JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 35, 749-767 (2019)
DOI: 10.6688/JISE.201907_35(4).0003

SRTM: A Sparse RNN-Topic Model for Discovering Bursty Topics in Big Data of Social Networks

LEI SHI, JUN-PING DU, MEI-YU LIANG AND FEI-FEI KOU
Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, 100876 P.R. China
E-mail: leikyshi@qq.com; {junpingdu; meiyu-1210; koufeifei000}@126.com

Social networks such as Twitter, Facebook, and Sina microblog have become major sources for generating big data and bursty topics. As bursty topics discovery is helpful to guide public opinion and control network rumors, it is necessary to design an effective method to detect the quickly-updated bursty topics. However, bursty topics discovery is challenging. This main reason is that big data is both high dimensional and sparse in social networks. In this study, we propose a Sparse RNN-Topic Model (SRTM) named SRTM, to deal with the task. First, we leverage RNN to learn the inside relationship between words and IDF to measuring high-frequency words. Second, the model distinguishes modeling between the bursty topic and the common topic to detect the variety of word in time. Third, we introduce "Spike and Slab" prior to decouple the sparsity and smoothness of the topic distribution. The burstiness of word pair is leveraged to achieve automatic bursty topics discovery. Finally, to verify the effectiveness of the proposed SRTM method, we collect Sina microblog dataset to conduct various experiments. Both qualitative and quantitative evaluations demonstrate that our proposed SRTM method outperforms favorably against several state-of-the-art methods.

Keywords: social networks, bursty topic discovery, topic model, RNN, big data

1. INTRODUCTION

With the development of the Internet, the social networks generate massive data every second, just take Sina microblog (the largest microblog platform in China) as an example, users will produce more than one hundred million data every day. In social networks such as Twitter, Sina microblog and Facebook, many users extensively use these platforms to discuss daily chatting, spreading bursts of world news. These platforms have many times been the first publisher of significant bursty topics, such as natural disaster and violent terrorist incidents. Thus, discovering bursty topics not only can provide people understand public attention, but also benefit many related applications such as public opinion mining, emerging topic detecting and topic clues tracking.

However, bursty topic discovery in social networks has the following challenges: (1) The contents are particularly short in social networks, and suffering the feature sparsity problem. How to alleviate sparsity problem and extract high-quality bursty topics from massive data is a much-watched challenge; (2) The topics are noisy and diverse, with a lot of spam information and misleading topics in social networks. Thus, it is necessary and challenging to distinguish bursty topics from common contents.

Received December 28, 2017; accepted September 19, 2018.
Communicated by Sun-Yuan Hsieh.

In previous studies, a typical method for bursty topic discovery is based on topic model, such as probabilistic latent semantic analysis (PLSA) [1] and Latent Dirichlet Allocation (LDA) [2], which are widely used to discover the latent topics from the normal text corpus. However, origin topic model method is used to detect the top-*N* topics in a normal text corpus or new corpus. It is not directly applying to detect bursty topic in social networks. Although many researchers leverage the post-processing method to discover bursty topics from clustering result. It is not very effective because discovered topics may be the common topic.

Meanwhile, these methods ignore the quantifiable relationship between words. To deal with these problems, many researchers have proposed temporal topic model [3, 4]. Unfortunately, these methods still require post-processing method for the discovered topic.

Furthermore, the above methods are initially designed to handle the regular texts, so it is less effective for social networks texts.

Recently, another study to detect bursty topics are based on feature clustering. However, these methods need related processing like heuristic tuning, pre-processing and post-processing, because the detected bursty feature are ambiguous and sparse, so it is difficult to cluster. Meanwhile, representing these bursty topics only via bursty features might lose a lot of important topics, making them difficult to distinguish between two similar topics.

In this paper, we propose a novel sparse RNN-topic model (SRTM) to effectively modeling short text and discovering bursty topics.

According to the actual situation of the social networks, the topic is defined as being bursty in a time step if it is shared and talked by a large number of users in a time slice. But it has little discussion at other times. The key of SRTM is to use burstiness of word pair as the prior introduction to the topic model. Meanwhile, recurrent neural networks (RNN) and the "Spike and Slab" prior are leveraged to learn the relationship of word pair and decouple the smoothness of the bursty topic distribution. To reduce the influence common words, we leverage famous inverse document frequency (IDF) to measure all words in the datasets. It can not only implement bursty topic discovery without any post-processing but also learn the relationship of word pair and address the social networks texts sparsity.

We have conducted extensive experiments over a Sina microblog dataset. The experimental results demonstrate that our proposed SRTM obtained better results than the state-of-the-art methods. The main contributions as follows:

(1) Our proposed SRTM distinguish modeling between the bursty topic and the common topic to detect the variety of words in time. The burstiness of word pair is leveraged to automatic discover bursty topics from social network data.
(2) Our proposed SRTM model can learn the quantifiable relationship between word pair from the corpus and constructing weight prior to optimize the results of topic discovery.
(3) Our SRTM introduce "Spike and Slab" prior to decouple the sparsity and smoothness of a topic distribution, which can focus on bursty topics.

2. RELATED WORK

Research on bursty topic discovery over big data of social networks has attracted great attention from many researchers in the late years because it widely uses in bursty topics detection like natural disaster, spatial-temporal information detection, election prediction, and so on. In this section, we briefly review the related work which is most related to our work including topic model methods, feature clustering, and document clustering methods.

