7-2015

Topic Modeling with Document Relative Similarities

Jianguang DU

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Dandan SONG

Lejian LIAO

Topic Modeling with Document Relative Similarities

Jianguang Du[†], Jing Jiang[‡], Dandan Song[†], Lejian Liao[†]
[†]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[‡]School of Information Systems, Singapore Management University, Singapore
dujianguang@bit.edu.cn; jingjiang@smu.edu.sg; {sdd,liaolj}@bit.edu.cn

Abstract

Topic modeling has been widely used in text mining. Previous topic models such as Latent Dirichlet Allocation (LDA) are successful in learning hidden topics but they do not take into account metadata of documents. To tackle this problem, many augmented topic models have been proposed to jointly model text and metadata. But most existing models handle only categorical and numerical types of metadata. We identify another type of metadata that can be more natural to obtain in some scenarios. These are relative similarities among documents. In this paper, we propose a general model that links LDA with constraints derived from document relative similarities. Specifically, in our model, the constraints act as a regularizer of the log likelihood of LDA. We fit the proposed model using Gibbs-EM. Experiments with two real world datasets show that our model is able to learn meaningful topics. The results also show that our model outperforms the baselines in terms of topic coherence and a document classification task.

1 Introduction

Topic modeling has been a popular text analysis method [Blei *et al.*, 2003]. In topic models, it is assumed that a document is generated by a mixture of topics, each of which is a distribution over words in the vocabulary. By fitting the models, we can represent each document through the learned topics as well as understand the topics in the corpus through the most probable words of each topic. Classical topic models such as Probabilistic Latent Semantic Analysis (PLSA) [Hofmann, 1999] and Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] have shown impressive success in modeling text documents.

Despite the success of the above two topic models, they cannot directly incorporate metadata such as the author or the posting time of a document to improve the quality of the learned topics. To tackle this problem, many augmented topic models have been developed (e.g. [Rosen-Zvi *et al.*, 2004; Wang and McCallum, 2006; Mcauliffe and Blei, 2008]). Depending on the nature of the metadata, different models make different assumptions about the relations between the metadata and the underlying topic mixture of each document and/or the word distributions of topics.

Generally, most of these models rely on the assumption that documents with the same attribute values (e.g. same authors or same publishing time) tend to have similar topic distributions. This assumption can be captured by establishing a dependency between the topic distributions of documents and the metadata, where the metadata can be either categorical (e.g. [Rosen-Zvi *et al.*, 2004]) or numerical (e.g. [Wang and McCallum, 2006]). However, metadata may not always be in the form of categorical or numerical attribute values. For example, given documents `A`, `B` and `C`, one may point out that `A` is more similar to `B` than to `C`. Presumably, this kind of *relative similarity* can also help topic modeling; document `A`'s topic distribution should be closer to `B`'s than to `C`'s.

In many scenarios, relative similarities may be easier to obtain than categorical or numerical attributes. For example, to an online user with a particular information need, it is easier to ask her to compare two Web pages and point out which one is more relevant than to ask her to score each Web page. Given a set of research papers, it might be easier to identify similar papers than to assign categorical labels to the papers. Even for documents with categorical or numerical attributes, sometimes it may also be meaningful to consider relative similarities derived from the attribute values. For example, a document written by a 10-year-old is more similar to a document written by a 12-year-old than by a 20-year-old. The value of the absolute age difference is not so important; we generally would not consider the difference between a 10-year-old and a 20-year-old to be 5 times that of the difference between a 10-year-old and a 12-year-old. Instead, the relative difference or similarity is more useful.

In this paper, we study topic modeling on documents with this kind of relative similarities. Our goal is to explore a general model that can utilize such constraints without manipulating the graphical structure of the classical topic models. The intuition is to facilitate the collaboration between topic models and the constraints. Specifically, we first transform the constraints into a loss function, and then design an objective function that combines the log likelihood of the corpus with the loss function. Essentially, the loss function acts as a *regularizer* for the log likelihood of text, ensuring that the parameters that minimize the overall objective function balance between fitting the text and satisfying the relative

similarity-based constraints. To fit our model, we use the Gibbs-EM algorithm, an algorithm that alternates between collapsed Gibbs sampling and gradient descent.

