic Modeling Ensembles

Zhiyong Shen[1,2], Ping Luo[1], Shengwen Yang[1], Xukun Shen[2]
[1] Hewlett Packard Labs China,{zhiyongs,ping.luo,shengwen.yang}@hp.com
[2]State Key Laboratory of Virtual Reality Technology and system, China, xkshen@vrlab.buaa.edu.cn

***Abstract*—In this paper we propose a framework of *topic modeling ensembles*, a novel solution to combine the models learned by topic modeling over each partition of the whole corpus. It has the potentials for applications such as distributed topic modeling for large corpora, and incremental topic modeling for rapidly growing corpora. Since only the base models, not the original documents, are required in the ensemble, all these applications can be performed in a privacy preserving manner. We explore the theoretical foundation of the proposed framework, give its geometric interpretation, and implement it for both PLSA and LDA. The evaluation of the implementations over the synthetic and real-life data sets shows that the proposed framework is much more efficient than modeling the original corpus directly while achieves comparable effectiveness in terms of perplexity and classification accuracy.**

***Keywords*-Topic model, Ensemble**

I. Introduction

Recent years have witnessed an increasing interest on ensemble learning in the area of data mining and machine learning. The idea of ensemble is to combine the base models from multiple local data nodes to achieve the similar (or even better) effectiveness with the model learned from the whole data. Ensemble learning was firstly introduced to supervised methods, i.e. *ensembles of classifiers* [1]. Then, unsupervised ensemble learning techniques [2], [3] were proposed as *clustering ensembles*, as shown in Figure 1(a). It reconciles multiple clustering results ($\lambda^{(1)} \cdots \lambda^{(r)}$) of a data set into a single consolidated clustering result (λ), without accessing the original data. Illuminated by the success of ensemble learning for classification and clustering, in this paper we explore how to apply ensemble to topic modeling.

As a generalization of clustering, topic modeling, such as PLSA [4] and LDA [5], [6], has been successfully used for analyzing sparse vectors of count data, such as bag of words for documents, bag of features for images, or bag of activities for human daily routines. It provides a compact and interpretable statistical summary for the original corpus. However, due to the fast evolution of information technology in the past decade, applying topic modeling to real applications faces the following new challenges.

• **Large scale data:** Text data sets such as Web pages are growing overwhelmingly large. Topic modeling on such large scale data might be intractable due to memory and time issues.

• **Incremental data:** Corpora such as news articles grow rapidly over time. Traditional topic modeling needs to access the entire corpus, including the old and new data, for model update. However, it is really time-consuming to perform topic modeling from the scratch, and also achieving the old data consume lots of storage.

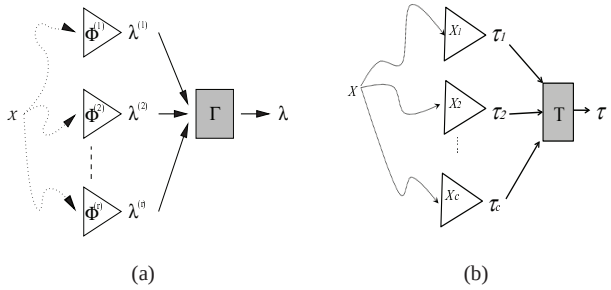

Figure 1. (a)Clustering ensembles [2]. (b)Topic modeling ensembles

• **Privacy concern:** When the data for processing are distributed separately over multiple organizations, we may face privacy concern. Namely, the participating organizations would not like to reveal their original data to public.

Motivated by these challenges we propose a novel ensemble framework, so-called *topic modeling ensembles*. As shown in Figure 1(b), it integrates multiple base topic models ($\tau_1 \cdots \tau_c$), learned from the disjoint sub-corpora ($X_1 \cdots X_c$), into a single ensemble topic model (τ). This framework helps to address the challenges above. More specifically, two of the application scenarios for topic modeling ensembles are

• **Distributed topic modeling:** Based on topic modeling ensembles, we could learn base topics models from each data node and then combine these base topic models for the entire data set. Different from previous distributed topic modeling techniques, we do not need any communication during the learning of base models.

• **Incremental topic modeling:** In the scenario of incremental topic modeling, data could be regard as distributed in different time slices. We can learn a base topic model for the new time slice and then combine it with other base topic models learned from the past time slices. Note that in incremental topic modeling we need only achieve the base topic models rather than the original data, which greatly saves the storage.

It is worth mentioning that only the base topic models

(the statistic summary of the local data), rather than the original data, are accessed by topic modeling ensembles. Thus, it naturally preserve the privacy of the participating organizations when the original data are not allowed to be disclosed. Note also that it is not required that the base topic models be learned from the disjoint sub-corpora of the whole data. This is only required in this study for the applications of distributed topic modeling and incremental topic modeling. What we need are only based topic models, no matter where they come from. Actually, some preliminary experiments show that if we perform topic modeling methods with different parameters and different initialization techniques over the whole corpus, the ensemble of these resultant base topic models may improve the effectiveness of each base model.

The main contributions of the paper include:

• We propose the framework of topic modeling ensembles and theoretically analyze the relationship between the ensemble approach and the direct approach.

• We implement the proposed framework for the two most popular topic models: PLSA and LDA.

• We conduct extensive experiments to evaluate the proposed ensemble approach using both synthetic and real-life benchmark data sets. The experimental results demonstrate the scalability of the proposed work with comparable performance on effectiveness.

The rest of paper is organized as follows. Section II gives the basic theoretical analysis, which illuminates topic modeling ensembles. Sections III and IV details its implementations for PLSA and LDA respectively. Its geometric interpretation is also discussed here. Section V present the experimental results on the time efficiency and effectiveness of topic modeling ensemble. Finally, Section VI discusses the related works, followed by the conclusion in Section VII.