In topic-model-based, the traditional topic model is designed to detect the topic of news events [1, 2]. These topic models are designed for modeling regular text, and fail to model short text topic in social networks. To overcome above problem Lin *et al.* [5] utilize sparse constraints for document-topic distributions to model short texts. Zuo *et al.* [6] proposed a word network model (WNTM) to enhance the semantic density of data space. Wang *et al.* [7] utilized hashtag relation information in hashtag graphs to discover word semantic relations. Yan *et al.* [8, 9] proposed a word pairs topic model, namely BTM based on mixture of unigrams, which effectively solves the sparseness problem. Zuo *et al.* [10] proposed a pseudo-document topic model for short text topic modeling Mehrotra *et al.* [11] proposed a hash pool scheme to automatically discover events. Hoffman *et al.* [12] proposed an online LDA model, which can directly analyze online data flowing. Li *et al.* [13] proposed an incremental temporal topic model namely BEE to discover bursty topics. Gao *et al.* [14] proposed a novel hierarchical Bayesian model to capture the dependency of the words, which has achieved better results in multi-document topic discovery.

In topic-model-based, bursty topic can also be detected by tackling a global optimization problem. Xie *et al.* [15] proposed TopicSketch framework to detect bursty topics, which formulates a task of bursty topic discovery as an optimization problem, and achieved better performance in efficiency and effectiveness. Huang *et al.* [16] calculate word novelty by formulating a linear regression with weight, which can detect more novelty bursty topic.

In the clustering method, the documents are usually clustered according to the topic similarity of the corpus. The typical method is incremental clustering and dictionary learning. Zhang *et al.* [17] utilized term frequency and user's social relation to discover bursty events and predicted the popularity event from social networks. Petrovic *et al.* [18] applied a local sensitive hash (LSH) method to search the neighbor for each incoming text. Fang *et al.* [19] used multi-view with semantic relations, social tag relations and temporal relations clustering to detect topic Petrovic *et al.* [20] improved the algorithm by applying fragment-level phrases and LSH. Becker *et al.* [21] proposed an incremental clustering method to detect emergency events in social networks. Other similar research is to apply dictionary learning method to discover new topics. McMinn *et al.* [22] proposed a method for detecting and tracking named entities in a bursty event. Dong *et al.* [23] proposed a novel multiscale topics detection method for social networks data, which can automatically handle the interaction of temporal-spatial attribute. However, the above methods require complex heuristic adjustments and post-processing, because the detected bursty characteristics are noisy and sparse, so it is very difficult to cluster.

Feature clustering methods try to extract features of topics from documents, and then topics are detected by clustering features based on their semantic relatedness. Michael *et al.* [24] proposed a TwitterMonitor system to perform trend detection in social networks. Weng *et al.* [25] used wavelet analysis method to filters the unrelated words and then clustered the remaining words for topics detection. Li *et al.* [26] proposed a

Twevent method to discover bursty event. Schubert *et al.* [27] proposed a novel measure to discovery bursty topics early, and leverage term clustering approaches to detect co-trends into larger topics.

3. MODEL INTRODUCTION AND INFERENCE

Motivated by the promising potential of RNN and BTM method in dealing with data sparsity, we propose a sparse recurrent neural network (RNN)-topic model for bursty topic discovery in big data of social networks.

As discussed before, most variants of topic model like BBTM, WNTM ignore the internal relationship between words in corpus. However, the relationship is play an important role in topic model because if the two words are closely related, they have very large probability to appear on the same topic.

3.1 Prior Knowledge Learning Based on RNN

Inspired by leveraging RNN to represent short text [28, 29], we also learn the relationship between words by Elman RNN. The Elman RNN network is shown as follows:

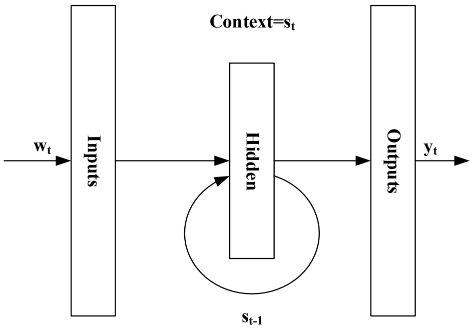


Fig. 1. An Elman RNN for text representation.

In Fig. 1, $w_t \in \mathbb{R}^T$ is the current word where T the size of vectorized $s_t \in \mathbb{R}^S$ represent a hidden unit, $y_t \in \mathbb{R}^N$ represent the output unit at time steps t. $x_t = [w_t, s_{t-1}]$ is the input layer, where $x_t \in \mathbb{R}^{T+S}$, and then hidden and output layers can be computed by x_t:

$$s_t = \delta(\mathrm{U}x_t) \tag{1}$$

$$y_t = h(\mathrm{V}s_t) \tag{2}$$

Where $\mathrm{U} \in \mathbb{R}^{S\times(T+S)}$ and $\mathrm{V} \in \mathbb{R}^{N\times S}$ are parameter matrices and vector. $\delta(\cdot)$ is the sigmoid function:

$$\delta(z) = \frac{1}{1+e^{-z}}. \tag{3}$$

$h(\cdot)$ is the softmax function:

$$h(z_m) = \frac{e^{z_m}}{\sum_k e^{z_k}}. \quad (4)$$

For the output results, we define y_i represent the relationship between word pair $w_{i,1}$ and $w_{i,2}$

$$y_i(j) = P(w_{i,2} \mid w_{i,1}, s_{i-1}). \quad (5)$$

Where $y_i(j)$ is the probability of $w_{i,1}$ appears given that $w_{i,2}$ has appeared. Since the hidden layer S_t and S_{t-1} can save necessary past all words. Thus, we can leverage RNN to learn the relationship between past all words and the current word.

