The Dual-Sparse Topic Model: Mining Focused Topics and Focused Terms in Short Text

Tianyi Lin*
The Chinese University of Hong Kong
tylin@se.cuhk.edu.hk

Wentao Tian*
The Chinese University of Hong Kong
wttian@se.cuhk.edu.hk

Qiaozhu Mei
School of Information
University of Michigan
qmei@umich.edu

Hong Cheng
The Chinese University of Hong Kong
hcheng@se.cuhk.edu.hk

ABSTRACT

Topic modeling has been proved to be an effective method for exploratory text mining. It is a common assumption of most topic models that a document is generated from a mixture of topics. In real-world scenarios, individual documents usually concentrate on several salient topics instead of covering a wide variety of topics. A real topic also adopts a narrow range of terms instead of a wide coverage of the vocabulary. Understanding this sparsity of information is especially important for analyzing user-generated Web content and social media, which are featured as extremely short posts and condensed discussions.

In this paper, we propose a dual-sparse topic model that addresses the sparsity in both the topic mixtures and the word usage. By applying a "Spike and Slab" prior to decouple the sparsity and smoothness of the document-topic and topic-word distributions, we allow individual documents to select a few focused topics and a topic to select focused terms, respectively. Experiments on different genres of large corpora demonstrate that the dual-sparse topic model outperforms both classical topic models and existing sparsity-enhanced topic models. This improvement is especially notable on collections of short documents.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining

General Terms

Algorithms, Experimentation

*Tianyi Lin and Wentao Tian contributed equally to this work.

WWW'14, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2744-2/14/04.
http://dx.doi.org/10.1145/2566486.2567980.

Keywords

Topic modeling; spike and slab; sparse representation; user-generated content

1. INTRODUCTION

We are living in an era of information revolution where social media is gradually substituting the role of traditional media. Online social network sites such as Facebook and Twitter have emerged as new mediums of information diffusion, allowing their users to create and spread information much more effectively than before. According to recent statistics, more than 500 million tweets are posted by Twitter users on a daily basis [1]. This huge volume of user-generated content, normally in the form of very short documents, contains rich and useful information that can hardly be found in traditional information sources [31]. As a result, discovering meaningful knowledge from the large-scale user-generated short text in social media has been recognized as a challenging and promising research problem.

Statistical topic models have been proved to be effective tools for exploratory analysis of the overload of text content [4]. It is the common assumption of most classical topic models (e.g., [13, 6, 5]) that a document is generated from a mixture of topics and a topic samples words from a distribution over the vocabulary. Once estimated, the topic proportions of a document (or a collection of documents) can be used as a high-level representation of the semantics of that document (or collection), and the top-ranked words in a topic-word distribution can be used to interpret the semantics of that topic. By doing this, a topic model can provide an effective organization of latent semantics to the unstructured text collection.

While topic models have enjoyed broad success on traditional media, the experience on social media is mixed. Unlike carefully edited articles, user-generated content in social media is characterized with an extremely short document length, a very large vocabulary, and a broad range of topics. Consequentially, the word co-occurrence information at individual document level becomes much sparser, inevitably compromising the performance of computational methods that utilize this co-occurrence information, including topic

[1]https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

modeling [19, 24]. As a result, classical topic models usually yield suboptimal performance when applied 'as is' to short documents. Heuristic treatments such as document pooling [19] or contextualization [24] have to be applied to improve the performance of topic modeling on short text, both of which require the availability of additional context information (e.g., authors) beyond individual documents.

Does this imply a curse of topic modeling on short text? What if we are dealing with a collection of de-identified documents where it is hard to link individual documents through context variables (a common requirement of privacy-preserving data analysis, e.g., analyzing patient records)? We do notice that although there is generally a large number of topics spanning in the collection, an individual post usually concentrates on only a small number of topics. Similarly, a user-generated topic usually has a very skewed distribution of words, i.e., a topic usually focuses on a narrow range of words instead of a wide coverage of the vocabulary.

Understanding this skewness in both the topic mixtures and the word distributions is especially important for analyzing the content in short text. This inspires us to reconsider the assumptions of classical topic models. Instead of letting the topic mixtures and the word distributions navigate freely in the simplex, can we tell the model that each document focuses on a few topics, and each topic focuses on a few words? When there are many topics in a document, the sparse information is too little to be shared. But maybe it is enough to be split by only a few of them?

With this motivation, we address the *skewness*, or the *sparsity* of both the topic mixtures and the word distributions in topic modeling. This is achieved by a *dual-sparse topic model*, which applies a "Spike and Slab" prior that decouples the *sparsity* and the *smoothness* of the document-topic and topic-word distributions. By doing this, we allow individual documents to select only a few focused topics and also each topic to select its focused terms. In order to avoid the ill-posed definitions of the topic mixtures and the word distributions under the direct application of the "Spike and Slab" prior, we further introduce a *weak smoothing prior* along with a *smoothing prior*, to ensure the probabilistic distributions in the generative process are well-defined.

Unlike existing sparsity-enhanced topic models, the inference of our model can be done through a novel inference procedure called *zero-order collapsed variational Bayes inference* (CVB0), thus allows the model to efficiently deal with large-scale corpora. Experimental results on three real-world data sets demonstrate that the dual-sparse topic model outperforms both classical topic models and the state-of-the-art sparsity-enhanced topic models. The improvement is especially notable on collections of short documents.

The rest of the paper is organized as follows. We discuss related work in Section 2. In Section 3, we formally define the problem of dual-sparsity in short text. The dual-sparse topic model and the inference procedure are introduced in Section 4. We present the experimental results in Section 5 and conclude this work in Section 6.

2. RELATED WORK

To the best of our knowledge, this is the first study that simultaneously mines focused topics and focused terms from short text. This is achieved through addressing the dual-sparsity of topic mixtures and topic-word distributions in topic modeling. Our work is related to the following lines of literature.

2.1 Classical Probabilistic Topic Models

Classical probabilistic topic models like the probabilistic latent semantic analysis (PLSA) [13] and the latent Dirichlet allocation (LDA) [6] have been widely adopted in text mining. Without utilizing auxiliary information such as higher-level context, the classical topic models generally regard each document as an admixture of topics where each topic is defined as a unigram distribution over all the terms in the vocabulary. In practice, the benefit of LDA over PLSA comes from the *smoothing* process of the document-topic distributions and the topic-word distributions introduced by the Dirichlet prior, which thus alleviates the overfitting problem of PLSA. In order to relax the assumption that the user knows the number of topics *a priori*, Blei et al. proposed a hierarchical topic model utilizing the *Chinese Restaurant Process* for the construction of infinite number of topics [5]. These classical models generally lack the ability of directly controlling the posterior sparsity [11] of the inferred representations, thus fail to address the skewness of the topic mixtures and the word distributions. Indeed, one could enhance the *sparsity* by approaching the Dirichlet prior in LDA to zero. However, such a cruel process also inevitably results in a weakened effect of *smoothing*. Previous work has shown that simply applying a small Dirichlet prior is not only ineffective in controlling the posterior sparsity [32], but also results in compromised, less smooth document-topic and topic-term distributions [27]. In other words, a weakened Dirichlet smoothing yields sparsity only because of the scarcity of information. As a result, when applied to short documents, classical topic models are usually unable to perform as well as for professional documents, even when the Dirichlet priors are optimized.

2.2 Sparsity-Enhanced Topic Models

Recently, there have been efforts to address the problem of sparsity in topic distributions. These sparsity-enhanced topic models aim at extracting focused topics or focused terms in text by imposing sparsity regularization [27, 8, 29, 23, 30, 1, 32, 14, 28, 17]. The practices of these models can be summarized under two camps: (1) non-probabilistic coding or matrix factorization, and (2) probabilistic graphical topic model using specific prior or infinite stochastic process.

In the first category, coding is used to represent the coefficients of corresponding topical basis in topic space to model the generative process of words. For example, the non-probabilistic sparse coding [32] and the non-negative matrix factorization (NMF) [17, 14] provide a feasible framework to impose various sparsity constraints directly, which are at the expense of losing the probabilistic representations of topics. Regularized Latent Semantic Index (RLSI) [28] imposes sparsity regularization into the framework of LSI. However, the intrinsic limitation of LSI (compared to LDA) has predicted the compromised performance as the number of topics increases. One of the state-of-the-art methods in this category is the recently proposed Sparse Topical Coding (STC) [32], which learns both focused topics and focused terms. The method relies on the ability of a *Laplacian prior* to induce sparsity into the topic-word associations and automatically learns the sparse representations of the vocabulary. STC performs significantly better than previous models in-

cluding the classical LDA, NMF, and RegLDA [23], even when the number of topics is large. However, STC does not achieve the sparse topic representations of documents, i.e., the structure of focused topics.