II. Topic Modeling Ensembles

In this section we give the basic theoretical analysis, which illuminates topic modeling ensembles. A glossary of notations used in this paper is given in Table I. Before describing the ensemble approach, we briefly describe PLSA first. PLSA (Figure 2(a)) assumes that each document $d \in \{1, 2, 3, ..., D\}$ is generated by a mixture of topics $t \in \{1, 2, ..., T\}$, where a topic is represented as a multinomial distribution over words $w \in \{1, 2, ..., W\}$. Then the learning of PLSA is to find appropriate t's that decompose the joint probability distribution of d, w as follows:

$$p(w, d) = p(d)p(w|d) = p(d) \sum_t p(t|d)p(w|t) \quad (1)$$

In this work, we assume the original corpus is separated into sub-corpora. Suppose each document corresponds to only one sub-corpus, we can denote c_d as the ID of the sub-corpus that contains d. Then we can learn a topic model

Table I
Notations

Symbol	Description
c	sub-corpus ID
C	number of sub-corpora
w	word
W	size of vocabulary
$\mathbf{w}$	word sequence
$\mathcal{D}_c$	c-th sub-corpus
D_c	number of documents in $\mathcal{D}_c$, $D_c = \lvert\mathcal{D}_c\rvert$
$\mathcal{D}$	corpus, $\mathcal{D} = \bigcup_c \mathcal{D}_c$
D	number of documents in $\mathcal{D}$, $D = \lvert\mathcal{D}\rvert$
d	document, $d \in \mathcal{D}$
$\mathbf{d}$	document sequence
$\mathcal{Z}_c$	base topic set of $\mathcal{D}_c$
Z_c	number of topics set for $\mathcal{D}_c$, $Z_c = \lvert\mathcal{Z}_c\rvert$
$\mathcal{Z}$	base topic set, $\mathcal{Z} = \bigcup_c \mathcal{Z}_c$
Z	the total number of local topics
z	base topic, $z \in \mathcal{Z}$
$\mathbf{z}$	sequence of local topics
t	global topic
T	number of global topics
y	ensemble topic
Y	number of ensemble topics
$\mathbf{y}$	ensemble topic sequence

separately on each sub-corpus c.

$$p(w, d|c) = p(d|c)p(w|d, c) = p(d|c) \sum_{z \in \mathcal{Z}_{c_d}} p(z|d)p(w|z) \quad (2)$$

where $z \in \mathcal{Z}_c$ is the locally learned *base topic*.

Note that $\mathcal{Z}_c$ with different c are disjoint. We can integrate all the base topics into $\mathcal{Z} = \bigcup_c \mathcal{Z}_c$. Now, if we consider the variable $z \in \mathcal{Z}$ as *pseudo document* we can apply another topic modeling over the co-occurrence of z and w in the whole corpus,

$$p(w, z) = p(z)p(w|z) = p(z) \sum_y p(y|z)p(w|y) \quad (3)$$

To avoid ambiguity we call y in the above equation *ensemble topic*, and the topics directly learned from the entire original corpus *global topic*, denoted by t as shown in (1). Exploring the relation among global topics, base topics and ensemble topics, we have the following proposition.

Proposition 1. *Given base topic z which satisfies (2), and ensemble topic y which satisfies (3), We have*

$$p(w, d) = p(d) \sum_y p(y|d)p(w|y) \quad (4)$$

where

$$p(y|d) = \sum_z p(z|d)p(y|z) \quad (5)$$

Proof: . See Appendix A ■

Comparing (4) and (1), Proposition 1 actually says that if Equations (2) and (3) hold the ensemble topic y represented by $p(w|y)$ can be viewed as a solution to the topic modeling over the original whole corpus. This motivates us to propose

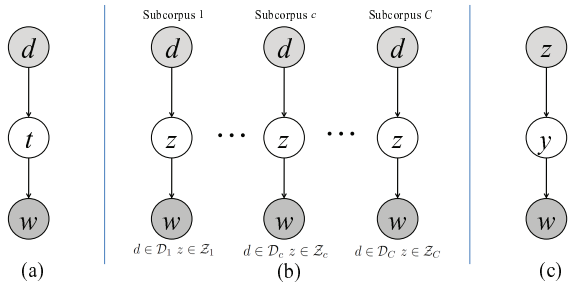

Figure 2. Graphical representation

the framework of distributed topic modeling ensembles as follows.

Phase 1. Base Topic Modeling: Learn base topics $p(w|z,c)$ from each sub-corpus c, as shown in Figure 2(b).

Phase 2. Ensemble Topic Modeling: Learn ensemble topics $p(w|y)$ over the co-occurrence of z and w in the whole corpus, as shown in Figure 2(c). Here, z is any topic in the union of all the base topics.

Phase 3. Inference: Take $p(w|y)$ as a resulting topic model and inference $p(y|d)$ for each document d. Phase 3 is optional because $p(y|d)$ can also be directly calculated via (5).

Note that Equations (2) and (3) only indicate the ensemble topic modeling are a solution for the entire corpus, but they give no demonstration on how "good" it is. In Section III-C we will show how good this solution is in terms of maximizing data likelihood.

III. Implementation For PLSA

In the previous section we take PLSA as an example to motivate the proposed framework. In this section we will detail the implementation of this framework for PLSA and give the theoretical analysis to validate our approach.

A PLSA model can be learned via a standard EM algorithm which maximizes the following log-likelihood.

$$\mathcal{L}_{t:d} = \sum_d \sum_w \left[\log\, p(d) \sum_t p(t|d)p(w|t) \right] \tag{6}$$

The update formulas used in the EM algorithm of parameter learning is shown as follows.

E-step:

$$p(t|w,d) = \frac{p(t)p(t|d)p(w|t)}{\sum_t p(t)p(t|d)p(w|t)} \tag{7}$$

M-step:

$$\begin{aligned} p(w|t) &= \frac{\sum_d n(d,w)p(t|w,d)}{\sum_{d,w'} n(d,w)p(t|w',d)} \\ p(t|d) &= \frac{\sum_w n(d,w)p(t|w,d)}{\sum_{d',w} n(d,w)p(t|w,d')} \\ p(t) &= \frac{\sum_{d,w} n(d,w)p(t|w,d)}{\sum_{d,w} n(d,w)} \end{aligned} \tag{8}$$

where $n(d,w)$ is the occurrence number of d and w.