To filter high-frequency words, such as [30] only deletes some common word pairs. Since sparsity and imbalance of short texts, deleting word pair may aggravate the problem so we apply inverse document frequency (IDF) to measure each word:

$$\text{IDF}_{w_i} = log \frac{|N_D|}{|d \in D : w_i \in d|}. \quad (6)$$

Where $|d \in D$: $w_i \in d|$ is number of documents where the term w_i appears. From Eq. (6) we can see the higher the number w_i occurrences in datasets, the smaller value of IDF. To reduce w_i's probability, we leverage this weight to achieve the step. Based on the above analysis, we can define prior knowledge β as follows:

$$\beta_i = \ell \times y_i(j) \times IDF_{w_i}, \quad (7)$$

$$\beta_j = \ell \times y_i(j) \times IDF_{w_{i,2}}. \quad (8)$$

Where ℓ is a relatively small positive number to avoid β being too small. Different from word pair definition in BBTM, we have introduced a priori knowledge in extracting word pairs. For each word pair $p \in P$, the definition is as follows: $p = (w_{i,1}, w_{i,2}, IDF_{w_{i,1}}, IDF_{w_{i,2}}, y_i(j))$. The extraction process of word pair is executing when reading the whole dataset.

3.2 Establishment of Sparse Topic Model

The "Spike and Slab" prior is a very effective established approach in Statistics and Mathematics, which is originally introduced by Wang *et al.* [31] into the topic model to implement sparse topic-word distribution. It can decouple the distribution of sparse and smooth. Especially Bernoulli variables are introduced into the prior, which determine "on" or "off" status of switch variables. Therefore, the model can judge whether a corresponding variable appears or not. In our approach, the switch variable indicates whether or not a topic is focused on the dataset. Since the "Spike and Slab" prior can produce null selection, which will lead to the probability distribution to be ill-defined. To tackle this problem, Lin *et al.* [5] proposed a weak smoothing prior to avoid the ill-defined distribution by the direct apply the Spike and Slab prior. Therefore, we also apply weak smoothing prior to avoid an ill-defined and simpler reasoning process, which can ensure the scalability of our model.

3.3 Model Formulation

Suppose the word pair P occurs n_w^t times in a time step T, since a word pair may be identified either used normally or in some bursty topic, so we decompose a word pair n_w^t into two parts: $n_{w,0}^t$ is the number of the word pair P occurred in normal usage, while $n_{w,1}^t$ is the number of the word pair P occurred in bursty topic. Where $n_{w,0}^t + n_{w,1}^t = n_w^t$, Such $n_{w,0}^t$ almost is constant over time, while $n_{w,1}^t$ may continuously change at different time steps. When some bursty topics related to word pair break out, $n_{w,1}^t$ might increase sharply, while there are no bursty topics generate in other time steps, $n_{w,1}^t$ will be nearly 0. Therefore, we can estimate $n_{w,0}^t$ by the mean value of n_w^t in the last M time steps $\bar{n}_w^t = \frac{1}{M}\sum_{M=1}^{M} n_w^{t-m}$. Then we can obtain $\hat{n}_{w,1}^t = \max[(n_w^t - \bar{n}_w^t), \tau]$ at the same time, where $n_{w,0}^t$ and $n_{w,1}^t$ cannot be observed, τ is a relatively small positive number to avoid the 0 value. We can apply the time and frequency to approximate the probabilistic of word pair generated in time steps T as follows:

$$\mu_w^t = \frac{\max\left[\left(n_w^t - \bar{n}_w^t\right), \tau\right]}{n_w^t}. \tag{9}$$

Where μ_w^t is the bursty probability of the word pair P in the time steps T. It suggests that word pair P appears more frequently than in a time step than other times step, and more likely to be generated from bursty topics. Table 1 lists the key notations of our proposed SRTM model.

Table 1. Variables and notations.

Notation	Meaning
D	collection of short documents
N_P	number of word pair
K	number of topics
P	set of word pair
ϕ_0	normal word distribution
θ	bursty topic distribution
b_z	topic selector
μ_w^t	bursty probability of word pair
z	topic assignment
α	bursty topic smoothing prior
$\bar{\alpha}$	weak topic smoothing prior
γ_0, γ_1	hyperparameter
π	binary variable
A_z	set of its focused topics
$I[\cdot]$	indicator function

Definition 1: Corpus contains two types of topics: bursty topic and common topic, the content of a bursty topic rapidly increase in the current time steps, while common topics almost are constant over time.

Definition 2: Given a short text corpus $D = \{d_1, d_2, \ldots, d_{N_d}\}$, a topic selector b_z is a binary switch variable to control whether the topic is a focused topic. b_z is sampled from the Bernoulli distribution.

Definition 3: The Smoothing Prior α is Dirichlet hyperparameter to smooth the topic is selected by the topic selector, while the weak Smoothing Prior $\bar{\alpha}$ is another Dirichlet hyperparameter to smooth the topic that is not appears in the topic. Since $\bar{\alpha} << \alpha$, the hyperparameter $\bar{\alpha}$ is called weak smoothing prior.

Definition 4: If the topic selector $b_z = 1$, the topic is a focused topic. For the dataset $A_z = \{z{:}b_z = 1, z \in \{1, \ldots, K\}\}$ is defined as the focus topic.