The second category of sparsity-enhanced topic models extends the graphical structure of classical topic models. This camp of models is inspired by the subtle effect of the variation of the Dirichlet prior on the topic models [26]. In order to achieve sparse representation in the document-topic and topic-term distributions, Wang and Blei [27] and Williamson et al. [29, 30] introduced a *Spike and Slab prior* and the *Indian Buffet Process* to model the sparsity in finite and infinite latent topic structures of text. Similarly, Chen et al. [8] proposed a *context focused topic model (cFTM)* by using the Hierarchy Beta Process. Even though Indian Buffet Process and Hierarchy Beta are theoretically sound, the inference procedures for these models are much more complicated and intractable on large document collections. These models typically extract focused topics or focused terms independently, rather than considering the dual-sparsity of topics per document and terms per topic. The closest work to ours is IBP-LDA proposed by Archambeau et al. [1] which tries to model the sparse representation of document-topic and topic-term distributions via Indian Buffet Process. However, over complicated inference procedure has limited its capability of handling large-scale document collections.

2.3 The "Spike and Slab" Prior

It is noticeable that although Wang and Blei only address the sparsity of topic-word distributions (as we address both sparse topic representations for documents and sparse term representations for topics), the "Spike and Slab" prior they introduce to topic modeling is related to the key practice of our treatment [27]. The "Spike and Slab" prior [15] is a well established method in mathematics, which has been used in real world applications since it can successfully decouple the *sparsity* and *smoothness* of a probabilistic distribution [9, 3]. Specifically, by using auxiliary Bernoulli variables to represent the "on" and "off" of particular variables, the model can determine if the corresponding variables appear or not. In topic modeling, this indicates whether a topic is "selected" by the document (i.e., a focused topic), or whether a term is "selected" by a topic (i.e., a focused term). Similar ideas appear in other models like noisy-OR model [22] and aspect Bernoulli model [16].

When directly applied to topic models, however, the "Spike and Slab" prior may cause the probabilistic distributions to be ill-defined. Indeed, such a process introduces *never appearing* terms to the distribution of a topic in Wang and Blei [27], which not only imposes unnecessary difficulty into the inference procedure but also compromises the quality of topics. In our work, we define a weak smoothing prior along with a smoothing prior to avoid the ill-posed definitions of distributions by the direct application of the "Spike and Slab" prior. Our model thus results in a much simpler inference procedure and a better performance on large collections of documents.

3. PROBLEM FORMULATION

In this section, we formally define the problem of **Dual-Sparse Topic Modeling**.

Let $D = \{\mathbf{w}^d\}_{d=1}^{|D|}$ be a collection of documents, where $\mathbf{w}^d = (w_1^d, w_2^d, \ldots, w_{|V|}^d)$ is a vector of terms representing the textual content of document d. w_i^d denotes the frequency of the i-th term in document d, and $|V|$ is the size of the vocabulary.

Definition 1. **(Topic, Topic Representation, Topic Modeling)** A *topic* $\vec{\phi}$ in a given collection D is defined as a multinomial distribution over the vocabulary V, i.e., $\{p(w|\vec{\phi})\}_{w \in V}$. It is a common assumption that there are K topics in D. The *topic representation* of a document d, $\vec{\theta_d}$, is defined as a multinomial distribution over K topics, i.e., $\{p(\vec{\phi_k}|\vec{\theta_d})\}_{k=1,\ldots,K}$. The general task of *topic modeling* aims to find K salient topics $\{\vec{\phi_k}\}_{k=1,\ldots,K}$ from D and to find the topic representation of each document.

Most classical probabilistic topic models adopt the Dirichlet prior for both the topics and the topic representation of documents, which are first proposed in LDA [6]. That is, $\vec{\theta_d} \sim \text{Dirichlet}(\vec{\alpha})$ and $\vec{\phi_k} \sim \text{Dirichlet}(\vec{\beta})$. In practice, the Dirichlet prior smooths the topic mixture in individual documents and the word distribution of each topic, which alleviates the overfitting problem of PLSA especially when the number of topics and the size of vocabulary increase. However, the Dirichlet prior itself does not formally control the posterior sparsity of the inferred representations as discussed before. Weakening the Dirichlet prior increases the sparsity at the expense of the smoothness of the document/topic representations. Specifically, for any term r and a topic k, $\phi_{kr} = 0$ is totally possible since each topic typically focuses on a subset of terms rather than covering all of them. Similarly, a document also focuses on a few topics instead of covering all the topics in the collection. These topics and terms are intuitively interpreted as *focused topics* of a document and *focused terms* of a topic. Before giving a formal definition of focused topics and terms, we define some auxiliary variables allowing each document to select its representative topics, and a topic to select its related terms.

Definition 2. **(Topic Selector, Term Selector)** For $d \in \{1, \ldots, |D|\}$, $k \in \{1, \ldots, K\}$, a *topic selector* α_{dk} is a binary variable that indicates whether topic k is relevant to document d. α_{dk} is sampled from Bernoulli(a_d), where a_d is a Bernoulli parameter. Similarly, for $k \in \{1, \ldots, K\}$, $r \in \{1, \ldots, |V|\}$, a *term selector* β_{kr} indicates if term r is included in topic k. β_{kr} is sampled from Bernoulli(b_k), where b_k is another Bernoulli parameter.

Definition 3. **(Smoothing Prior, Weak Smoothing Prior)** The *smoothing prior* is a pair of Dirichlet hyperparameters π and γ that are used to smooth those topics and terms that are selected by the topic selector and the term selector, respectively. The *weak smoothing prior* is another pair of Dirichlet hyperparameters $\bar{\pi}$ and $\bar{\gamma}$ that are used to smooth those topics and terms that are *not* appearing in the corresponding document and topic (i.e., not selected by the topic selector and the term selector), respectively.

The topic selector and term selector are referred to as "spikes," while smoothing prior and weak smoothing prior correspond to "slabs" in statistics. In this way, we are able to decouple the sparsity and smoothness of topic proportion and topic distribution by applying a "Spike and Slab" prior. The application of the "Spike and Slab" prior to topic modeling is however not trivial. The Bernoulli selectors may cause

the multinomial distributions to be ill-defined, where some topics never appear in the collection, and some terms never appear in the topics. Indeed, this problem can be alleviated by the "slab," however too little "slabbing" brings little effect of smoothing while too much "slabbing" compromises the effect of "spiking" (sparsity).

We propose a treatment to separate the *smoothing prior* and *weak smoothing prior*, which is a key contribution of the proposed model. The topics and terms selected by the "spike" are smoothed using the stronger *smoothing prior* π and γ, which successfully avoid the "*less smooth*" problem as in [27]. Topics and terms which are not selected by the spike are smoothed through the weak smoothing prior $\bar{\pi}$ and $\bar{\gamma}$. Since $\bar{\pi} \ll \pi$ and $\bar{\gamma} \ll \gamma$, we can easily maintain the effect of sparsity while also fixing the ill-definition of the distributions. By doing this, the sparse representation achieved is not induced from data scarcity [32] since we have selected topics and terms by "spike" before smoothing. Clearly, through this treatment the enhancement of sparsity no longer compromises the effect of smoothing in the document/topic representations. More details will be discussed in Section 4.1.

With the definition of topic selectors and term selectors, we are now able to formally define focused topics and focused terms in topic modeling.

Definition 4. **(Focused Topic, Focused Term)** Topic k is a *focused topic* of document d if the topic selector $\alpha_{dk} = 1$, and term r is a *focused term* of topic k if the term selector $\beta_{kr} = 1$. For document d, $A_d = \{k : \alpha_{dk} = 1\}$ is defined as the set of its focused topics, and for topic k, $B_k = \{r : \beta_{kr} = 1\}$ is defined as the set of its focused terms.

Clearly, the set of focused topics provides a sparse representation of the semantics of a document, and the set of focused terms provides a sparse representation of the semantics of a topic.