A. PLSA Ensembles

In the proposed framework, the base topics are learned via maximizing the following log-likelihood for each sub-corpus c:

$$\mathcal{L}_{z,d_c} = \sum_{d \in \mathcal{D}_c} \sum_w \left[\log\, p(d) \sum_{z \in \mathcal{Z}_c} p(z|d)p(w|z) \right] \tag{9}$$

By replacing t with z in (7) and (8), we can learn $p(w|z)$ and $p(z|d)$ in a sub-corpus. Then, given $p(w,z)$ over the whole corpus, we can learn the ensemble topics y's by maximizing the following log-likelihood:

$$\begin{aligned} \mathcal{L}_{y:z} &= \sum_z \sum_w \log\, p(w,z) \\ &= \sum_z \sum_w \log \left\{ p(z) \sum_y \left[p(y|z)p(w|y) \right] \right\} \end{aligned} \tag{10}$$

Specifically, by replacing t, d with y, z respectively in (7) and (8), we can learn $p(w|y)$ and $p(y|z)$ by the EM algorithm. This time z is used as *pseudo document*. In this EM process the occurrence number $n(w,z)$ can be replaced by $p(w,z)$, where

$$p(w,z) = p(w|z)p(z) = p(w|z) \sum_d p(z|d)p(d).$$

and $p(d)$ is proportional to d's length, $p(w|z), p(z|d)$ are obtained from the base topics.

The complexity of EM algorithm for learning global topics is $\mathbf{O}(DWY)$ while if the learning of base topics is performed in parallel, the complexity of PLSA ensembles is

$$\max_c \left[\mathbf{O}(D_c W Z_c) \right] + \mathbf{O}(ZWY) \tag{11}$$

B. Geometric Interpretation

Topic modeling has an elegant geometric interpretation [4], [5], [6] as shown in Figure 3(a). In topic modeling, a document and a topic can be both represented as distributions over words, say $p(w|d)$ and $p(w|t)$, which can both be viewed as points on the $(W-1)$-simplex (*word simplex*). The T points corresponding to the topics can span another $(T-1)$-simplex (*topic simplex*). Then, the documents are

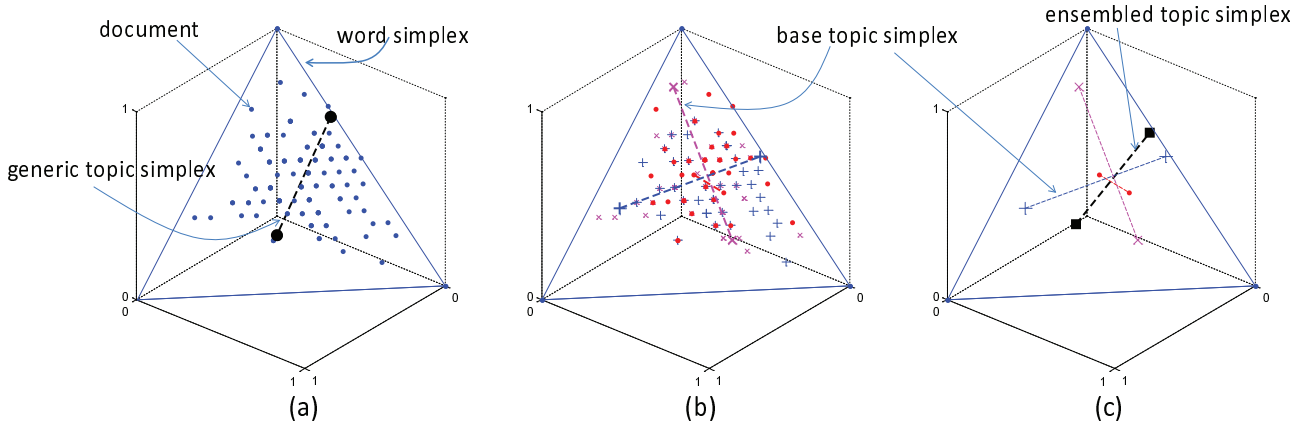

Figure 3. Geometric interpretation.

projected onto the topic simplex via maximizing (6). We set $W = 3$ and $T = 2$ in the illustrative example of Figure 3(a) where the word simplex is a triangle and the documents and topics are represented as points in this triangle. We use 120 points corresponding to 120 documents with $p(w|d)$ as the coordinates. The two black points corresponding to the topics span a 1-simplex (a line segment).

We can also perform topic modeling ensemble as follows. In Phase 1 we learn a base topic simplex on each of the three sub-corpora, which are represented with different markers in Figure 3(b). The resultant three topic simplices, represented by the three line segments with six vertices, are also shown in Figure 3(b). In this phase all the documents actually are projected onto the corresponding base topic simplices. In Phase 2, instead of considering their image points in the base topic simplices we use the vertices of the base topic simplices as the pseudo documents to learn the ensemble topics via maximizing (10). Here, we have six vertices for the three base topic simplices. As we can see, the learned ensemble topic simplex in Figure 3(c) is close to the global topic simplex in Figure 3(a). Next, we will theoretically show that using only the vertices of the base topic simplices in Phase 2 is reasonable.

C. Discussions on the Approximate in Topic Modeling Ensemble

In Phase 2, rather than consider all the documents' image points in the base topic simplices, we propose to maximize the log-likelihood in (10), which only use the vertices of the base topic simplices. The following proposition shows that this approximate is reasonable.

Proposition 2. *Let $\mathcal{L}_{y:d}$ denote the log-likelihood based on the ensemble topics, where*

$$\mathcal{L}_{y:d} = \sum_w \sum_d \log\, p(d) \sum_y p(y|d)p(w|y) \qquad (12)$$

Then, maximizing $\mathcal{L}_{y:z}$ in (10) is equivalent to maximize a lower bound of $\mathcal{L}_{y:d}$.

Proof: See Appendix B ■

It indicates that the EM algorithm for the ensemble process actually maximize a lower bound of the log-likelihood over the whole corpus. In this sense we argue that the output from the ensemble process is a good solution. The experiments in Section V further demonstrate that the effectiveness of this approximation is acceptable in practice.

IV. Implementation For LDA

In this section we will detail how to implement this ensemble framework for LDA.