We evaluate our model with two datasets, 20 newsgroups and TDT2. Experimental results show that our model outperforms LDA in terms of topic coherence. When we use the learned topics for document classification, our model also achieves higher accuracy than other baselines including supervised LDA [Mcauliffe and Blei, 2008].

In summary, the main contributions of this paper are the following:

- We identify an unexplored type of metadata, which we call relative similarities, for topic modeling. We point out that relative similarities can also be derived from traditional categorical and numerical metadata.
- We propose a general model that combines LDA with relative similarity-based constraints (Section 3).
- We provide an efficient inference algorithm to fit the proposed model (Section 4).
- With two standard datasets, we show that our model is able to learn more meaningful topics than LDA (Section 5).

2 Related Work

Our work is mostly related to topic modeling with additional document metadata. We review two general approaches to such topic modeling.

2.1 Augmented Topic Models

The first approach to modeling document metadata is through modifying the structure of the graphical model. There are mainly two ways to incorporate metadata into the graphical model. One way is to generate the metadata from the topic distributions of each document. In such models, typically a topic is associated with not only a word distribution but also a distribution over different values of an attribute [Wang and McCallum, 2006; Mcauliffe and Blei, 2008; Qiu *et al.*, 2013; Liao *et al.*, 2014]. For example, in [Mcauliffe and Blei, 2008] the authors proposed a general supervised LDA model for scenarios where each document has a response variable. The response variable is assumed to be generated from a Gaussian distribution, the mean of which is determined by the topic assignment of the words in the document through a generalized linear model. This supervised LDA model can be used for predicting response variables such as star ratings of reviews and scores of student essays.

Another approach assumes that the topic distribution of each document is a mixture of metadata-specific topic distributions [Rosen-Zvi *et al.*, 2004; McCallum *et al.*, 2005; Eisenstein *et al.*, 2010]. A typical example of this approach is the author-topic model [Rosen-Zvi *et al.*, 2004]. Here to sample the topic of a word, first one of the authors of the document is chosen, and then a topic is sampled from that author's topic distribution. This approach generally only works for metadata with categorical values.

All the models above assume either categorical or numerical type of metadata. For our work, we focus on metadata in the form of relative similarities. It is not easy to modify standard graphical models to incorporate this kind of metadata.

2.2 Topic Modeling with Regularization

Another line of literature related to our work is topic modeling with regularization terms [Cai *et al.*, 2008; Huh and Fienberg, 2010; Tang *et al.*, 2013; McAuley and Leskovec, 2013; Andrzejewski *et al.*, 2011; Mei *et al.*, 2014]. Here the idea is to turn additional information about the documents into constraints or a loss function that do not necessarily follow a generative model. For example, in [Cai *et al.*, 2008] and [Huh and Fienberg, 2010], regularization terms were used to capture the manifold structures of documents. In [Tang *et al.*, 2013], the authors added a regularization term that pushes context-specific topic distributions close to the consensus topic distributions. In [McAuley and Leskovec, 2013], rating data were incorporated into the model and rating prediction errors became the additional loss function to regularize standard LDA. Recently, some researchers have also proposed to incorporate domain knowledge in the form of First-Order Logic (FOL) to standard LDA using a regularization framework [Andrzejewski *et al.*, 2011; Mei *et al.*, 2014]. While FOL rules are more expressive in representing knowledge, they require the analysts writing the rules to have background in FOL. In comparison, relative similarities as we use in this paper are generally easier to obtain.

In our work, we borrow the general idea of using a regularization term to incorporate additional knowledge. However, the type of metadata we model is very different from the work discussed above.

3 Model

We address the problem of topic modeling on documents where relative similarities are given. In this section, we first formulate the constraints derived from relative similarities. We then briefly review Latent Dirichlet Allocation (LDA) and propose our model, which combines LDA with relative similarities.