Definition 5. **(Dual-Sparsity)** The sparsity of a document-topic distribution exists if $|A_d| < K$, which is defined as $\text{sparsity}(d) \triangleq 1 - \frac{|A_d|}{K}$. The sparsity of a topic-term distribution exists if $|B_k| < |V|$, which is defined as $\text{sparsity}(k) \triangleq 1 - \frac{|B_k|}{|V|}$. The *Dual-Sparsity* denotes the joint effect when both types of sparsity exist.

Given a collection of documents D, the vocabulary V, and the predefined number of topics K, the major tasks of **Dual-Sparse Topic Modeling** can be defined as to:

1. determine the set of focused topics and focused terms by estimating the topic selector $\vec{\alpha}$ and term selector $\vec{\beta}$;
2. further infer the Bernoulli parameter $\vec{a}$ and $\vec{b}$ to present dual-sparsity quantitatively;
3. learn the sparse word representation of topics $\vec{\phi}$;
4. learn the sparse topic representation of documents $\vec{\theta}$.

All the notations used in this paper are summarized in Table 1.

Table 1: Variables and Notations

Notation	Meaning
K	number of topics
V	vocabulary
D	collection of short documents
N_d	the length of document d
α_{dk}	topic selector
A_d^{Θ}	number of focused topics for document d
a_d	probability of topic selector
s, t	parameters of a_d
π	topic smoothing prior
$\bar{\pi}$	weak topic smoothing prior
$\vec{\theta}_d$	document-topic distribution
B_k^{Φ}	number of focused terms for topic k
$\mathbf{Z}$	topic assignments
b_k	probability of term selector
x, y	parameters of b_k
β_{kr}	term selector
γ	term smoothing prior
$\bar{\gamma}$	weak term smoothing prior
$\vec{\phi}_k$	topic distribution
$n_{d,r}^k$	frequency of term r assigned to topic k in document d
$\mathbf{W}$	words
$\mathrm{I}[\cdot]$	indicator function

4. LEARNING THE DUAL-SPARSITY

Dual-sparsity is commonly observed in short text, such as user-generated Web content and social media. It brings great challenges to the classical topic models in learning the sparse representations of documents and topics. To address this problem, we propose to modify the machinery of traditional topic models by imposing the sparsity. Importantly, we need to take the feasibility of the inference procedure into consideration in designing the new model.

Zhu and Xing [32] have shown it is unlikely to control the posterior sparsity effectively even if we impose weak Dirichlet smoothing prior to all topics. Such a method also results in *less smooth* expected document-topic and topic-term distributions [27]. One needs a better solution to decouple the sparsity and smoothness of document and topic representations in the graphical structure.

We propose a new topic model, named **Dual-Sparse Topic Model (DsparseTM)**, to find the focused topics and focused terms. This model is largely inspired by the "Spike and Slab" prior [15]. The most important difference between DsparseTM and prior work is that we successfully avoid the ill-posed singularity of document-topic and topic-term distributions by the utilization of a smoothing prior and a weak smoothing prior. Specifically, a pair of weak smoothing prior $\bar{\pi}$ and $\bar{\gamma}$ is proposed to deal with the problem that Dirichlet($\pi\vec{\alpha}_d$) and Dirichlet($\gamma\vec{\beta}_k$) are ill-posed when $\vec{\alpha}_d = \vec{\beta}_k = \mathbf{0}$.

4.1 The Dual-Sparse Topic Model

The key idea of the dual-sparse topic model is to restrict the size of the topic simplex and the word simplex over Dirichlet distributions in order to induce sparsity. This is done through auxiliary Bernoulli variables. Specifically, Bernoulli variables indicating "on" and "off" of given variables are used to determine whether a topic is a focused topic, or a term is a focused term. Smoothing priors are introduced to smooth focused topics and focused terms.

DsparseTM is depicted in Figure 1 and the probabilistic generative process is presented as follows:

For each topic $k \in \{1, 2, \ldots, K\}$:

1. $b_k \sim \text{Beta}(x, y)$;
2. For each term $r \in \{1, 2, \ldots, |V|\}$:
   (a) the term selector $\beta_{kr} \sim \text{Bernoulli}(b_k)$, $\vec{\beta}_k = \{\beta_{kr}\}_{r=1}^{|V|}$;

(b) the set of focused terms: $B_k = \{r : \beta_{kr} = 1\}$;

(c) the topic distribution $\vec{\phi}_k \sim \text{Dirichlet}(\gamma\vec{\beta}_k + \bar{\gamma}\vec{1})$;

For document $d \in \{1, 2, \ldots, |D|\}$:

1. $a_d \sim \text{Beta}(s, t)$;
2. For each topic $k \in \{1, 2, \ldots, K\}$:
   (a) the topic selector $\alpha_{dk} \sim \text{Bernoulli}(a_d)$, $\vec{\alpha}_d = \{\alpha_{dk}\}_{k=1}^{K}$;
   (b) the set of focused topics: $A_d = \{k : \alpha_{dk} = 1\}$;
3. the topic proportion $\vec{\theta}_d \sim \text{Dirichlet}(\pi\vec{\alpha}_d + \bar{\pi}\vec{1})$;
4. For each word $i \in \{1, 2, \ldots, N_d\}$:
   (a) sample z_{di} from Multinomial$(\{\vec{\theta}_k : k \in A_d\})$;
   (b) sample w_{di} from Multinomial$(\{\vec{\phi}_{z_{d,i}} : z_{di} \in B_k\})$;

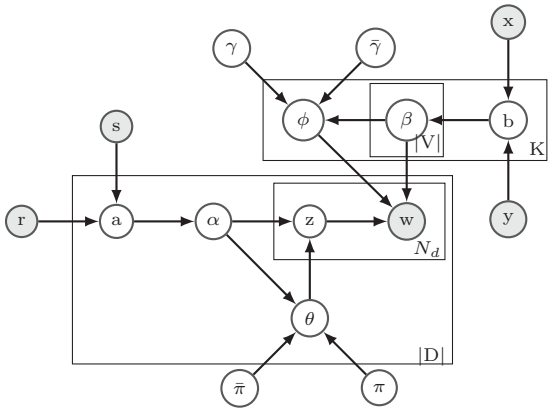

Figure 1: The graphical model of DsparseTM

We make the following remarks:

Smoothing Priors: It is necessary to understand the rationale behind the weak smoothing prior $\bar{\pi}$, $\bar{\gamma}$ along with the stronger smoothing prior π, γ. Previous work [6] has found it reasonable to use Dirichlet distribution to define both document-topic and topic-term distributions. However, Dirichlet distributions will be ill-posed as the Dirichlet hyperparameter approaches zero. Indeed, Dirichlet$(\pi\vec{\alpha}_d)$ or Dirichlet$(\gamma\vec{\beta}_k)$ are not well-defined when $\vec{\alpha}_d = 0$ or $\vec{\beta}_d = 0$. In order to address this nuisance, Wang and Blei [27] define $\beta_{k,|V|+1} = \text{I}[\sum_{r=1}^{|V|} \beta_{kr} = 0]$ to ensure that Dirichlet$(\gamma\vec{\beta}_k)$ is well-defined. In other words, they assume that there is a $|V| + 1$-th term which *never appears* in documents. When it comes to the Dirichlet distribution of topic proportions, however, to define a *never appearing* topic will lead to high computational cost. In this paper, we choose to introduce weak smoothing prior $\bar{\pi}$ and $\bar{\gamma}$ so that Dirichlet$(\gamma\vec{\beta}_k + \bar{\gamma}\vec{1})$ and Dirichlet$(\pi\vec{\alpha}_d + \bar{\pi}\vec{1})$ are well defined even when all $\vec{\alpha}_d = 0$ and $\vec{\beta}_k = 0$.

After defining well-posed document-topic and topic-word distributions, the model further samples z and w from the multinomial distributions restricted on $A_d = \{k : \alpha_{dk} = 1\}$ and $B_k = \{r : \beta_{kr} = 1\}$. Since $\bar{\pi} \ll \pi$ and $\bar{\gamma} \ll \gamma$, this does not contradict the definition of a multinomial distribution. Indeed, we can get $\sum_{k \in A_d} \theta_{dk} = 1$ and $\sum_{r \in B_k} \phi_{kr} = 1$ in numerical sense when $\bar{\pi} = 10^{-7}$ and $\bar{\gamma} = 10^{-7}$.