3.4 A Sparse Topic Model

Based on the above analysis, the word pair is generated by the topic. Therefore, the burstiness of word pair directly relates with the burstiness of the topic, we assume that a word pair is identified either normal usage or in some bursty topic. Our SRTM model applies the learned burstiness of word pair to discover bursty topics based on the above assumption. We define a binary switch variable π to represent the source of occurrence a word pair. Where $\pi = 0$ indicates word pair is generated from the normal topic, while $\pi = 1$ indicates word pair is generated from bursty topic so we apply bursty probability of a word pair to encode prior knowledge from bursty topic. Meanwhile, introduce a prior distribution with parameter μ_w^t for a binary switch variable π. Moreover, we introduce θ to denote the bursty topics distribution in the collection, ϕ_k denote the word distribution for bursty topics in the collection. A normal word distribution ϕ_c denote the normal usage. Then we apply smoothing prior and weak smoothing prior to decouple the topic distribution of sparse and smooth. In particular, given a short text data $D = \{d_1, d_2, \ldots, d_{N_d}\}$, the corresponding set of word pair is $P = \{p_1, p_2, \ldots, p_N\}$, and $p_i = (w_{i,1}, w_{i,2})$. The graphical model of our SRTM was shown in Fig. 2. The generative process for SRTM as follows:

```
1. For the collection
     sample $\pi \sim Beta(\gamma_0, \gamma_1)$
     sample topic selector $b_z \sim Bernoulli(\eta)$, $\vec{b} = \{b_z\}_{k=0}^{K}$
     sample a bursty topic distribution $\theta \sim Dir(\alpha\vec{b} + \bar{\alpha}\vec{1})$
2. For each bursty topic
     sample a word distribution: $\phi_{k,1} \sim Dir(\beta_i)$
     sample a word distribution: $\phi_{k,2} \sim Dir(\beta_j)$
     sample a normal word distribution $\phi_{c,1} \sim Dir(\beta_i)$
     sample a normal word distribution $\phi_{c,2} \sim Dir(\beta_j)$
3. For each word pair $p_i \in P$
     sample a binary switch $\pi \sim Bernoulli(\mu_w)$
     If $\pi = 0$
       sample words $w_{i,1} \sim Multi(\phi_{c,1})$
       sample words $w_{i,2} \sim Multi(\phi_{c,2})$
     If $\pi = 1$
       sample a bursty topic $z \sim Multi(\theta)$
```

sample words $w_{i,1} \sim Multi(\phi_{z,1})$
sample words $w_{i,2} \sim Multi(\phi_{z,2})$

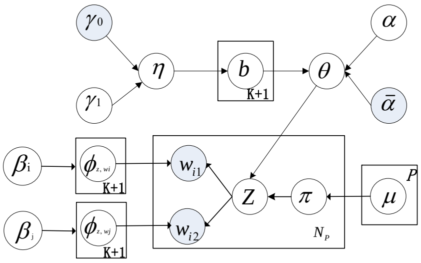


Fig. 2. The graphical model of SRTM.

3.5 Parameter Estimation

The exacting inference is often intractable in many topic models and appropriate methods must be used, such as Collapsed Gibbs sampling and variational inference. Consequently, we employ the collapsed Gibbs sampling algorithm [32] to approximate to obtain samples of latent variables and estimate unknown parameters in the SRTM, which is simple to derive, comparable in speed to other estimators, and can approximate a global maximum. The key idea is to alternately estimate random variables for posterior sampling, where each random variable is sampled based on the assignment of other random variables.

In SRTM, we sample a topic for each word pair. Integrating out θ, ϕ, and η analytically, the latent variables needed by the Gibbs sampling algorithm are switching variables π and topic selector b_z. We also sample Dirichlet hyper-parameter α and Beta hyper-parameter γ_1, and fix $\bar{\alpha}$ equal to 10^{-8} and γ_0 equal to 1. According to sampling following conditional distribution:

$$P(\pi = 0 \mid rest) \propto (1-\mu_i)\frac{\left(n_{0,w_{i,1}}^{\neg i} + \beta_i\right)\left(n_{0,w_{i,2}}^{\neg i} + \beta_j\right)}{\left(n_{0,\cdot}^{\neg i} + W\beta\right)\left(n_{0,\cdot}^{\neg i} + 1 + W\beta\right)} \tag{10}$$

$$P(\pi = 1, z_i = k \mid rest) \propto \mu_i \frac{\left(n_k^{\neg i} + b_z\alpha + \bar{\alpha}\right)}{\left(n_\cdot^{\neg i} + |A_z|\alpha + K\bar{\alpha}\right)} \frac{\left(n_{k,w_{i,1}}^{\neg i} + \beta_i\right)\left(n_{k,w_{i,2}}^{\neg i} + \beta_j\right)}{\left(n_{k,\cdot}^{\neg i} + W\beta\right)\left(n_{k,\cdot}^{\neg i} + 1 + W\beta\right)} \tag{11}$$

where $\pi = \{\pi_i\}_{i=0}^{N_P}$, $Z = \{z_i\}_{i=0}^{N_P}$, $\mu = \{\mu_i\}_{i=0}^{N_P}$, $n_{0,w}$ represents the number of tokens of word pair is assigned to the normal word distribution, $n_{0,\cdot} = \sum_{w=1}^{W} n_{0,w}$ is the total number of words assigned to the normal word distribution, n_k is the number of word pair assigned to bursty topics, $A_z = \{z: b_z = 1, z \in \{1, \ldots, K\}\}$ is the set of indices of $\vec{b}$ that is "on", $|A_z|$ is the size of A_z, $n. = \sum_{k=1}^{K} n_k$ is the total number of word pair assigned to bursty topics, α is topic smoothing prior, $\bar{\alpha}$ is weak topic smoothing prior, $n_{k,w}$ represents the number of tokens of word w is assigned to bursty topic k, $n_{k,\cdot} = \sum_{w=1}^{W} n_{k,w}$ represents the total number of words assigned to bursty topic k, and $\neg i$ represents excluding word pair.

Sampling the topic selector b_z: For sampling, we leverage π as an auxiliary variable.