3.1 Constraints

We want to model documents where relative similarities are known for some documents if not all. Such relative similarities can come from human judges under a particular application setting or automatically derived from other data. First of all, we assume that we are given a set of documents denoted as $\mathcal{D}$. We further assume that there is a distance function $dist(d_i, d_j)$ defined for any pair of $d_i, d_j \in \mathcal{D}$. To formally formulate the relative similarities, we assume that we are given a set of T triplets as follows:

$$\mathcal{S} = \{(d_i, d_i^+, d_i^-)\}_{i=1}^T, \tag{1}$$

where $d_i, d_i^+, d_i^- \in \mathcal{D}$, and d_i is more similar to d_i^+ than to d_i^-. In other words,

$$dist(d_i, d_i^-) > dist(d_i, d_i^+), \quad \forall (d_i, d_i^+, d_i^-) \in \mathcal{S}. \tag{2}$$

Borrowing the idea from max-margin methods, we further rewrite the constraints as follows:

$$dist(d_i, d_i^-) \geq dist(d_i, d_i^+) + C, \quad \forall (d_i, d_i^+, d_i^-) \in \mathcal{S}. \tag{3}$$

where C, a positive constant, is a margin to ensure that $dist(d_i, d_i^-)$ is sufficiently larger than $dist(d_i, d_i^+)$.

Given the definitions above, our goal is to minimize a loss function defined below:

$$\mathcal{L} = \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-), \tag{4}$$

where

$$\mathcal{L}_i(d_i, d_i^+, d_i^-) = \max(0, dist(d_i, d_i^+) + C - dist(d_i, d_i^-)). \tag{5}$$

The next question is how to define the distance function *dist*. As we will see, the distance function will be based on the topic distributions learned by LDA.

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] is a well-developed and widely used topic model. Evolved from Latent Semantic Analysis (PLSA) [Hofmann, 1999], LDA defines a proper Bayesian model which overcomes the overfitting problem in PLSA.

Given a document set $\mathcal{D}$ where each document $d \in \mathcal{D}$ contains N_d words $\{w_{d,1}, w_{d,2}, \ldots, w_{d,N_d}\}$, we assume that there exist K topics, each associated with a multinomial word distribution $\boldsymbol{\varphi}_k$. Each document has a topic distribution $\boldsymbol{\theta}_d$ in the K-dimensional topic space. Each word in a document has a hidden topic label drawn from the document's topic distribution. Formally, the generative process of LDA is described as follows:

- For each topic $k = 1, \ldots, K$, draw $\boldsymbol{\varphi}_k \sim \text{Dir}(\boldsymbol{\beta})$
- For each document $d \in \mathcal{D}$
  - Draw $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})$
  - For each word $w_{d,n}, n = 1, \ldots, N_d$
    - Draw $z_{d,n} \sim \text{Multi}(\boldsymbol{\theta}_d)$
    - Draw $w_{d,n} \sim \text{Multi}(\boldsymbol{\varphi}_{z_{d,n}})$

Here $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are parameters of the Dirichlet priors.

The parameters of LDA can be learned in several ways including variational methods and Gibbs sampling.

3.3 The Regularized Model

In this subsection, we present our proposed model which combines LDA with the constraints defined in Section 3.1. Our motivation is to learn better topics by incorporating these constraints.

First of all, without considering the constraints and with only standard LDA, our objective is to find $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ that maximize the following objective function:

$$\log\Big(p(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{\varphi})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\varphi}|\boldsymbol{\beta})\Big).$$

To link this objective function with the constraints presented in Section 3.1, we can simply add the two terms together:

$$\log\Big(p(\boldsymbol{w}|\boldsymbol{\theta}, \boldsymbol{\varphi})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\varphi}|\boldsymbol{\beta})\Big) - \eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-), \tag{6}$$

where η is a constant to balance the two terms.

Now recall that we need to define the distance function *dist*, and this should be related to our model parameters, i.e. $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$. Since $\boldsymbol{\varphi}$ is not document-specific, we use just $\boldsymbol{\theta}$ to define *dist*. Intuitively, if two documents have similar $\boldsymbol{\theta}_d$, then their distance should be smaller. There are several choices we can consider. One is the Euclidean distance, where each $\boldsymbol{\theta}_d$ is treated as a k-dimensional vector and the standard Euclidean distance can be computed between them. In this paper, we experiment with squared Euclidean distance. Another choice is KL-divergence, defined as follows:

$$D_{\text{KL}}(\boldsymbol{\theta}||\boldsymbol{\theta}') \quad = \quad \sum_{k=1}^{K} \boldsymbol{\theta}_k \log \frac{\boldsymbol{\theta}_k}{\boldsymbol{\theta}'_k}.$$