Generative process: After sampling the Bernoulli parameter b_k from Beta(x, y) for the k-th topic, we can generate different binary variables β_{kr} for all the terms in vocabulary. Whether a term r is a focused term of topic k is determined by the value of β_{kr}. Then we can sample the topic-term distribution $\vec{\phi}_k$ from Dirichlet$(\gamma\vec{\beta}_k + \bar{\gamma}\vec{1})$.

For document d, the corresponding Bernoulli parameter a_d sampled from Beta(s, t) will generate a series of binary α_{dk} to select subset of focused topics. $\vec{\theta}_d \sim \text{Dirichlet}(\pi\vec{\alpha}_d + \bar{\pi}\vec{1})$ shows the admixture proportion of document d over all topics, and we only choose θ_{dk} when $\alpha_{dk} = 1$. Indeed, it is reasonable to assume that topics with $\alpha_{dk} = 0$ do not appear in document d since their smoothing prior $\bar{\pi}$ is so weak that $\bar{\pi} \ll \pi$. In other word, $A_d = \{k : \alpha_{dk} = 1\}$ is defined as a set of focused topics.

Finally we sample z from Multinomial$(\{\vec{\theta}_k : k \in A_d\})$. The number of words in d, N_d can be notated as a summation of $n^k_{d,(.)}\text{I}[k \in A_d]$. In this case, topics never appearing in document d do not contribute to N_d. Similar idea motivates the definition of $n^k_{(.),r}\text{I}[r \in B_k]$. In the end, we sample a word w from Multinomial$(\{\vec{\phi}_{z_{d,i}} : z_{di} \in B_k\})$.

4.2 Inference

Since the posterior inference is intractable for DsparseTM, we need to find an algorithm for posterior inference that is both effective and efficient for large document collections. We adopt the well-known **Zero-Order Collapsed Variational Bayes Inference Algorithm (CVB0)** [2] for this task. Indeed, the convergence of CVB0 has been theoretically proven to be stable [21], and its practical performance is better than the Collapsed Gibbs Sampling Algorithm [12] and the Collapsed Variational Bayes Inference Algorithm (CVB) [25].

Integrating out document-topic distribution $\vec{\theta}$, topic-term distribution $\vec{\phi}$, and Bernoulli parameters $\vec{a}$, $\vec{b}$ analytically, we get the remaining variables for inference: topic assignment $\mathbf{z}$, topic selector $\vec{\alpha}$, and term selector $\vec{\beta}$. Applying the framework of CVB0, we get the updated equations for variational parameters of $\mathbf{z}$, $\vec{\alpha}$ and $\vec{\beta}$:

Variational Bernoulli distribution for $\vec{\alpha}$:

$$\begin{aligned}
\hat{a}_{jk} =& \frac{a^1_{jk}}{a^1_{jk} + a^0_{jk}} \\
\tilde{a}^1_{jk} =& (s + A_j^{\Theta\neg jk})\Gamma(N^{\Theta}_{jk} + \pi + \bar{\pi}) \\
& B(\pi + K\bar{\pi} + \pi A_j^{\Theta\neg jk}, N^{\Theta}_j + \pi A_j^{\Theta\neg jk} + K\bar{\pi}) \\
\tilde{a}^0_{jk} =& (t + K - 1 - A_j^{\Theta\neg jk})\Gamma(\pi + \bar{\pi}) \\
& B(K\bar{\pi} + \pi A_j^{\Theta\neg jk}, N^{\Theta}_j + \pi + \pi A_j^{\Theta\neg jk} + K\bar{\pi})
\end{aligned} \tag{1}$$

where $A^{\Theta}_j = \sum_{k'} \hat{a}_{jk'}$, $N^{\Theta}_{jk} = \sum_{i'} \gamma_{i'jk}$, $N^{\Theta}_j = \sum_{i'k'} \gamma_{i'jk'}$ and $\neg jk$ means without α_{jk}.

Variational Bernoulli distribution for $\vec{\beta}$:

$$\begin{aligned}
\hat{b}_{kr} =& \frac{b^1_{kr}}{b^1_{kr} + b^0_{kr}} \\
\tilde{b}^1_{kr} =& (x + B_k^{\Phi\neg kr})\Gamma(N^{\Phi}_{kr} + \gamma + \bar{\gamma}) \\
& B(\gamma + |V|\bar{\gamma} + \gamma B_k^{\Phi\neg kr}, N^{\Phi}_k + \gamma B_k^{\Phi\neg kr} + |V|\bar{\gamma}) \\
\tilde{b}^0_{kr} =& (y + |V| - 1 - B_k^{\Phi\neg kr})\Gamma(\gamma + \bar{\gamma}) \\
& B(|V|\bar{\gamma} + \gamma B_k^{\Phi\neg kr}, N^{\Phi}_k + \gamma + \gamma B_k^{\Phi\neg kr} + |V|\bar{\gamma})
\end{aligned} \tag{2}$$

where $B_k^{\Phi} = \sum_{r'} \hat{b}_{kr'}$, $N_{kr}^{\Phi} = \sum_{i'j'} \mathrm{I}[w_{i'j'} = r]\gamma_{i'j'k}$, $N_k^{\Phi} = \sum_{i'j'} \gamma_{i'j'k}$ and $\neg kr$ means without β_{kr}.

Variational Multinomial distribution for z:

$$\gamma_{ijk} = \hat{q}(z_{ij} = k) \propto \frac{N_{kw_{ij}}^{\Phi\neg ij} + \gamma\hat{b}_{kw_{i,j}} + \bar{\gamma}}{N_k^{\Phi\neg ij} + \gamma B_k^{\Phi} + |V|\bar{\gamma}}(N_{jk}^{\Theta\neg ij} + \pi\hat{a}_{jk} + \bar{\pi}) \quad (3)$$

Please refer to the Appendix for the details of derivation.

5. EXPERIMENT

In this section, we investigate the performance of DsparseTM on three large collections of documents. We are particularly interested in the effectiveness of DsparseTM on short text. The objectives of the experiments include: (1) a quantitative evaluation of the quality of extracted topics; (2) a quantitative evaluation of the sparse topic representations of documents; and (3) interpreting focused topics, focused terms, and dual-sparsity discovered by DsparseTM.

5.1 Data Sets

We adopt three different genres of real-world data sets for our experiments. We design text classification tasks to evaluate the performance of the sparse topic representations of documents. To do this, we select document collections with explicit class labels (articles from 20 Newsgroups [2] and titles of papers from selected computer science conferences). In order to evaluate the performance of the proposed model on short text collections, we select titles of scientific papers and microblogs (tweets) from Twitter.com. Stop words are removed from each data set according to a standard list of stop words [3].

- **DBLP**. Titles of scientific papers are good examples of short documents. We collect titles of all conference papers from the DBLP database [4] in three research areas: (1) database/data mining/information retrieval (**DB/DM/IR**), (2) theoretical computer science (**TCS**), and (3) computer networks/systems. This data set contains 40,190 short documents and 9,393 unique words, with labels of 22 different conferences.
- **20 Newsgroups**. This data set, denoted as 20NG, contains 18,774 newsgroup documents labeled in 20 categories, with a vocabulary of 60,698 unique words.
- **Twitter**. We sample a collection of 1,119,464 tweets posted in June 2009 from the Twitter data set released by the Stanford Network Analysis Project [5], denoted as TWITTER. After removing words that appeared less than 15 times, we yield a vocabulary of 32,641 words. In addition, we sample another collection with 13,080 users who have posted at least 50 tweets, resulting in 1,110,303 tweets. The second collection of tweets can be converted into "pseudo-documents" by author-wise pooling, which is denoted as TWITTER-A.

The statistics of the data sets are summarized in Table 2.

[2]http://qwone.com/~jason/20Newsgroups/
[3]http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop
[4]http://www.informatik.uni-trier.de/~ley/db/
[5]http://snap.stanford.edu/data/twitter7.html

Table 2: Statistics of the data sets

Data set	# Documents	Vocabulary size	Avg doc len by words
DBLP	40,190	9,393	5.7
20NG	18,774	60,698	114.8
TWITTER	1,119,464	32,641	4.9
TWITTER-A	13,080	15,952	451.4

5.2 Metrics

Finding an objective metric to compare the quality of topics is hard. Some commonly used metrics such as the perplexity or the likelihood of held-out data cannot directly measure the semantic coherence of the learned topics. Chang et al. [7] present quantitative methods to measure the topical coherence of the learned topics. They found that the likelihood of the held-out data is not always a good indicator of topic coherence. Recently, measuring the semantic coherence of the learned topics has received increasing attention [20, 24]. We adopt the same topic coherence metric for the comparison of topic models.