Give the joint conditional distribution as follows:

$$P\left(\eta,\vec{b}_z \mid rest\right) \propto \prod_z P\left(b_z \mid \eta\right) P\left(\eta \mid \gamma_0,\gamma_1\right) \frac{I\left[B_l\right]\Gamma\left(\left|A_z\right|\alpha + K\bar{\alpha}\right)}{\Gamma\left(n_{\cdot} + \left|A_z\right|\alpha + K\bar{\alpha}\right)}. \tag{12}$$

With the joint conditional distribution, we iteratively sample b_z condition on η and eventually obtain a sample for b_z. Then we integrate out η and sample b_z using the reverse method [5], for hyper-parameter α, we apply Metropolis-Hastings with a symmetric Gaussian as proposal distribution. For concentration parameter γ_1, we apply previously developed approaches for Gamma priors [33], $I[\cdot]$ is an indicator function $B_l = \{z: n_k > 0, z \in \{1, \ldots, K\}\}$.

We randomly assign a topic to each word as the initial state. Then, we sample latent variables according to Eqs. (10)-(12) in each iteration process. After a sufficient number of iterations, we can estimate the parameters by the learned parameter mean. The distributions are obtained by:

$$\theta_k = \frac{n_k^{-i} + b_z\alpha + \bar{\alpha}}{n_{\cdot}^{-i} + \left|A_z\right|\alpha + K\bar{\alpha}}, \tag{13}$$

$$\phi_{k,w_i} = \frac{n_{k,w_{i,1}} + \beta_i}{\left(n_{k,\cdot} + W\beta\right)}, \tag{14}$$

$$\phi_{k,w_i} = \frac{n_{k,w_{i,1}} + \beta_i}{\left(n_{k,\cdot} + W\beta\right)}. \tag{15}$$

Suppose document d contains N_P word pairs. We can estimate $P(w_{d_j}|d)$ via the method of maximum likelihood:

$$P(w_{d_j} \mid d) = \frac{n_p(w_{d_j})}{N_p}. \tag{16}$$

Where $n_p(w_{dj})$ is the frequency of word pair P in document d. According to Eq. (16), we can derive the percentage of all bursty topics in document d:

$$P\left(\pi = 1 \mid d\right) = \frac{1}{N_p}\sum_{j=1}^{Np} n_p\left(w_{d_j}\right)\hat{u}_i, \tag{17}$$

$$\hat{u}_i = P\left(\pi = 1 \mid d\right) = \frac{1}{Z_i}\mu_i \sum_{k=1}^{K} \theta_k \phi_{k,w_{i,1}} \phi_{k,w_{i,2}}, \tag{18}$$

$$Z_i = \phi_{0,w_{i,1}} \phi_{0,w_{i,2}} \left(1 - \mu_i\right) + \sum_{k=1}^{K} \theta_k \phi_{k,w_{i,1}} \phi_{k,w_{i,2}} \mu_i. \tag{19}$$

Where θ and ϕ are parameters learned in our SRTM.

4. EXPERIMENT

In this Section, we report extensive experimental results on our collected Sina microblog dataset to implement the effectiveness of the propose SRTM. The experiments are conducted to demonstrate the effectiveness and efficiency of our proposed method in bursty topic discovery.

4.1 Dataset

We collect data from Sina microblog, which are the largest microblog platform in China. A total of about 2 million microblog data were collected from February 26, 2014 to March 15, 2014. Then (1) Removing duplicate documents; (2) Word segmentation and removing stop words; (3) Removing the number of occurrences less than 8; (4) Removing documents with less than 3 words.

4.2 Baseline Method

OnlineLDA: OnlineLDA [3] is a typical bursty topic discovery method based on topic learning which model text by dividing the text stream into a set of textbooks with sequential relationships in successive time slices. Specifically, the OnlineLDA calculates the Jensen-Shannon divergence for the word distribution of the corresponding topic in two time periods. If the Jensen-Shannon divergence is greater than a threshold, it is considered a bursty topic.

Twevent: Twevent [26] is the latest method of emergency detection based on feature clustering. It consists of four steps: (1) The microblog is segmented and the segmented slices are extracted as features; (2) Calculate the burstiness of features; (3) Clustering the burst characteristics; (4) Use Wikipedia to filter topics. Because our goal is to test the effectiveness of discovery of the bursty topic, we do not cut the content of microblog, then bursty words are extracted as features and clustered.

BBTM: BBTM [9] is a bursty topic discovery model based on the BTM model. it introduces binary switching variables to determine whether the topic is a bursty topic based on the burstiness of the word.

BEE: BEE [13] is an incremental temporal topic model based on PLSA. It is able to detect bursty topic from social networks dataset and model the temporal information. It uses post-processing steps incrementally to track the topic drifting of events over time. the latent semantic indices are preserved from one period to the next.

4.3 Parameter Setting

In the experiments, we set the time step to be 1 day, $\alpha = 0.1$, $\bar{\alpha} = 10^{-12}$, $\beta = 0.01$, $\gamma_0 = 0.1$ and the value of K is varied from 10 to 50. The parameter settings for the other algorithms are based on the default parameters described in their paper.