However, because KL-divergence is not symmetric, i.e. generally $D_{\text{KL}}(\boldsymbol{\theta}||\boldsymbol{\theta}') \neq D_{\text{KL}}(\boldsymbol{\theta}'||\boldsymbol{\theta})$, here we consider the symmetric KL-divergence instead:

$$dist(d_i, d_j) \quad = \quad D_{\text{KL}}(\boldsymbol{\theta}_{d_i}||\boldsymbol{\theta}_{d_j}) + D_{\text{KL}}(\boldsymbol{\theta}_{d_j}||\boldsymbol{\theta}_{d_i}).$$

However, Eqn (6) is a constrained optimization problem because both $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are probability distributions. To transform the objective function into an unconstrained optimization problem, we first define the following transformation function for $\theta_{d,k}$:

$$\theta_{d,k} = \frac{e^{\lambda_{d,k}}}{\sum_{k'=1}^{K} e^{\lambda_{d,k'}}}. \tag{7}$$

We then change the Dirichlet prior on $\boldsymbol{\theta}_d$ into a Gaussian prior on $\boldsymbol{\lambda}_d$, that is, each $\lambda_{d,k}$ follows a Gaussian distribution with a zero mean and a variance of σ^2. Next we leave out $\boldsymbol{\varphi}$ from our objective function and try to estimate it later based on the hidden variables $\boldsymbol{z}$ (explained in the next section). Our modified objective function becomes the following:

$$\mathfrak{L}(\boldsymbol{\lambda}) = \underbrace{\log p(\boldsymbol{w}|\boldsymbol{\lambda}, \boldsymbol{\beta})}_{\text{log likelihood}} + \underbrace{\log p(\boldsymbol{\lambda}|(\mathbf{0}, \sigma^2\mathbf{I}))}_{\text{prior}} - \underbrace{\eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-)}_{\text{hinge loss}} \tag{8}$$

Note that the loss $\mathcal{L}_i(d_i, d_i^+, d_i^-)$ is also a function of $\boldsymbol{\lambda}$.

4 Model Fitting with Gibbs-EM

To optimize the objective function given in Eqn (8), we adopt the Gibbs-EM algorithm [Wallach, 2006]. Recall that we have defined hidden variables $\boldsymbol{z}$ that represent topic assignment. With the hidden variables, the objective function can be optimized in the following way. During E-step, we fix the parameter $\boldsymbol{\lambda}^{(t)}$ learned in the t^{th} iteration and obtain the conditional distribution of the hidden variables $p(\boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\beta})$. During the M-step, we solve the following optimization problem:

$$\boldsymbol{\lambda}^{(t+1)} = \arg\max \mathbb{E}_{\boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\beta}}[\mathfrak{L}'(\boldsymbol{\lambda})],$$

where $\mathbb{E}_q[f]$ is the expected value of f with respect to the distribution q, and

$$\begin{aligned} \mathfrak{L}'(\boldsymbol{\lambda}) \quad &= \quad \log p(\boldsymbol{w}, \boldsymbol{z}|\boldsymbol{\lambda}, \boldsymbol{\beta}) + \log p(\boldsymbol{\lambda}|(\mathbf{0}, \sigma^2\mathbf{I})) \\ &\quad -\eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-). \end{aligned} \tag{9}$$

With Gibbs-EM, instead of evaluating the exact conditional distribution $p(\boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\lambda}^{(t)}, \boldsymbol{\beta})$, we use Gibbs sampling to approximate it.

4.1 E-step

In the E-step, we use Gibbs sampling to sample the hidden topic variables $\boldsymbol{z}$ for all words $\boldsymbol{w}$ by fixing $\boldsymbol{\lambda}^{(t)}$. To simplify the discussion, we will use $\boldsymbol{\theta}$ when referring to topic distributions and $\boldsymbol{\lambda}$ otherwise. The deterministic relation between them is given in Eqn (7). To perform Gibbs sampling, we need to compute the probability of assigning a topic $z_{d,n}$ to a specific word $w_{d,n}$ given all the other topic assignment to all the other words:

$$p(z_{d,n} = k|\mathbf{w}, \boldsymbol{z}_{-(d,n)}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{w}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\beta})}{p(\boldsymbol{w}_{-(d,n)}, \boldsymbol{z}_{-(d,n)}|\boldsymbol{\theta}, \boldsymbol{\beta})},$$