Topic Coherence. In [20], Newman et al. propose to use the point-wise mutual information (PMI) to measure the semantic coherence of topics. For a given topic T, we choose the top-N most probable words $w_1, w_2, \ldots, w_N$, and calculate the average relatedness of each pair of these words as the PMI score:

$$\text{PMI-Score}(T) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \quad (4)$$

where $p(w_i, w_j)$ is the joint probability of words w_i and w_j co-occurring in the same document, while $p(w_i)$ is the marginal probability of word w_i appearing in a document. These probabilities are computed from a much larger corpus. We set $N = 15$ in our analysis.

Classification Accuracy. One of the most important products of topic modeling is the topic proportion of each document, which provides a latent semantic representation of that document. Such a low-dimensional representation can be used in external text mining tasks such as text classification [18]. We evaluate the effectiveness of the topic representation of documents through the accuracy of a text classification task on the document collection D:

$$\text{Accuracy}(D) = \frac{1}{|D|} \sum_{d \in D} \mathrm{I}[\text{Label}_d = \text{Prediction}_d], \quad (5)$$

where $\mathrm{I}[\cdot]$ is the indicator function, Label_d and Prediction_d are the true label and the predicted label of document d in a text classification task, respectively.

Dual-Sparsity Ratio. Dual-sparsity refers to the joint sparsity of document-topic and topic-term distributions. The sparsity ratio of a document-topic distribution is defined as the expectation of sparsity(d) conditioned on Bernoulli parameter a_d, and the sparsity ratio of a topic-term distribution is similarly defined as the expectation of sparsity(k) conditioned on Bernoulli parameter b_k, i.e.,

$$\begin{aligned} \text{Sparsity-ratio}(d) &\triangleq \mathbf{E}[\text{sparsity}(d)] = 1 - a_d \\ \text{Sparsity-ratio}(k) &\triangleq \mathbf{E}[\text{sparsity}(k)] = 1 - b_k \end{aligned} \quad (6)$$

The Dual-Sparsity Ratio provides a direct measurement of the sparsity of the topic representation of documents and the word representation of topics.

5.3 Candidate Models for Comparison

We compare DsparseTM with the following models.

- **LDA**. The classical topic model LDA can induce sparsity as the Dirichlet prior approaches zero. We use the LDA package [6] with variational bayes inference, which automatically optimizes the Dirichlet hyperparameter α by using the Newton-Raphson method [6].
- **Sparse Topical Coding (STC)**. STC is a recently published sparsity-enhanced topic model which has been proven to perform better than many existing models, including NMF and RegLDA [32]. We use the implementation of STC with ℓ_2-norm provided by the authors [7].
- **Mixture of Unigrams**. It is also interesting to include the following extreme scenario into comparison. The mixture of unigrams [6] assumes that each document is generated by only one topic z which generates N words independently from the conditional multinomial $p(w|z)$. The probability of a document is:

$$p(\mathbf{w}) = \sum_z p(z) \prod_{n=1}^{N} p(w_n|z)$$

Clearly, this simple model forces the topic representation of a document to adopt the largest sparsity. One may imagine that for some certain collections of short documents, it may be a reasonable assumption that each document only contains one topic.

Note that the inclusion of the state-of-the-art sparsity-enhanced topic model (i.e., STC) has allowed us to omit other existing models from comparison. The comparison between STC and other models is reported in [32]. Some related models such as sparseTM [27] are also excluded from the comparison because of their nonparametric machinery. Specifically, sparseTM can learn the number of topics via Hierarchical Dirichlet Process. Furthermore, it only provides a sparse word representation of topics but no sparse topic representation of documents. Therefore, it is hard to compare DsparseTM and sparseTM fairly.

For the proposed DsparseTM, we sample the Bernoulli parameters $\vec{a}$ and $\vec{b}$ from uniform Beta distribution Beta(1,1) one by one and simply fix $\bar{\pi} = \bar{\gamma} = 10^{-7}$, although these hyperparameters may be further optimized.

5.4 Experimental Results

5.4.1 Topic coherence

The PMI scores of all candidate methods are presented in Table 3. The numbers of topics extracted from DBLP, 20 Newsgroups and Twitter are 15, 120 and 200, respectively.

Table 3: Topic coherence (PMI) on four data sets

	DBLP	20NG	Twitter	Twitter-A
Number of topics	15	120	200	200
DsparseTM	**0.871**	**1.621**	1.051	**1.939**
LDA	0.622	1.336	0.562	1.757
STC	0.088	1.515	0.378	1.192
Mixture of unigrams	0.532	0.691	**1.121**	0.823
Conference topic	0.586	-	-	-

[6]http://www.cs.princeton.edu/∼blei/lda-c/
[7]http://www.ml-thu.net/∼jun/stc.shtml

We make the following remarks:

DBLP. The proposed DsparseTM model yields the highest PMI score, followed by LDA and the mixture of unigrams, all of which outperform STC by a large margin. DBLP is an interesting data set where the documents (titles of papers) are short but still cover multiple topics. How good are the topic coherence scores? We provide a reference coherence score by aggregating the documents from each conference and treating that as a "topic." Clearly, this gives us a "naturally" defined topic that is interpretable by human. The average PMI score of the 22 conference based topics is 0.586. It is interesting to see that both DsparseTM and LDA achieve higher PMI scores than the conference based topics.

Even though the hyperparameters of LDA are optimized, it is still outperformed by DsparseTM. This is reasonable as short documents fail to provide sufficient statistics of word co-occurrence. The mixture of unigrams does not perform well, because although the titles of scientific papers are short, they often still cover more than one topic. The poor performance of STC possibly indicates that their sparsity-induced prior is not able to detect serious document-topic sparsity in DBLP titles. Note that STC is a non-probabilistic topic model. It is also likely that losing the probabilistic representation compromises the interpretability of topics.

20 Newsgroups. Again DsparseTM achieves the best topic coherence, followed by STC and LDA, and all outperform the mixture of unigrams by a large margin. As documents in 20NG are typically longer and may contain more topics, the performance of the mixture of unigrams suffers.

Both DsparseTM and STC harvest more coherence topics than LDA, which provides a strong evidence of the existence of dual-sparsity in 20 Newsgroups. It is nice to see that DsparseTM performs well on both short documents and long documents. To further investigate the behavior of DsparseTM on short text, we vary the length of documents in 20NG by randomly sampling words from the original documents. The topic coherence scores on these collections of shortened documents are reported in Figure 2 [8].

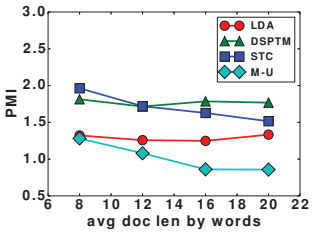

Figure 2: PMI on shortened 20NG documents. DsparseTM achieves high topic coherence that is robust to the variation of document length.

Interestingly, when the length of documents increases, the performance of STC and the mixture of unigrams drops and the performance of LDA increases. The performance of DsparseTM remains high and stable. This promising result indicates that DsparseTM successfully adapts to the sparsity of the data, and is not affected by the insufficient observations of word co-occurrence.

Note that STC achieves the best performance when there are only 8 words in a document. The reason that STC works well with extremely insufficient word co-occurrence

[8]Because of the much shorter documents, we learned 30 topics instead of 120 in this experiment.

may attribute to its use of the sparsity-induced prior, which is indeed simpler than the use of weak smoothing prior and smoothing prior in DsparseTM.

Twitter. It is surprising that the simplest model, the mixture of unigrams, performs the best on the collection of tweets. One could imagine that the topic proportions in short tweets are close to the extreme case where each tweet contains only one topic (compared to DBLP titles that cover multiple topics). Even on this data set, DsparseTM achieves a very close PMI score, which significantly outperforms LDA as well as STC.