4.4 Accuracy Bursty Topics Discovered

To evaluate the accuracy of bursty topics discovery for each approach, five volunteers are invited to manually label discovered bursty topics as true or false by all of baseline methods. To ensure fair, just, therefore, before labeling, we randomly mixed all the bursty topics discovered. Meanwhile, for all bursty topic discovered, we provide information including: date, the probability of the largest 10 words, time slice information and 40 most relevant terms. External tools, such as Google, Baidu (China's most famous search engine) and Sina microblog search can be leveraged to help judgment. Criteria for identifying bursty topics: a topic is labelled true if nearly all the words discuss a topic that appears in the current step but does not appears in the previous step. In addition, if a topic contains words that come from different topic or daily communication, it will be judged "false". A bursty topic is treated as "bursty" if more than half of the volunteers label it "true". Finally, we evaluate the accuracy of bursty topic discovery by P@K for different methods. Table 2 present the results of P@K for different methods.

Table 2. Accuracy of the bursty topics discovered.

	P@10	P@20	P@30	P@40	P@50
SRTM	0.803	0.808	0.822	0.825	0.827
BBTM	0.720	0.724	0.732	0.728	0.724
Twevent	0.711	0.715	0.725	0.693	0.689
OnlineLDA	0.228	0.221	0.213	0.209	0.186
BEE	0.612	0.552	0.481	0.473	0.467

From Table 2 we can see: (1) The accuracy of the proposed SRTM model is always greater than 0.8, which is significantly and consistently outperforms the baseline methods. It indicates that our SRTM can more accurately discover the bursty topic. Compared the accuracy of all baseline methods with different settings of different bursty topic K, we also found that the proposed SRTM method is slightly less effective at $K=10$, this is mainly because the number of topics is too few, which leads to the topic is more dispersed; (2) BBTM achieves higher accuracy than the other four baseline methods, but compared to the proposed SRTM, BBTM is relatively worst This shows that our proposed method is helpful for discovering bursty topic by prior knowledge learned from RNN, leveraging "Spike and Slab" prior to decouple the sparsity and smoothness of a distribution; (3) Twevent performs better than OnlineLDA and BEE, the major reason is that Twevent only detects bursty topics by clustering bursty features, which makes the bursty topic more centralized; (4) OnlineLDA that based on common topic model always performs the worst. This is due to the common topic model failure to model burstiness of the topic, and cannot effectively distinguish between common topics and bursty topics.

4.5 Novelty of Bursty Topics Discovered

In social networks, the bursty topic is constantly changing. We introduce Novelty [9] to evaluate the sensitivity and novelty of different algorithms for discovering bursty topics. We collect the more likely word from topic Z to construct a set of keywords in each time slice, $W^{(s)}$ and $W^{(s\text{-}1)}$ is word pair set of two adjacent time steps, the Novelty of the bursty topics is defined as follows:

$$Novelty\left(Z^{(s)}\right)=\frac{\left|W^{(s)}\right|-\left|W^{(s)}\cap W^{(s-1)}\right|}{M*L}. \tag{20}$$

Where |·| is the number of elements in the sets, M is the number of words contained in each topic and L is the number of bursty topics. In our experiments, we only leverage top-10 terms of each topic to calculate the Novelty. The result is shown in Fig. 3.

Based on the results of the comparison of novelty with different settings of bursty topic number K was shown in Fig. 3. From the results, we can observe that: (1) The variety of Novelty is obvious with increasing K; (2) Our proposed SRTM always outperforms other baseline methods on novelty, especially when the K is large. This is because the proposed SRTM model is more sensitive to bursty topics by incorporating the burstiness of word pair as prior and introducing RNN to learn relationship than baseline methods; (3) Twevent obtains better performance than other baseline method when the K is small, since it detects bursty topic only by clustering the bursty word. However, the performance of Twevent decreases fast with increasing bursty topic number K. The major reason is that more noisy topics are generated with increasing in the number of bursty topics; (4) BBTM significantly outperforms Twevent. This is because the BBTM employ word pair to model bursty topic, effectively improves the handling ability on short texts and discovering topics.

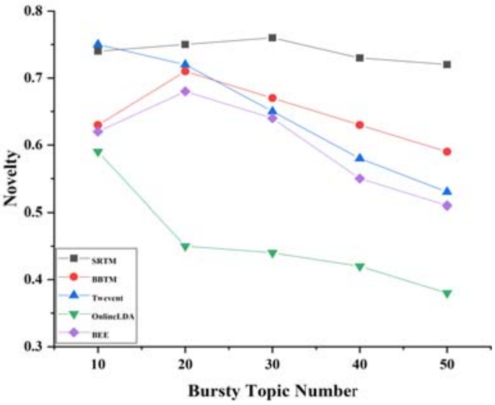


Fig. 3. Comparison of novelty results.

4.6 Topic Coherence Bursty Topics Discovered

Evaluation of the topic model has always been an open problem in academia. Perplexity is a commonly used evaluation measure, but the result has been proved less correlated to human interpretability, which better perplexity cannot understand the topic. At present, the latest topic model such as BTM, BBTM and PTM, no longer use perplexity to evaluate the model.

Based on the above analysis, we apply PMI-Score topic coherence to evaluate our model [34]. The PMI-Score uses point mutual information to evaluate the topic coherence. Given topic z, we choose the top-N possible words, w_1, w_2, …, and calculate the PMI scores for each word pair. The PMI-Score uses a large amount of external data to compute the average PMI. The higher the PMI, the more relevant the words are. Therefore, if the higher the PMI-score of a topic, the better the expandability of the topic, the formula is as follows:

$$PMI(z)=\frac{2}{N(N-1)}\sum_{1\le i\le j\le N}\log\frac{p(w_i,w_j)}{p(w_i)p(w_j)}. \tag{21}$$

Where $p(w_i, w_j)$ is the joint probability distribution of word pair w_i and w_j co-occurring the same sliding window, $p(w_i)$ is the marginal probability of word w_i appears in the sliding window within the edge probability distribution. We estimate the value of the relevant probability by Wikipedia. In our experiment, the value of N is set to10.