When we regard $p(t|d)$ and $p(w|t)$ as random variables and assume they have Dirichlet prior with hyper-parameters α and β respectively, we get the LDA model. We then denote the posterior of $p(t|d)$ and $p(w|t)$ as $p(t|d;\alpha)$ and $p(w|t;\beta)$. In this paper we consider the LDA model with prefixed α and β and the posteriors $p(t|d;\alpha)$ and $p(w|t;\beta)$ could be estimated via Collapsed Gibbs Sampling (CGS) [6]. Specifically, the input data for CGS are two aligned vectors:

$$\begin{bmatrix}\mathbf{d}\\ \mathbf{w}\end{bmatrix} = \begin{bmatrix}d_1, & ..., & d_n\\ w_1, & ..., & w_n\end{bmatrix}$$

where n denotes total number of tokens, $d_i \in \{1, 2, ..., D\}$ and $w_i \in \{1, 2, ..., W\}$. The tuple (d_i, w_i) denotes an occurrence of word w_i in document d_i. The output of CGS is another vector $\mathbf{t} = t_1, t_2, ..., t_n$, where each $t_i \in \{1, 2, ..., T\}$ is a topic assignment for tuple (d_i, w_i). The states in $\mathbf{t}$ are randomly initialized. Then, the assignment of each t_i is iteratively updated by sampling from a distribution as follows:

$$p(t|\mathbf{t}_{\neg i}, w_i, d_i, \alpha, \beta) \propto \frac{O^{WT}_{w_i t} - 1 + \beta}{\sum_{w=1}^{W}(O^{WT}_{wt} + \beta) - 1} \times \frac{O^{DT}_{d_i t} - 1 + \alpha}{\sum_{t=1}^{T}(O^{DT}_{dt} + \beta) - 1} \qquad (13)$$

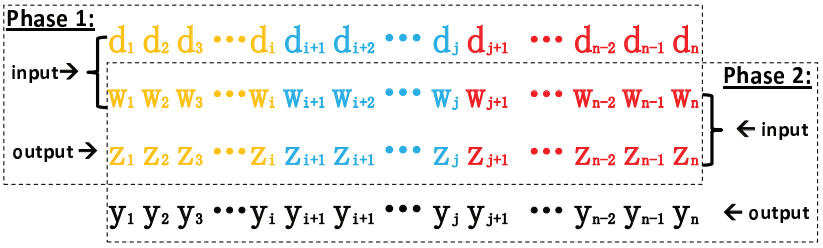

Figure 4. Illustration of Gibbs Sampling

where $O_{dt}^{DT} = \#(d,t)$, $O_{wt}^{WT} = \#(w,t)$ and $\neg i$ denotes the exclusion of the current one.

After a sufficient number of sampling iterations, the posterior of $p(t|d;\alpha)$ and $p(w|t;\beta)$ could be estimated based on O^{DT}, O^{WT}, α and β.

A. LDA Ensembles

We apply CGS in LDA ensembles as illustrated in Figure 4. In Phase 1 for base topic modeling, we split the original $[\mathbf{d},\mathbf{w}]^T$ into distributed segments (indicated with different colors), each of which corresponds to the data from a sub-corpus. Then, for each segment we learn the base topics $z \in \mathcal{Z}_c$ separately. In Phase 2 for ensemble topic modeling, we combine the segments of base topics from all the sub-corpora into a single vector $\mathbf{z}$ and take $[\mathbf{w},\mathbf{z}]^T$ as input for another CGS, where we can regard z as *pseudo document* again. Note that the base topics with different colors should be indexed distinctly. The output from Phase 2 is the vector of $\mathbf{y}$ based on which it is easy to get the ensemble topics $p(w|y)$.

B. Rescale on Co-occurrence Number

It is clear that the complexity of CGS is proportional to the number of tokens in the corpus, namely

$$\sum_{d,w} O_{dw}^{DW} = |\mathbf{d}| = |\mathbf{w}| = |\mathbf{z}| = |\mathbf{y}| \tag{14}$$

Thus, Phase 2 has the same complexity with that in topic modeling over the whole corpus. To achieve better efficiency in Phase 2, we can obviously employ PLSA or the variational EM proposed in [5] over O^{ZW}, whose complexity are both $\mathbf{O}(Z \times W)$. Here, we propose another strategy to accelerate the CGS process in Phase 2.

After Phase 1, we observe that due to $\sum_{z,w} O_{zw}^{ZW} = \sum_{d,w} O_{dw}^{DW}$ and $Z \ll D$, some counts in O^{ZW} are very large. We can rescale O^{ZW} via $\lceil \frac{O^{ZW}}{R} \rceil$ as the input for Phase 2 with less tokens, where R is a rescaling coefficient. Then, if the learning of base topics is conducted in parallel the total complexity of Phases 1 and 2 is

$$\max_c [\mathbf{O}(\sum_{z \in \mathcal{Z}_c, w} O_{zw}^{Z_c W})] + \mathbf{O}(\sum_{z,w} \lceil \frac{O_{zw}^{ZW}}{R} \rceil) \tag{15}$$

We set $R = 2C$ in our experiments and find in experiments that this setting significantly improves the efficiency of the ensemble topic modeling phase while achieves acceptable effectiveness on the large data corpus.

V. Experimental Results

In this section with various data sets we evaluate the ensemble framework for distributed topic modeling. For each data set we randomly divide it into several sub-corpora, lean the base topics over the sub-corpora separately, and then combine these base topics by ensemble. Incremental topic modeling can be viewed as a special case of distributed topic modeling, thus is not evaluated individually.

We set different topic numbers to be the same, i.e. $T = Y = Z_c$. In the EM procedures for PLSA, we terminate the iteration at round p, if the relative change of log-likelihood $\Delta\mathcal{L}/\mathcal{L}^{(p-1)} < 10^{-4}$. In the CGS procedure of LDA, we set $\alpha = 50/T$ and $\beta = 0.01$ if there is no extra declaration, and run 100 iterations for each algorithm.

A. Illustrative Examples on Synthetic Data

In Section III-B, we've illustrated the implementation for PLSA ensembles by simplex examples. Here we borrow the *bar* graphical example [6] for LDA ensembles. In this synthetic data set, documents and topics are represented by images, each containing 9 pixels in a 3×3 square. These 9 pixels can be viewed as words and the intensity of a pixel in a image encodes the frequency of the corresponding word in a document or the word's weight in a topic. We firstly give 6 topics (Figure 5(a)) corresponding to horizontal and vertical bars and then generate 600 documents (Figure 5(b)) following a standard LDA generative process based on these 6 topics with $\alpha = 1, \beta = 1$ and for each document we sample 100 words. We can learn the global topics (Figure 5(c)) via apply CGS directly to all the 600 pseudo documents. Comparing panel (c) to panel (a) of Figure 5, we see the learned topics approximately reveal the underline structure of these documents.