where $-(d,n)$ indicates that $z_{d,n}$ or $w_{d,n}$ is excluded. By using the conjugacy property of Dirichlet and multinomial distributions, the Gibbs updating rule of our model can be represented as follows:

$$p(z_{d,n} = k|\mathbf{w}, \boldsymbol{z}_{-(d,n)}, \boldsymbol{\theta}, \boldsymbol{\beta}) \propto \frac{n_{k,w_{d,n}} + \beta - 1}{\sum_{v=1}^{V} n_{k,v} + V\beta - 1} \cdot \theta_{d,k}, \tag{10}$$

where $n_{k,v}$ denotes the number of times word v is assigned to topic k.

As we pointed out in the previous section, we do not directly estimate $\boldsymbol{\varphi}$ by optimizing the objective function. But with Gibbs sampling, $\varphi_{k,v}$ can be estimated as follows:

$$\hat{\varphi}_{k,v} = \frac{n_{k,v} + \beta}{\sum_{v'=1}^{V} n_{k,v'} + V\beta}. \tag{11}$$

Algorithm 1 Gibbs-EM for our model.

Input:
D documents, # topics K, size of the vocabulary V, regularization parameter η, margin C, max # EM iterations nEM, # Gibbs sampling iterations nGS, max # gradient descent in each M-step nGD.

Output:
$\lambda_{d,k}$ and $\varphi_{k,v}$, $d = 1, .., D; k = 1, .., K; v = 1, .., V$

```
Randomly initialize z and λ
t ← 0
while (t < nEM) do
    E-step:
    Sample z_{d,n} as in Eqn (10) with nGS iterations
    M-step:
    n ← 0
    while (n < nGD) do
        Compute the objective function ℒ'(λ) as in Eqn (9)
        Set the learning rate ξ
        for (d = 1 to D) do
            for (k = 1 to K) do
                Compute the partial derivative ∂ℒ'(λ)/∂λ_{d,k}
                λ_{d,k}^{(t)}_{(n+1)} ← λ_{d,k}^{(t)}_{(n)} + ξ ∂ℒ'(λ)/∂λ_{d,k}
            end for
        end for
        n ← n + 1
    end while
    t ← t + 1
end while
Compute each φ_{k,v} as in Eqn (11)
```

4.2 M-step

In the M-step, we use the last sample of $\boldsymbol{z}^{(t)}$ obtained from the previous E-step and use gradient descent to learn $\boldsymbol{\lambda}^{(t+1)}$:

$$\begin{aligned}\boldsymbol{\lambda}^{(t+1)} &= \arg\max_{\boldsymbol{\lambda}} \Big(\log p(\boldsymbol{w}, \boldsymbol{z}^{(t+1)}|\boldsymbol{\lambda}, \boldsymbol{\beta}) \\ &\quad + \log p(\boldsymbol{\lambda}|(\mathbf{0}, \sigma^2\mathbf{I})) - \eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-) \Big).\end{aligned} \tag{12}$$

By computing the first-order patrial derivatives of the objective function in Eqn (12) with respect to each $\lambda_{d,k}$, we can use gradient descent to optimize the objective function. The model fitting algorithm is summarized in Algorithm 1.

5 Experiments

In this section, we present the experiments to evaluate our model. We first describe our datasets. We then conduct experiments both quantitatively and qualitatively.

5.1 Datasets

We use two widely used text corpora, 20 newspapers[1] and TDT2 [Cai *et al.*, 2008].

The 20 newsgroups text corpus is a collection of approximately 20,000 newsgroup documents, partitioned evenly across 20 different newsgroups. We used a preprocessed version of this dataset[2], where the documents are divided into a training set and a test set. Among the documents in the training set, we randomly selected 100 documents from each category and removed stop words and very short documents (documents with fewer than 3 words), thus leaving us with 1,997 documents with 14,538 distinct words.

The TDT2 corpus consists of 11,201 documents which are classified into 96 categories. Only the largest 20 categories were kept, and those documents appearing in more than one category were removed. We also randomly sampled 100 documents from each category as the strategy for the 20 newsgroup dataset. Finally, there are 1,998 documents with 12,166 distinct words left.