Twitter-A. It is also interesting to investigate how much benefit the availability of auxiliary context information can bring to topic models. Indeed, when individual tweets written by the same author are pooled into "pseudo-documents," the performance of LDA improved significantly. Note that the mixture of unigrams, the best performer on short tweets, suffered from a drop of performance after pooling. This is because although it is a fair assumption that a tweet contains only one topic, a twitter user may still write about multiple topics. After author-wise pooling, DsparseTM still achieves the best performance, which once again demonstrates the effectiveness of the proposed model on both short and long documents.

5.4.2 Classification accuracy

The second task is to evaluate the effectiveness of the topic representation of documents. To do this, we perform text classification tasks on DBLP and 20NG, using the topic proportions of a document as the feature representation of that document, as an alternative to the conventional representation using bag of words. The latent semantic representation of documents enhances the classification performance when the training data is rare [18]. We use the 3 areas of DBLP conferences and the 20 newsgroups as classification categories, respectively.

On DBLP, we use 80% documents for training and 20% for testing. On the 20NG data set, we use 60% documents for training and 40% for testing, which is the same configuration as in [32]. With the topic feature representation, the documents are classified by a multi-class SVM [10]. A 5-fold cross-validation on the training data is used to select the optimal parameters of SVM. To better understand the behavior of topic representations in classification, we vary the ratio of labeled documents by sampling from the *training* set (from 0.2% to 100%). Besides the topic models, we also include a baseline which represents documents using the conventional term frequencies (TF).

Figure 3 reports the classification accuracy under different sampling ratios of training data. We can observe that DsparseTM consistently outperforms LDA, STC, and the mixture of unigrams in most settings and on both data sets. When there are sufficient training examples, the simple representation using term frequencies outperforms all the latent topic representations. Topic representation of documents plays an important role when the training examples are rare, where keyword features are likely to overfit the classifier to the data. This result is consistent with the conclusion in literature [18].

Shortened 20NG Documents. Like in the analysis of topic coherence, we also repeat this experiment by varying the length of 20 Newsgroups documents. Here we only report the setting with 1% sample of training documents. The classification results are reported in Figure 4. Clearly, DsparseTM consistently outperforms the other candidate models including the TF representation when the average document length varies among 8, 12, 16, and 20 words. Again, this demonstrates that DsparseTM can successfully model dual-sparsity in text, which is robust to the variation of document lengths. This indicates that DsparseTM is potentially very useful in text classification tasks with short documents and limited training data.

5.4.3 Characters of dual sparse representation

Finally, we present selected focused topics, focused terms, and dual sparse representations discovered by DsparseTM.

Tables 4 and 5 present the average sparsity ratio of topic representation for documents in selected categories in DBLP and 20 Newsgroups, respectively. We also list the most common focused topics of documents in each category. In Table 4, the average sparsity ratio of titles in the DB/DM/IR area is lower than that of the Theory area and the Networking/System area. This result is reasonable since DB/DM/IR covers a wider range of topics. Similarly, in Table 5, it is reasonable to see that the sparsity of topic mixture of "comp.graphics" and "comp.os.ms-windows.misc" is higher than that of "soc.religion.christian" and "talk.politics.mideast" since the number of topics related to computer technology is smaller than that related to religion and politics. Compared with DBLP, the sparsity of topic representations in 20 Newsgroups is lower. This reassures our observation that a newsgroup document covers a wider range of topics than a title of scientific paper.

Table 6: Focused terms and sparsity ratio of selected topics on DBLP

T7	sparsity: 0.9210	# Focused terms: 124
	detection social-network analysis network data online	
T9	sparsity: 0.9010	# Focused terms: 144
	algorithms approximation problems trees parallel algorithm	
T15	sparsity: 0.8873	# Focused terms: 185
	web search information semantic content user mining	

Table 7: Focused terms and sparsity ratio of selected topics on 20NG

T72	sparsity: 0.9536	# Focused terms: 288
	god religion atheists exist atheism evidence people	
T63	sparsity: 0.9540	# Focused terms: 214
	space nasa earth orbit shuttle mission lunar spacecraft	
T45	sparsity: 0.9661	# Focused terms: 125
	pain doctor day disease medical patients treatment blood	

Tables 6 and 7 present selected topics, with the sparsity ratio of the corresponding topic-word distribution as well as focused terms. In DBLP, topic 7 representing "social network analysis" focuses on a smaller set of terms than topic 15, a broader research topic on Web mining. In contrast to the higher sparsity of topic representations of documents, the sparsity of word distributions of topics in DBLP is lower than that in 20 Newsgroups. A possible explanation is that the vocabulary of social media is much larger than that of academic titles.

Tables 8 and 9 show the most frequent terms in each category and the topics which select these terms as focused terms. Firstly, we can observe that some words (e.g., "jpeg" and "wireless") are selected as focused terms by very few topics, indicating a narrow usage of those terms. Instead, some words (e.g., "system") are selected as focused terms by

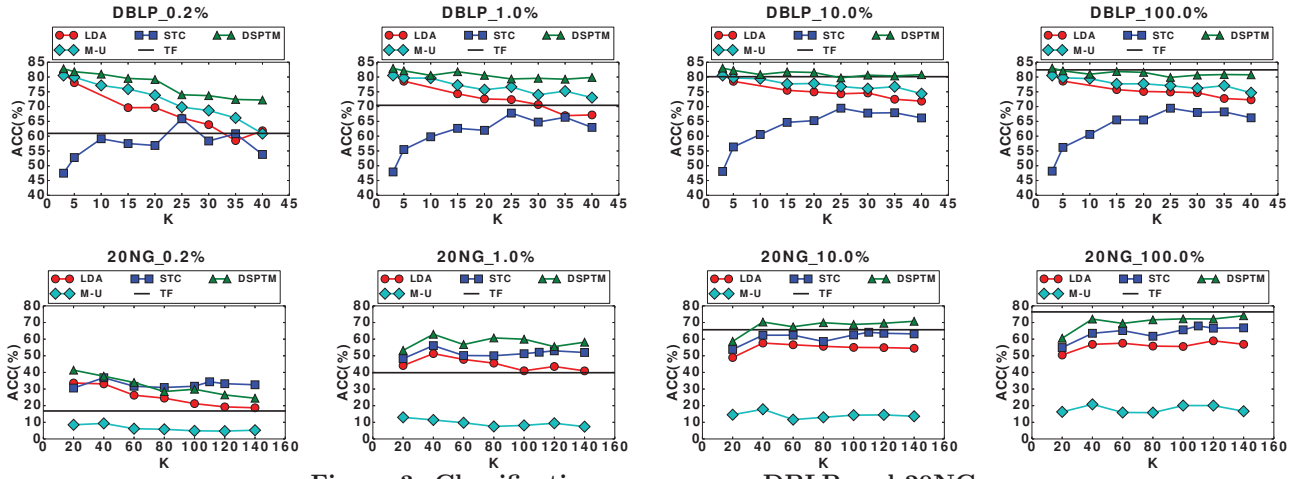

Figure 3: Classification accuracy on DBLP and 20NG

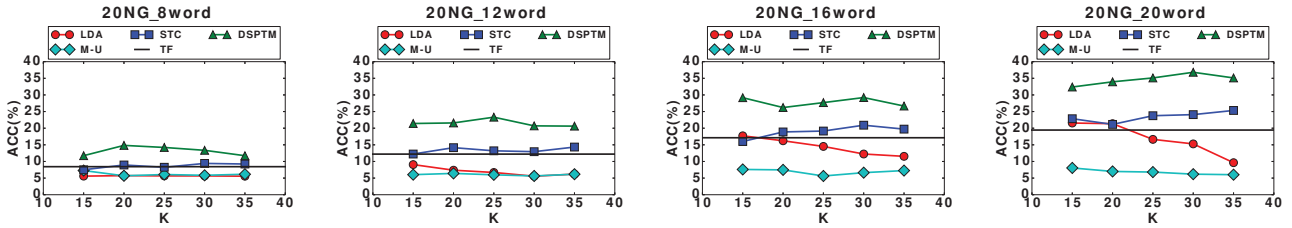

Figure 4: Classification accuracy on shortened 20NG documents with 1% training data

Table 4: Focused topics and average sparsity ratio for different DBLP categories

DB/DM/IR		TCS		Networking/System	
Avg. sparsity ratio: 0.9927		Avg. sparsity ratio: 0.9963		Avg. sparsity ratio: 0.9966	
T15	T6	T13	T9	T12	T5
web	retrieval	time	algorithms	networks	networks
search	information	polynomial	approximation	control	wireless
information	search	algorithm	problems	performance	sensor
semantic	query	linear	trees	analysis	mobile
content	models	computing	parallel	traffic	scheduling
user	queries	algorithms	algorithm	atm	distributed
mining	relevance	systems	preliminary	packet	routing
knowledge	evaluation	computation	problem	network	multihop

many topics, indicating a broad, and even ambiguous usage of those terms.