Now we apply the LDA ensembles to this synthetic data set. In the base topic modeling phase (Arrow 3 in Figure 5), we split the documents into 3 parts, and then learn 3

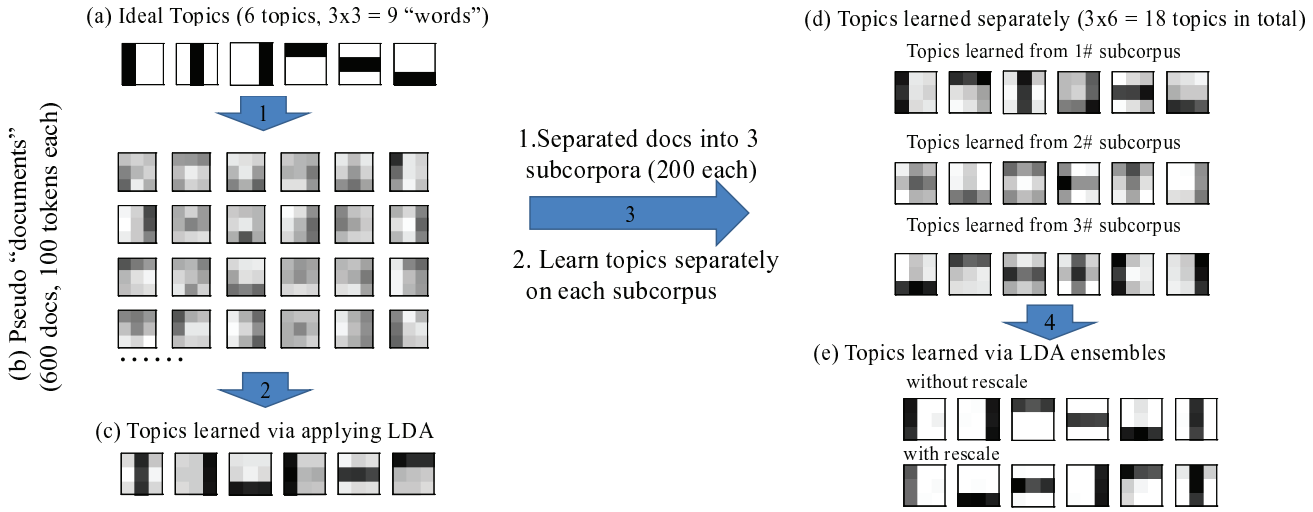

Figure 5. Illustration with bar example.

sets of base topics (Figure 5(d)). There are totally 3×6 = 18 based topics. In the ensemble topic modeling phase, we treat these base topics as another set of *pseudo documents* and then learn the ensemble topics as shown in Figure 5(e). We show both the results with or without the rescale strategy introduced in Section IV-B. Compare these results with directly learned topics in Figure 5(c) and the ground truth topics in Figure 5(a), we can see the ensemble topics are even better than the topics directly learned from the original data. This superior result is based on the fact that the number of instances in each sub-corpus is sufficient enough to get the good base topics. This superiority will not hold for all real-life applications and in the experiments on real-world document sets we only demonstrate that the proposed distributed ensemble framework can approximate the non-distributed topic modeling.

B. Experiments on Real-life Data

In this section, evaluate the proposed topic modeling ensembles over three real-life data set for document modeling and document classification task.

1) Data Sets: The real-life data sets are generated from three text sources[1], including *Industry Sector*(*Sector* for short), *20-Newsgroups*(*Newsgroup* for short) and *SRAA*.

Sector: The *Sector* data set is a collection of web pages classified into a class hierarchy and we use the 12 classes in the 2nd level.

Newsgroup: The *Newsgroup* data set is a text collection of about 20,000 UseNet postings from 20 newsgroups considered as 20 classes.

SRAA: The *SRAA* data set contain 73,218 UseNet articles from four discussion groups regarded as four classes. For each data set, we choose 2,000 words[2] with highest information gains according to the known categories.

Note that the ranking of these three data sets in the increase order of corpus size is: *Sector*, *Newsgroup* and *SRAA*.

2) Evaluation Metrics: Perplexity is a common measure for the document modeling effectiveness which evaluates the model generalization performance on a held-out document set. Formally, for a test corpus with M documents, the perplexity is defined as

$$\text{Perplexity}(D_{\text{test}}) = exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M}N_d}\right\} \quad (16)$$

Let $\boldsymbol{C}$ denote the result of classification and $\boldsymbol{B}$ denote the "true" class labels. The number of classes is K. Suppose n_{ij} is the number of documents which are labeled as i in $\boldsymbol{C}$ and j in $\boldsymbol{B}$,the classification accuracy is defined as

$$\text{Accuracy}(\boldsymbol{C},\boldsymbol{B}) = \frac{\sum_{i=1}^{K} n_{ii}}{M} \quad (17)$$

Note that smaller perplexity means better performance, and bigger accuracy means better performance.

3) Results in Document modeling: We evaluate the effectiveness and efficiency of document modeling over two dimensions: number of topics and number of sub-corpora, and train the following models: 1) *PLSA* for non-distributed PLSA; 2) *PLSA-E* for PLSA ensembles; 3) *LDA* for non-distributed LDA; 4) *LDA-E with rescale* for LDA ensembles with the rescale step and 5) *LDA-E without rescale* for that without the rescale step.

[1]http://www.cs.umass.edu/∼mccallum/code-data.html

[2]We use such small vocabularies that the non-distributed algorithms can work on large corpus such as *SRAA* in a tolerable time.