5.2 Experimental Setup

In our experiments, we performed 200 runs of Gibbs-EM. In each run, we ran 100 iterations of Gibbs sampling and another 10 iterations of gradient descent. We set the Dirichlet prior $\boldsymbol{\beta} = \mathbf{0.1}$, the variance of Gaussian prior $\sigma = 1$. We also fixed the number of topics to be 20 (same as the number of categories in each dataset). Note that we do not tune this parameter since our goal is not to find the optimal number of topics.

To automatically obtain the triplet constraints for our model, we sampled a set of triplets from the documents in the training set according to their ground truth category labels. Specifically, we generated a triplet instance by randomly sampling two documents in the same category and one document from another different category. In our experiments, we generated 100K, 50K and 10K triplet instancnes (about 1%, 0.5%

[1] http://qwone.com/~jason/20Newsgroups/

[2] http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html

and 0.1% of all the triplet instances, respectively) from the training set of each dataset. Experiments were run on a machine with 4 cores and 4GB of memory.

5.3 Quantitative Evaluation

In this subsection, we give quantitative evaluation of our model on topic coherence and document classification. We evaluate our model with different number of sampled triplets (i.e. 100K, 50K and 10K) against other baselines. Specifically, the methods used in our experiments are:

- DRS-KL. Our model with the symmetric KL-divergence distance metric.
- DRS-SE. Our model with squared Euclidean distance as the distance function.
- LDA. The standard LDA model.
- sLDA. The supervised LDA model [Mcauliffe and Blei, 2008]. This is a strong baseline where metadata is also used. Here we treat the category labels of all the training documents as the response values. This baseline is used for comparison for the classification task as well as computational costs presented later.

Note that for DRS-KL and DRS-SE, we tune the regularization parameter η and margin C, and report the results with the best performance. To test the robustness of our model, we used 5-fold cross validation for all methods. We should emphasize that we sampled triplet instances only from the training documents, i.e. no metadata from the test documents is used. This ensures the fairness of the comparison of our model and the baselines.

Topic Coherence

In this experiment, we want to compare the performance of different topic models. Previous studies usually utilized perplexity (likelihood on held-out data) as the metric. However, such metric cannot measure the coherence of learned topics. Chang et al. [Chang *et al.*, 2009] found that perplexity is not always a good indicator of topic coherence [Mimno *et al.*, 2011]. To tackle this problem, we explore another metric to measure the quality of learned topics. Specifically, to measure the semantic coherence of topics, we use the point-wise mutual information (PMI) [Newman *et al.*, 2010]. PMI measures the co-occurrence of a number of words, which is defined as:

$$\text{PMI}(\boldsymbol{w}) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \quad (13)$$

where $\boldsymbol{w}$ are the top-N words of a topic, $p(w_i, w_j)$ is the probability that words w_i and w_j co-occur in the same document, while $p(w_i)$ is the probability that word w_i appears in a document. In order to compute PMI, we need a larger dataset. In our setting, we used 9,394 documents of TDT2 and 19,997 documents of 20 newsgroups to compute the PMI scores. N is set to 10 in our analysis.

The PMI scores of all candidate methods are presented in Table 1. From the results we can see that our model achieves higher PMI scores compared with the LDA model in all settings. So we can conclude that our model is more successful in learning coherence topics in both corpora. In comparison, LDA does not show such capabilities because it does not capture the constraints of relative similarities among documents.

Document Classification

With topic modeling, each document can be represented by its topic distribution. In this subsection, we investigate the use of hidden topics learned by different topic models for document classification.

We use the topic distributions of documents as features and train a RBF kernel Support Vector Machine (SVM) [Chang and Lin, 2011]. Table 2 shows the classification accuracy with 5-fold cross validation on both datasets.

The results reveal a number of points:

1. Compared with LDA, sLDA performs better. This is because sLDA makes use of the categorical information of the training documents, and thus learns topics that are more discriminative for the categories. Our method also consistently outperforms LDA in all settings. This shows that our way of using the metadata in the training documents is also effective.
2. The comparison between our method and sLDA is mixed, but generally when we use more triplets (100K) and when we use the symmetric KL-divergence distance, our method is more likely to outperform sLDA. Note that our method relies on relative document similarities and sLDA needs document labels. Moreover, in sLDA, the labels of all training documents are used, while in our model, only a small set of sampled labels are used. Hypothetically, in scenarios where only relative document similarities are available, we would not be able to apply sLDA but our method can still be used.