Table 8: Most frequent terms of a category, and most relevant topics on DBLP

DB/DM/IR		TCS					Networking/System		
web	mining	algorithm			complexity		network		wireless
6	7	1	4	11	1	13	2	8	5
8	11	2	5	12	9		5	9	
15	15	3	9	13	10		7	12	

Table 9: Most frequent terms of a category, and most relevant topics on 20NG

comp.graphics					comp.os.ms-windows.misc					
image		graphics		jpeg	windows		system			
21	82	21	56	56	21	56	2	28	47	65	85
37		35	60		34	61	9	30	54	67	87
50		37	61		37	76	16	31	56	69	90
56		48	71		47		17	34	60	70	97
58		50	76		50		21	37	61	71	99
71		52	99		52		27	41	63	81	100

One can also observe that there are some topics that select all the frequent terms in a category as focused terms (underlined in Tables 8 and 9, e.g., topic 15 for DB/DM/IR, topic 9 and 13 for TCS, and topic 56 for "comp.graphics"). These topics are likely to be highly related to that category. Interestingly, these topics also match well with the most common focused topics of that category (see Tables 4 and 5). This correlation provides evidence of the existence of the joint effect of dual-sparsity.

Summary: We have presented comprehensive experiments using three different genres of text collections, documents with various lengths and concentrations of topics, and multiple tasks to evaluate the effectiveness of the proposed model DsparseTM. The proposed model successfully discovers focused topics, focused terms, and dual-sparsity from data. The model results in an effective sparse topic representation of documents and a coherent word representation of topics. The performance outperforms classical topic models (i.e., LDA) and the state-of-the-art sparsity-enhanced topic models (i.e., STC). The effectiveness is especially significant on short text.

6. CONCLUSION

In this paper, we address the dual sparsity of the topic representation for documents and the word representation for topics in topic modeling. This problem is especially important for analyzing short text such as user-generated content on the Web.

We propose a novel topic model, DsparseTM, which employs a "Spike and Slab" process and introduces a smoothing prior and a weak smoothing prior for focused/unfocused topics and focused/unfocused terms. DsparseTM can effec-

Table 5: Focused topics and average sparsity ratio for different 20NG categories

comp.graphics			comp.os.ms-windows.misc			soc.religion.christian			talk.politics.mideast		
Avg. sparsity ratio: 0.9624			Avg. sparsity ratio: 0.9631			Avg. sparsity ratio: 0.9514			Avg. sparsity ratio: 0.9521		
T56	T48	T37	T34	T52	T76	T3	T58	T72	T87	T80	T68
image	points	ftp	windows	hp	writes	god	god	god	israel	israeli	armenian
color	line	pub	dos	printer	articles	jesus	church	religion	israeli	israel	turkish
jpeg	point	version	file	print	zip	christ	bible	athelists	arab	arab	armenians
gif	higgins	graphics	driver	font	ftp	christians	jesus	exist	lebanese	true	genocide
file	find	tar	system	fonts	risc	christian	father	atheism	peace	center	turks
format	writes	software	drivers	postscript	computer	law	spirit	evidence	writes	policy	people
files	polygon	based	problem	good	files	people	catholic	people	arabs	questions	armenia
images	radius	subject	files	windows	instruction	hell	holy	existence	lebanon	jews	soviet

tively model the dual sparsity of document-topic and topic-term distributions and successfully discover focused topics of a document and focused terms of a topic. The sparse topic representation provides a nice low-dimensional latent semantic representation of documents, which is useful in many applications such as text classification. Experimental results on a variety of real-world data sets demonstrate the advantage of DsparseTM over classical topic models and the state-of-the-art sparsity-enhanced topic models.

Due to its simple but effective model structure and inference procedure, DsparseTM can be integrated with additional regularities and/or stochastic online inference procedures. By doing this, we anticipate that DsparseTM can scale to handle very large collections of documents. It is also a natural future direction to optimize the hyperparameters of DsparseTM, or to investigate its nonparametric counterpart.

7. ACKNOWLEDGMENTS

This work is supported by the Hong Kong Research Grants Council (RGC) General Research Fund (GRF) Project No. CUHK 411211, 411310, the Chinese University of Hong Kong Direct Grant No. 4055015, and the National Science Foundation under grant numbers IIS-0968489, IIS-1054199, CCF-1048168.

8. REFERENCES

[1] C. Archambeau, B. Lakshminarayanan, and G. Bouchard. Latent IBP compound dirichlet allocation. In *NIPS Bayesian Nonparametrics Workshop*, 2011.

[2] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *UAI*, pages 27–34, 2009.

[3] Y. Bengio, A. C. Courville, and J. S. Bergstra. Unsupervised models of images by spike-and-slab rbms. In *ICML*, pages 1145–1152, 2011.

[4] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[5] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, pages 106–114, 2003.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[7] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296, 2009.

[8] X. Chen, M. Zhou, and L. Carin. The contextual focused topic model. In *KDD*, pages 96–104, 2012.

[9] A. C. Courville, J. Bergstra, and Y. Bengio. A spike and slab restricted boltzmann machine. In *International Conference on Artificial Intelligence and Statistics*, pages 233–241, 2011.

[10] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2002.

[11] J. V. Graca, K. Ganchev, B. Taskar, and F. Pereira. Posterior vs. parameter sparsity in latent variable models. In *NIPS*, pages 664–672, 2009.

[12] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

[13] T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, pages 289–296, 1999.

[14] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *JMLR*, 5:1457–1469, 2004.

[15] H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

[16] A. Kabán, E. Bingham, and T. Hirsimäki. Learning to read between the lines: The aspect bernoulli model. In *SDM*, pages 462–466, 2004.

[17] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[18] Y. Lu, Q. Mei, and C. Zhai. Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203, 2011.

[19] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *SIGIR*, pages 889–892, 2013.

[20] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *NAACL*, pages 100–108, 2010.

[21] I. Sato and H. Nakagawa. Rethinking collapsed variational bayes inference for lda. In *ICML*, 2012.

[22] E. Saund. A multiply cause mixture model for unsupervised learning. *Neural Comput.*, 7(1):51–71, 1995.

[23] M. Shashanka, B. Raj, and P. Smaragdis. Sparse overcomplete latent variable decomposition of counts data. In *NIPS*, pages 1313–1320, 2007.

[24] J. Tang, M. Zhang, and Q. Mei. One theme in all views: Modeling consensus topics in multiple contexts authors. In *KDD*, pages 5–13, 2013.

[25] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, pages 1353–1360, 2006.

[26] H. M. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. In *NIPS*, pages 1973–1981, 2009.

[27] C. Wang and D. M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In *NIPS*, pages 1982–1989, 2009.

[28] Q. Wang, J. Xu, H. Li, and N. Craswell. Regularized latent semantic indexing. In *SIGIR*, pages 685–694, 2011.

[29] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. Focused topic models. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.

[30] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei. The ibp compound dirichlet process and its application to focused topic modeling. In *ICML*, pages 1151–1158, 2010.

[31] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *ECIR*, pages 338–349, 2011.

[32] J. Zhu and E. P. Xing. Sparse topical coding. In *UAI*, pages 831–838, 2011.