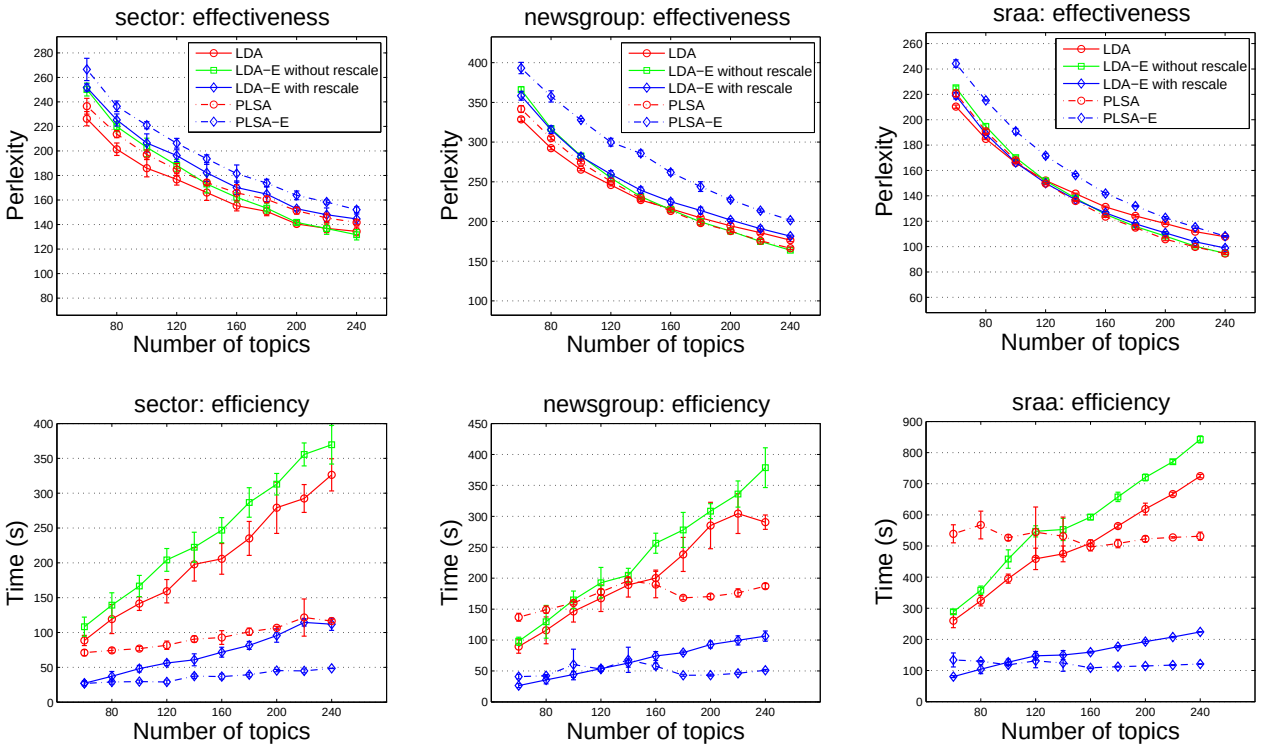

Figure 6. Results with varying number of topics.

Figure 6 illustrates the perplexity results (upper row) and the time costs (lower row) with respect to different numbers of topics (increasing T from 60 to 240 by 20). For each T, we do five-fold cross validation and plot the mean value together with a error bar for the standard deviation. It shows that the perplexity values strictly decrease along the increase of topic numbers T. We also find that along with the increase of the corpus size, the perplexity values of the ensemble methods are more and more close to (even better than for the large corpus of *SRAA*) their corresponding original topic modeling. Meanwhile, we plot the absolute time costs for the models. The efficiency and scalability of the ensemble methods are significantly better than applying topic modeling directly to the original corpus.

Figure 7 shows the impact of increasing the number C of sub-corpora. From these results we can see the perplexity values of the ensemble methods, except *LDA-E with rescale*, are stable along with the increase of sub-corpora number. It shows that the rescale process sacrifices more when the sub-corpora number increases. We also find that when the corpus is large, e.g. on *SRAA*, this sacrifice begins later (from a larger sub-corpora number). It indicates again that our ensemble methods prefer large data sets. We also measure the speed-up for the ensemble methods as plotted in the right column of Figure 7. The speed-up is not significant when the corpus size is relatively small, while for the corpus as large as *SRAA*, *PLSA-E* and *LDA-E with rescale* can almost achieve linear speed-up.

4) Results in Document Classification: Since all the corpora we used in experiments have class labels we can conduct binary classification problems on them. For a corpus with K classes, we can conduct $K(K-1)/2$ binary classification problems, and the values of $p(t|d)$ over different topics can be viewed as the features of the document d for classification. So we can compare the classification accuracy over the topic spaces from different topic modeling methods with that over the original bag-of-words space as the baseline. Logistic regression is adopted as the binary classifier. We rank all the classification problems from a corpus in the increase order of their accuracy from the bag-of-words baseline. All these results are included in Figure 8 where two values of sub-corpora number, $C = 5$ and $C = 10$, are tested. For each corpus we also give the mean accuracy values together with standard deviation in the legends of the figures. It shows that the ensemble methods of topic modeling are very close to those directly modeling the original corpora in terms of classification accuracy. We also find that the accuracy values do not significantly decrease