Computational Costs

Table 3: Comparison of training time (minutes).

# tps	method	time	
		20 newsgroups	TDT2
100K	DRS-KL	105	47
	DRS-SE	156	76
50K	DRS-KL	77	39
	DRS-SE	89	46
10K	DRS-KL	40	29
	DRS-SE	36	29
–	sLDA	38	58

In this subsection, we empirically compare the computational costs of our method and sLDA. The version of sLDA we used was implemented in C++[3] and our method was implemented in Java. Table 3 shows the results. We can see that when the number of triplets used is relatively small, our method takes similar amount of time as sLDA, and when the number of triplets increases to 100K, our method may take up to 4 times the computational cost of sLDA. Note that our method is not meant to outperform sLDA in terms of computational efficiency. The advantage of our method is its special

[3] http://www.cs.cmu.edu/~chongw/slda/

Table 1: Topic coherence based on PMI of topics on both datasets. The larger the metric is, the better the topics.

# tps*	method	20 newsgroups						TDT2					
		fold					avg*	fold					avg*
		0	1	2	3	4		0	1	2	3	4	
100K	DRS-KL	-1.43	-3.97	-4.75	-4.52	-4.16	-3.77	-0.96	0.36	-1.18	-1.16	-1.17	-0.82
	DRS-SE	-2.41	-2.79	-2.59	-1.28	-2.22	-2.26	0.40	-0.39	-4.09	-0.98	-0.20	-1.05
50K	DRS-KL	-3.41	-1.07	-3.17	-1.63	-4.55	-2.77	-1.76	-1.57	0.21	-1.38	-3.31	-1.56
	DRS-SE	-3.58	-0.44	-1.08	-2.45	-0.84	-1.68	-2.72	0.41	-0.98	-3.12	-3.32	-1.95
10K	DRS-KL	-3.18	-2.59	-4.16	-0.28	-3.41	-2.72	-1.17	0.01	-0.75	-0.19	-2.36	-0.89
	DRS-SE	-3.94	-6.27	-1.44	-5.50	-3.21	-4.07	-3.50	-4.49	-0.79	-0.40	-0.22	-1.88
–	LDA	–	–	–	–	–	-5.71	–	–	–	–	–	-1.98

* tps and avg represent triplets and average, respectively.

Table 2: Classification accuracy on both datasets.

# tps	method	20 newsgroups						TDT2					
		fold					avg	fold					avg
		0	1	2	3	4		0	1	2	3	4	
100K	DRS-KL	0.529	0.558	0.642	0.626	0.674	**0.606**	0.824	0.871	0.929	0.926	0.939	**0.898**
	DRS-SE	0.532	0.508	0.518	0.576	0.587	0.544	0.861	0.863	0.908	0.932	0.934	**0.899**
50K	DRS-KL	0.542	0.592	0.605	0.579	0.658	0.595	0.834	0.882	0.932	0.905	0.895	0.889
	DRS-SE	0.555	0.526	0.561	0.547	0.618	0.562	0.845	0.897	0.900	0.926	0.868	0.887
10K	DRS-KL	0.589	0.563	0.537	0.516	0.587	0.558	0.787	0.866	0.911	0.932	0.926	0.884
	DRS-SE	0.568	0.576	0.550	0.539	0.582	0.563	0.863	0.824	0.884	0.932	0.918	0.884
–	LDA	0.518	0.482	0.518	0.539	0.545	0.521	0.784	0.853	0.887	0.871	0.895	0.858
	sLDA	0.518	0.550	0.542	0.582	0.616	0.562	0.837	0.839	0.884	0.932	0.939	0.886

way of dealing with relative similarities, which sLDA cannot handle. Therefore, the computational costs in Table 3 show that our method can handle special kinds of metadata, i.e. relative similarities, and achieve good accuracy, all with reasonable computational costs compared with sLDA.

5.4 Qualitative Evaluation

In this subsection, we show the hidden topics learned by our model. For each dataset, we randomly choose four topics and show the top 10 words of each topic. Tables 4 and 5 show the top words generated by our model with 100K triplets on 20 newsgroups dataset and TDT2 dataset, respectively.