We calculate the average PMI of the top-10 words by using Chinese Wikipedia articles as an auxiliary corpus. Fig. 4 is the results of coherence with K varying from 10 to 50.

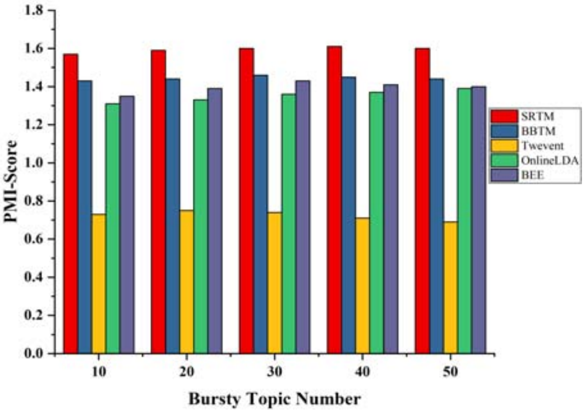


Fig. 4. Coherence of the bursty topics discovered.

From results in Fig. 4, we can make the following conclusions: (1) Our proposed SRTM consistently outperforms other state-of-the-art methods, and indicate can learn higher coherence from social networks. The major reason is that our SRTM leveraged RNN and the "Spike and Slab" prior can learn more focus bursty topics; (2) BBTM consistently outperforms OnlineLDA and Twevent, because it can learn more bursry topic by utilizing burstiness of the word to detect bursty topic; (3) BEE models achieve better results than OnlineLDA, due to the sparseness of short text can be solved; (4) Twevent is always performs the worst, because it ignores a lot of word co-occurrence patterns by simply clustering the burst characteristics.

To further analyze effectiveness of our proposed model, we will qualitatively analyze bursty topic discovery. We first randomly select two bursty strong and high-frequency hashtags[1]. The hashtags are "# KunMing Railway Station violent terrorist event #", which occurred in March. 1, 2014 and "# Malaysia flight missing event #" which occurred in March. 8, 2014. For each hashtag, extract the microblogs that contain these hashtags, statistical word frequency and normalization. Then, for all comparison method,

[1] In Sina microblog, the hashtags are expressed as "# #"

we select the most similar words with empirical word of the hashtag. Tables 3 and 4 list the top-10 words of the most similar topics with the hashtag, where the second line represents the hashtag corresponding topic content.

From Table 3, we can make the following observations: (1) The word in SRTM is most similar to the word distribution corresponding to the hashtag; (2) BBTM is also closer to the topic hashtag word distribution; (3) Twevent contains some irrelevant words. It indicates that bursty word clustering is more sensitive to noise; (4) The topics discovered by OnlineLDA contain many common words, and only part of words is related to "# Kunming Railway Station violent terrorist event #", it shows the similarity is the lowest; (5) The topics discovery of BEE is similar to OnlineLDA, where multiple different topics are mixed together. This shows that the basic topic model cannot distinguish well between the bursty topic and common topic.

Table 3. The most similar bursty topics to "#昆明火车站暴恐案 (KunMing Railway Station violent terrorist event) #" on March 1, 2014.

Empirical	SRTM	BBTM	TWevent	OnlineLDA	BEE
昆明 (kunminng)	火车站 (railway station)	嫌疑人 (suspect)	暴力 (violence)	暴力 (violence)	火车站 (railway station)
火车站(railway station)	昆明 (kunminng)	火车站(railway station)	昆明 (kunming)	危险 (danger)	袭击 (attack)
暴力 (violence)	遇难 (victims)	救治 (treatment)	砍人 (killing)	昆明 (kunming)	进站口 (Entrance)
恐怖 (terror)	暴力 (violence)	警察 (police)	袭击 (attack)	情况 (situation)	手机 (mobile phone)
袭击 (attack)	嫌疑人 (suspect)	嫌疑犯 (suspect)	进站口 (Entrance)	救护车（ambulance）	乘客 (passenger)
遇难 (victims)	打击 (combat)	新疆 (xinjiang)	恐怖 (terror)	乘务员 (attendant)	旅游 (tourism)
现场 (scene)	死亡 (death)	遇难 (victims)	购物 (shopping)	警察 (police)	现场 (scene)
嫌疑人 (suspect)	救治 (treatment)	祈祷 (pray)	美食 (delicious food)	百货大楼 (department store)	景点(tourist attractions)
打击 (combat)	紧急 (emergency)	亲人 (relatives)	祈祷 (pray)	新疆 (xinjiang)	祈祷 (pray)
救治 (treatment)	砍人 (killing)	进站 (Entrance)	云南 (yunnan)	晚点 (late)	事件 (event)

Table 4. The most similar bursty topics to "#马来西亚航班失踪事件 (Malaysia flight missing event)#" on March 8, 2014.