APPENDIX

Derivations of Zero-Order Collapsed Variational Bayes Inference

Assuming $\Theta = \{s, t, \pi, \gamma, \bar{\pi}, \bar{\gamma}, x, y\}$. We begin with the joint probability $P(\vec{W}, \vec{Z}, \vec{\alpha}, \vec{\beta}; \Theta)$ by taking the advantages of conjugate priors to simplify the derivation. All symbols have been defined in Section 3:

$$
\begin{aligned}
&P(\vec{W}, \vec{Z}, \vec{\alpha}, \vec{\beta}; \Theta) \\
=&\int\int\int\int p(\vec{W}, \vec{Z}, \vec{\theta}, \vec{\phi}, \vec{\alpha}, \vec{\beta}, \vec{a}, \vec{b}; \Theta) d\vec{\theta} d\vec{\phi} d\vec{a} d\vec{b} \\
=&(\prod_{d=1}^{|D|}\prod_{w=1}^{N_d} p(w_{d,w}; \vec{\phi}_{z_{d,w}}, \vec{\beta}))(\prod_{k=1}^{K} p(\vec{\phi}_k; \gamma, \bar{\gamma}, \vec{\beta}_k))(\prod_{k=1}^{K} p(\vec{\beta}_k; b_k)) \\
&(\prod_{k=1}^{K} p(b_k; x, y))(\prod_{d=1}^{|D|}\prod_{w=1}^{N_d} p(z_{d,w}; \vec{\theta}_d, \vec{\alpha}_d))(\prod_{d=1}^{|D|} p(\vec{\theta}_d; \pi, \bar{\pi}, \vec{\alpha}_d)) \\
&(\prod_{d=1}^{|D|} p(\vec{\alpha}_d; a_d))(\prod_{d=1}^{|D|} p(a_d; s, t)) d\vec{\theta} d\vec{\phi} d\vec{a} d\vec{b} \\
=&\prod_{k=1}^{K} \frac{\Gamma(\gamma \sum_{r=1}^{|V|} \beta_{k,r} + |V|\bar{\gamma})}{\prod_{r=1}^{|V|} \Gamma(\gamma\beta_{k,r} + \bar{\gamma})} \frac{\prod_{r=1}^{|V|} \Gamma(n^k_{(.),r} \mathrm{I}[r \in B_k] + \gamma\beta_{k,r} + \bar{\gamma})}{\Gamma(\sum_{r=1}^{|V|} (n^k_{(.),r} \mathrm{I}[r \in B_k] + \gamma\beta_{k,r}) + |V|\bar{\gamma})} \\
&\frac{B(x + \sharp\{1 \le r \le |V|, \beta_{k,r} = 1\}, y + \sharp\{1 \le r \le |V|, \beta_{k,r} = 0\})}{B(x, y)} \\
&\prod_{d=1}^{|D|} \frac{\Gamma(\pi \sum_{k=1}^{K} \alpha_{d,k} + K\bar{\pi})}{\prod_{k=1}^{K} \Gamma(\pi\alpha_{d,k} + \bar{\pi})} \frac{\prod_{k=1}^{K} \Gamma(n^k_{d,(.)} \mathrm{I}[k \in A_d] + \pi\alpha_{d,k} + \bar{\pi})}{\Gamma(\sum_{k=1}^{K} (n^k_{d,(.)} \mathrm{I}[k \in A_d] + \pi\alpha_{d,k}) + K\bar{\pi})} \\
&\frac{B(s + \sharp\{1 \le k \le K, \alpha_{d,k} = 1\}, t + \sharp\{1 \le k \le K, \alpha_{d,k} = 0\})}{B(s, t)}
\end{aligned}
$$

Under the framework of collapsed bayes inference framework, the variational distribution and the variational free energy can be defined as:

$$
\begin{aligned}
&\hat{q}(\mathbf{z}, \alpha, \beta) = \prod_{ij} \hat{q}(z_{ij}|\hat{\gamma}_{ij}) \prod_{jk} \hat{q}(\alpha_{jk}|\hat{a}_{jk}) \prod_{kr} \hat{q}(\beta_{kr}|\hat{b}_{kr}) \\
&\hat{\Im}(\hat{q}(\mathbf{z}, \alpha, \beta)) \triangleq E_{\hat{q}(\mathbf{z},\alpha,\beta)}[-\log p(\mathbf{z}, \mathbf{w}, \alpha, \beta|\Theta)] - H(\hat{q}(\mathbf{z}, \alpha, \beta))
\end{aligned}
$$

Through minimizing the variational free energy respect to different variational parameters, we derive the updates for the variational parameters, $\hat{a}_{jk}$,$\hat{b}_{kr}$ and $\hat{\gamma}_{ij}$.

Variational Bernoulli distribution for $\vec{\alpha}$: Minimizing (7) with respect to $\hat{a}_{jk}$, we get

$$
\begin{aligned}
\hat{a}_{jk} &= \hat{q}(\alpha_{jk} = 1) \\
&= \frac{\exp(E_{\hat{q}(\mathbf{z},\alpha^{\neg jk},\beta)}[\log p(\alpha_{jk} = 1|\alpha^{\neg jk}, \mathbf{z}, \beta : \Theta)]}{\sum_{m=0,1} \exp(E_{\hat{q}(\mathbf{z},\alpha^{\neg jk},\beta)}[\log p(\alpha_{jk} = m|\alpha^{\neg jk}, \mathbf{z}, \beta : \Theta)]}
\end{aligned}
$$

Plugging in $P(\vec{W}, \vec{Z}, \vec{\alpha}, \vec{\beta}; \Theta)$, cancelling those appearing in both numerator and denominator, and applying the Gaussian approximation to the above equation as well as [25], we can get:

$$
\begin{aligned}
\hat{a}_{jk} =& \frac{a^1_{jk}}{a^1_{jk} + a^0_{jk}} \\
\tilde{a}^1_{jk} =& (r + A_j^{\Theta\neg jk})\Gamma(N^{\Theta}_{jk} + \pi + \bar{\pi}) \\
& B(\pi + K\bar{\pi} + \pi A_j^{\Theta\neg jk}, N^{\Theta}_j + \pi A_j^{\Theta\neg jk} + K\bar{\pi}) \\
\tilde{a}^0_{jk} =& (s + K - 1 - A_j^{\Theta\neg jk})\Gamma(\pi + \bar{\pi}) \\
& B(K\bar{\pi} + \pi A_j^{\Theta\neg jk}, N^{\Theta}_j + \pi + \pi A_j^{\Theta\neg jk} + K\bar{\pi})
\end{aligned}
$$

where $A^{\Theta}_j = \sum_{k'} \alpha_{jk'}$, $N^{\Theta}_{jk} = \sum_{i'} \gamma_{i'jk}$, $N^{\Theta}_j = \sum_{i'k'} \gamma_{i'jk'}$ and $\neg jk$ means without α_{jk}.

Variational Bernoulli distribution for $\vec{\beta}$: Similar as calculation above,the update equation for variational parameter $\hat{b}_{kr}$ is:

$$
\begin{aligned}
\hat{b}_{kr} =& \frac{b^1_{kr}}{b^1_{kr} + b^0_{kr}} \\
\tilde{b}^1_{kr} =& (x + B_k^{\Phi\neg kr})\Gamma(N^{\Phi}_{kr} + \gamma + \bar{\gamma}) \\
& B(\gamma + V\bar{\gamma} + \gamma B_k^{\Phi\neg kr}, N^{\Phi}_k + \gamma B_k^{\Phi\neg kr} + V\bar{\gamma}) \\
\tilde{b}^0_{kr} =& (y + V - 1 - B_k^{\Phi\neg kr})\Gamma(\gamma + \bar{\gamma}) \\
& B(V\bar{\gamma} + \gamma B_k^{\Phi\neg kr}, N^{\Phi}_k + \gamma + \gamma B_k^{\Phi\neg kr} + V\bar{\gamma})
\end{aligned}
$$

where $B^{\Phi}_k = \sum_{r'} \beta_{kr'}$, $N^{\Phi}_{kr} = \sum_{i'j'} \mathrm{I}[w_{i'j'} = r]\gamma_{i'j'k}$, $N^{\Phi}_k = \sum_{i'j'} \gamma_{i'j'k}$ and $\neg kr$ means without β_{kr}.

Variational Multinomial distribution for z:

The same as zero-order collapsed variational bayes inference (CVB0) for LDA [2], we can directly derive the the distribution to update γ_{ijk}:

$$
\begin{aligned}
\gamma_{ijk} &= \hat{q}(z_{ij} = k) \\
&\propto \frac{N^{\Phi\neg ij}_{kw_{ij}} + \gamma\hat{b}_{kw_{i,j}} + \bar{\gamma}}{N^{\Phi\neg ij}_k + \gamma B^{\Phi}_k + V\bar{\gamma}} (N^{\Theta\neg ij}_{jk} + \pi\hat{a}_{jk} + \bar{\pi})
\end{aligned}
$$

where all notations have been mentioned before.