Table 4: Sample topics by our model on 20 newsgroups.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
gun	god	privacy	drive
control	people	encryption	disk
crime	jesus	internet	hard
guns	christian	anonymous	card
rate	bible	information	mb
weapons	man	government	drives
people	religion	email	apple
police	way	technology	system
manes	christ	mail	know
rates	religious	access	dos

We can see from both tables that the discovered topics are generally meaningful. For example, from the top words like "gun," "control" and "crime," it is easy to justify that Topic 1 in Table 4 is about "gun control"; Top topic words like "tobacco," "smoking" and "tax" demonstrate that Topic 4 in Table 5 is about "tobacco control".

Table 5: Sample topics by our model on TDT2.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**
iraq	israel	spkr	tobacco
united	israeli	voice	smoking
weapons	palestinian	president	tax
gulf	netanyahu	clinton	industry
iraqi	peace	news	companies
oil	talks	peterjennings	congress
saddam	palestinians	camera	money
gas	arafat	white	billion
war	west	house	settlement
ap	bank	abcnews	bill

6 Conclusions

In this paper, we perform topic modeling where relative similarities among documents are given. We formulate the constraints as a loss function and propose a general probabilistic model that combines LDA with such constraints. Our model treats the constraints as a regularizer of the log likelihood of text. Extensive experiments are conducted on two real world datasets, in which the empirical results show that our model not only learns meaningful topics, but also outperforms the baselines in terms of topic coherence and a document classification task.

The constraints we captured in this work are commonly used in the field of Distance Metric Learning (DML) [Gao *et al.*, 2014]. An interesting direction of future work is to combine the theory of DML with our model to better model documents with such constraints.

Acknowledgments

This work was done during Jianguang Du's visit to Singapore Management University. Dandan Song is the corresponding author. This work was partially supported by the National Program on Key Basic Research Project (973 Program, Grant No. 2013CB329600), National Natural Science Foundation of China (NSFC, Grant Nos. 61472040 and 60873237), and Beijing Higher Education Young Elite Teacher Project (Grant No. YETP1198).

References

[Andrzejewski *et al.*, 2011] David Andrzejewski, Xiaojin Zhu, Mark Craven, and Benjamin Recht. A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic. In *IJCAI*, volume 22, page 1171, 2011.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Cai *et al.*, 2008] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. Modeling hidden topics on document manifold. In *Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08)*, pages 911–920, 2008.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Chang *et al.*, 2009] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[Eisenstein *et al.*, 2010] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.

[Gao *et al.*, 2014] Xingyu Gao, Steven CH Hoi, Yongdong Zhang, Ji Wan, and Jintao Li. Soml: Sparse online metric learning with application to image retrieval. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[Huh and Fienberg, 2010] Seungil Huh and Stephen E. Fienberg. Discriminative topic modeling based on manifold learning. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 653–662. ACM, 2010.

[Liao *et al.*, 2014] Lizi Liao, Jing Jiang, Ying Ding, Heyan Huang, and Ee Peng LIM. Lifetime lexical variation in social media. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[McAuley and Leskovec, 2013] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[Mcauliffe and Blei, 2008] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[McCallum *et al.*, 2005] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. In *IJCAI*, 2005.

[Mei *et al.*, 2014] Shike Mei, Jun Zhu, and Jerry Zhu. Robust regbayes: Selectively incorporating first-order logic domain knowledge into bayesian models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 253–261, 2014.

[Mimno *et al.*, 2011] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[Newman *et al.*, 2010] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, 2010.

[Qiu *et al.*, 2013] Minghui Qiu, Feida Zhu, and Jing Jiang. It is not just what we say, but how we say them: Lda-based behavior-topic model. In *Proceedings of the 13th SIAM International Conference on Data Mining*, pages 794–802. SIAM, 2013.

[Rosen-Zvi *et al.*, 2004] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[Tang *et al.*, 2013] Jian Tang, Ming Zhang, and Qiaozhu Mei. One theme in all views: modeling consensus topics in multiple contexts. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 5–13. ACM, 2013.

[Wallach, 2006] Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

[Wang and McCallum, 2006] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.