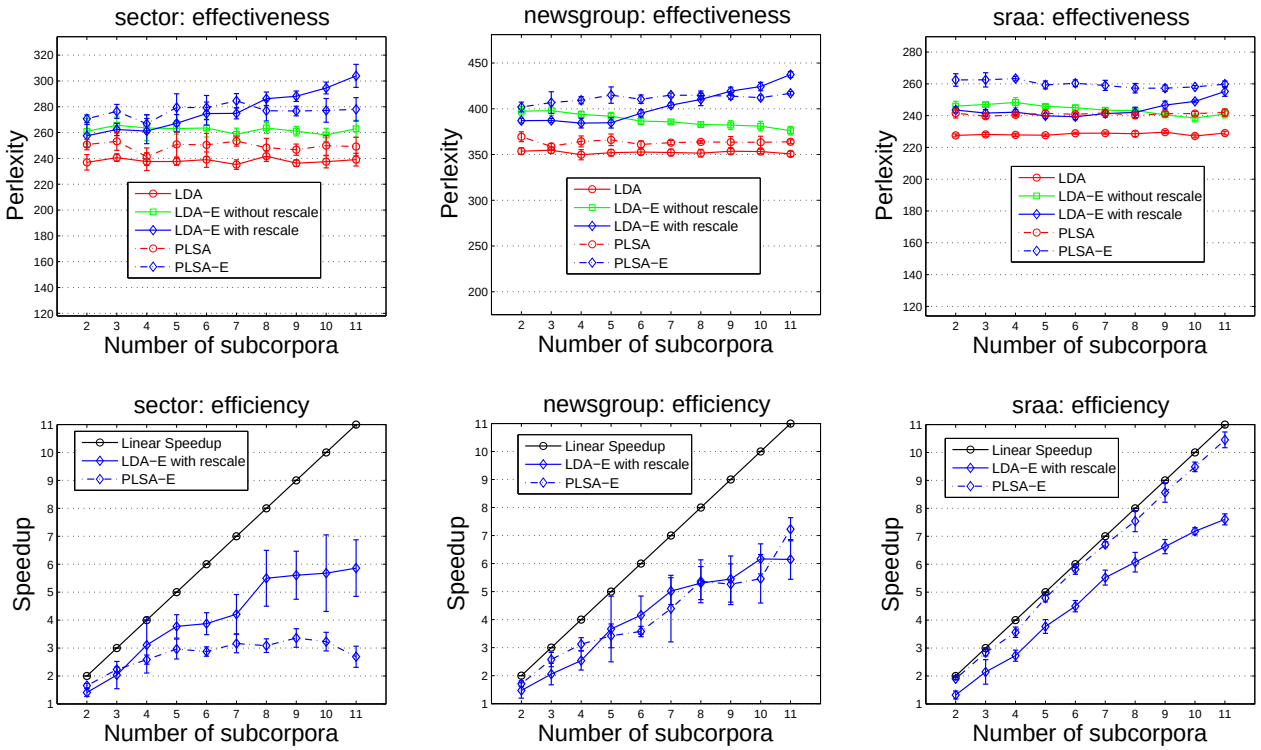

Figure 7. Results with varying number of subcorpora.

when we increase the sub-corpora number from 5 to 10.

C. Discussions

According to the complexity analysis in Section III and Section IV the speed-up will be significant when $D \gg \max_c(D_c), D \gg W$, $D_c \gg Z_c$. This is validated especially on the large corpus $SRAA$. In terms of effectiveness our ensemble methods also prefer large corpora with great co-occurrences, which leads to 1) data instances in each sub-corpus are sufficient for base topic learning; 2) base topics learned from different sub-corpora have enough overlaps which benefit the subsequent ensemble phase. In cases D_c's are sufficient but the sub-corpus is too large to fit in the memory of node c, we can reduce W by feature selection to reduce the sub-corpus size.

VI. Related Works

The related works to the proposed framework can be separated into two parts: clustering ensembles, distributed topic modeling and incremental topic modeling. As to the best of our knowledge, there is no close related work to the privacy preserving topic modeling.

Topic modeling could be viewed as soft co-clustering for both documents and words. The proposed topic modeling ensembles could also be viewed as an extension of clustering ensembles [2], [7], [8], [9], soft clustering ensembles [10], [11], [3] and co-clustering ensembles [12]. The extension is not only in conceptual but also methodological. A popular method for clustering ensembles is to take multiple clustering results as "pseudo features", and apply another phase of clustering over them to get a consensus clustering result. In topic modeling ensembles, multiple local topic models learned from the partitions of the original corpus are used as sets of "pseudo documents" for the second phase of topic modeling, which generate the global topic models.

Comparing with distributed topic modeling [13], [14], [15], the proposed framework has a complete distributed manner, i.e. it needs no communication overhead in the local computing phase. Moreover, unlike the traditional distributed topic modeling techniques, which rely on elaborately designed parallel computing algorithms, the proposed framework employs the original PLSA or LDA algorithms. The incremental topic modeling [16] and other related works, such as dynamic topic modeling [17], [18], [19] or online topic modeling [20] can also handle the text data with growing size. However, most of them pay more attention to tracking the topic evolution in the text streams while the proposed frame work aim to learn a global topic model as if the data is static.

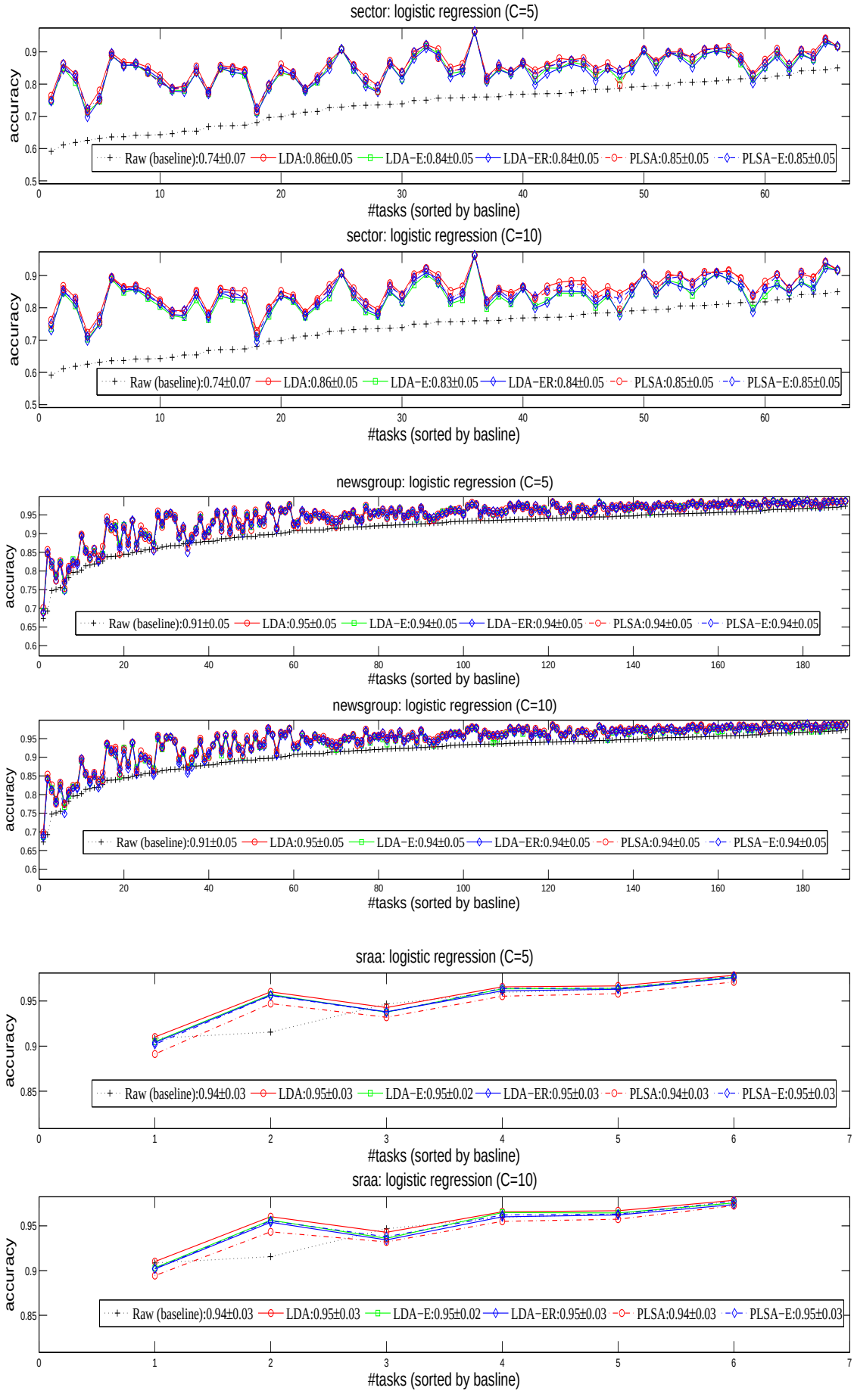

Figure 8. Document classification via logistic regression.

VII. Conclusions

In this paper we propose topic modeling ensembles, an novel solution to combine the base topic models from disjoint subsets of a corpus. The proposed framework has no communication overhead in the distributed computing phase and is easy to implement. We apply our approach to both PLSA and LDA with the discussion of the theoretical foundation. The experiments validate the effectiveness and efficiency of the proposed framework.

References

[1] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 26, 1996.

[2] A. Strehl and J. Ghosh, "Cluster ensembles – a knowledge reuse framework for combining multiple partitions," *JMLR*, 2002.

[3] K. Punera and J. Ghosh, "Consensus based ensembles of soft clusterings," *Appli. Artif. Intel.*, 2008.

[4] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of UAI*, 1999.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, 2003.

[6] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Latent Semantic Analysis: A Road to Meaning*, 2007.

[7] H. Wang, H. Shan, and A. Banerjee, "Bayesian cluster ensembles," in *Proc. of SDM*, 2009.

[8] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering:," in *Proc. of ICML*, 2003.

[9] ——, "Solving cluster ensemble problems by bipartite graph partitioning," in *Proc. of ICML*, 2004.

[10] C. Domeniconi and M. Al-Razgan, "Weighted cluster ensembles: methods and analysis," *ACM Trans. Knowl. Discov. Data*, 2009.

[11] K. Punera and J. Ghosh, "Soft clusterings ensembles," *Advances in fuzzy clustering and its applications*, 2007.

[12] P. Wang, C. Domeniconi, and K. B. Laskey, "Nonparametric bayesian co-clustering ensembles," in *Proc. of NIPS Workshops of NP Bayes*, 2009.

[13] R. Nallapati, W. W. Cohen, and J. D. Lafferty, "Parallelized variational em for latent dirichlet allocation: an experimental evaluation of speed and scalability," in *Proc. of ICDM Workshops*, 2007.

[14] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed inference for latent dirichlet allocation," in *Proc. of NIPS*, 2007.

[15] Y. Wang, H. Bai, M. Stanton, W.-Y. Chen, and E. Y. Chang, "Plda: Parallel latent dirichlet allocation for large-scale applications," in *Proc. of AAIM*, 2009.

[16] A. Surendran and S. Sra, "Incremental aspect models for mining document streams," in *Proc. of PKDD*, 2006.

[17] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *Proc. of SIGKDD*, 2006.

[18] C. Wang, D. M. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Proc. of UAI*, 2008.

[19] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. of ICML*, 2006.

[20] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking," in *Proc. of ICDM*, 2008.

Appendix A.
Proof of Proposition 1

Proof: For (5): we have

$$p(y|d) = \sum_z p(y, z|d) = \sum_z p(y|z, d)p(z|d) \tag{18}$$

Since y and d are independent when z is fixed which means

$$p(y|z, d) = p(y|z) \tag{19}$$

Thus, (5) holds.

For (4): substitute (3) into (2) we have

$$\begin{aligned} p(w, d) &= p(d) \sum_z p(z|d)[\sum_y p(w|y)p(y|z)] \\ &= p(d) \sum_y p(w|y) \sum_z p(z|d)p(y|z) \\ &= p(d) \sum_y p(w|y)p(y|d) \end{aligned}$$

■

Appendix B.
Proof of Proposition 2

Proof: First, we have the following inequality

$$\begin{aligned} \mathcal{L}_{y:d} &= \sum_w \sum_d \log\, p(d) \sum_y p(y|d)p(w|y) \\ &= \sum_w \sum_d \log \sum_y \sum_z p(d)p(z|d)p(y|z)p(w|y) \quad (20) \\ &\geq \sum_{w,d,z} p(z|d) \log \left\{ p(d) \sum_y [p(y|z)p(w|y)] \right\} \quad (21) \\ &= \sum_d p(z|d) \cdot \sum_z \sum_w \log \left\{ \sum_y [p(y|z)p(w|y)] \right\} \\ &+ \sum_z \sum_w \sum_d p(z|d) \log p(d) = \mathcal{L}'_{y:d} \end{aligned}$$

where (20) follows from Equation (5), and (21) follows from Jensen's inequality. Thus, $\mathcal{L}'_{y:d}$ is a lower-bound of $\mathcal{L}_{y:d}$. From (10) we have

$$\mathcal{L}_{y:z} = \frac{\mathcal{L}'_{y:d} - \sum_{z}\sum_{w}\sum_{d} p(z|d)\log p(d)}{\sum_{d} p(z|d)} + \sum_{z}\sum_{w} \log\, p(z) \tag{22}$$

Since z and d are observed in Phase 2, $\sum_{d} p(z|d)$, $\sum_{z}\sum_{w}\sum_{d} p(z|d)\log p(d)$ and $\sum_{z}\sum_{w} \log\, p(z)$ are all constant. Therefore, maximizing $\mathcal{L}_{y:z}$ in (10) is equivalent to maximizing $\mathcal{L}'_{y:d}$, a lower-bound of $\mathcal{L}_{y:d}$. ■