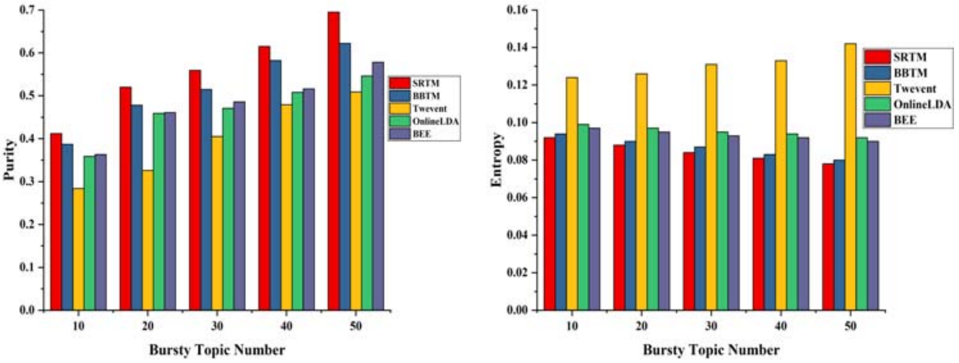
Empirical	SRTM	BBTM	TWevent	OnlineLDA	BEE
马航(Malaysia Airlines)	飞机 (aircraft)	客机 (airliner)	马来西亚 (Malaysia)	北京 (beijing)	祈祷 (pray)
飞机 (aircraft)	乘客 (Passenger)	击落 (shot down)	乌克兰 (Ukraine)	入境处(immigration department)	马航(Malaysia Airlines)
失联 (missing)	马航(Malaysia Airlines)	飞机 (aircraft)	恐怖 (terror)	乘务员(Flight attendant)	安息 (rest)
MH370	失联 (missing)	坠毁 (crash)	贵宾厅 (VIP hall)	MH370	手机 (Mobile phone)
声明 (statement)	MH370	马航(Malaysia Airlines)	航班 (flight)	护照 (passport)	天气 (weather)
遇难 (victims)	遇难 (victims)	服务 (service)	天气 (weather)	消息 (message)	旅游 (tourism)

乘客 (Passenger)	客机 (airliner)	俄罗斯 (Russia)	公司 (company)	日本 (Japan)	华为 (huawei)
祈福 (bless)	平安 (Safety)	乘客 (passenger)	护照 (passport)	马航(Malaysia Airlines)	北京 (beijing)
平安 (Safety)	声明 (statement)	中国 (China)	艾滋病 (AIDS)	报道 (report)	飞机 (aircraft)
中国 (China)	祈祷 (pray)	平安 (safety)	绝望 (despair)	事件 (event)	贵宾厅 (VIP hall)

4.7 Quality of Bursty Topic Discovered

We leverage the purity and entropy to evaluate the quality of bursty topic discovery. The purity and entropy are two standard evaluation measure for clustering quality. For OnlineLDA, BTM, BEE and SRTM. We set each bursty topic as the cluster, and then assign each document to the cluster of $P(\pi == 1|d)$. For Twevent, we assign the topics to the most similar cluster by the Jaccard coefficient between the cluster and topics.

In the experiment, we first manually select the hashtags that the daily occurrences are more than twice the average daily occurrences the first 2-15 days in microblog. Then, we sort them by the count of occurrences, and select top5 high frequency and clear hashtags as the category label for the message in the test set. We randomly sampled the 1/10 datasets to remove the hashtag as a test set. The results for different models are shown in Fig. 5.



(a) (b)

Fig. 5. comparison of cluster purity and entropy (a) Purity; (b) Entropy.

From the results of the above experiments, we observe that SRTM always outperforms existing other baseline methods both on purity and entropy. It indicates the most accurate analysis for bursty topics from social networks. The major reason is that it uses RNN to learned topic representation and IDF for filtering high-frequency words leads to better result in a text clustering task. BBTM also works better than BEE, Twevent and OnlineLDA, but it performs poorer than our SRTM. BEE also achieve better results than Twevent and OnlineLDA. This is because BEE can model the temporal information to depict burst characteristics well for analysis. Twevent always performs the worst. The major reason is that Twevent utilizes clustering the bursty words to express bursty topic which is difficult to exactly judge the similarity between the bursty topic and the message.

5. CONCLUSIONS

In this paper, we propose a novel sparse RNN-topic model (SRTM) to discover bursty topic in big data of social networks. Which can also effective alleviate sparsity problem in social networks.

Firstly, we exploit RNN to learn the semantic relationship between words and IDF for filtering high-frequency words. Secondly, SRTM introduces "Spike and Slab" priors to decouple the sparsity and smoothness of distribution and introduces the burstiness of word pair as prior knowledge to guide bursty topic modeling. Finally, SRTM utilizes the frequency of words as a prior to guide the discovery of bursty topic. Our approach can not only overcome the data sparsity of short texts in social networks, but also can effectively discover bursty topic.

Extensive experiments one real-world dataset demonstrates that our SRTM significantly outperforms all other baseline methods. However, social networks also include social contact relations, and our SRTM cannot model the social contact relations property of social topic. In our future work, we will focus on introducing the social contact relations to achieve the discovery of bursty topic based on multi-attribute topic models.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 61532006, No. 61320106006, No. 61772083).

REFERENCES

1. T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50-57.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, 2003, pp. 993-1022.
3. J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models: twitter trends detection topic model online," in *Proceedings of International Conference on Computational Linguistics*, 2012, pp. 1519-1534.
4. G. Stilo and P. Velardi, "Efficient temporal mining of micro-blog texts and its application to event discovery," *Data Mining and Knowledge Discovery*, Vol. 30, 2016, pp. 372-402.
5. T. Lin, W. Tian, and Q. Mei, "The dual-sparse topic model: mining focused topics and focused terms in short text," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 539-550.
6. Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: a simple but general solution for short and imbalanced texts," *Knowledge and Information Systems*, Vol. 48,

2016, pp. 379-398.

7. Y. Wang, J. Liu, and Y. Huang, "Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, 2016, pp. 1919-1933.
8. X. Cheng, X. Yan, and Y. Lan, "BTM: Topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, 2014, pp. 2928-2941.
9. X. Yan, J. Guo, and Y. Lan, "A probabilistic model for bursty topic discovery in microblogs," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015, pp. 353-359.
10. Y. Zuo, J. Wu, and H. Zhang, "Topic modeling of short texts: A pseudo-document view," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 2105-2114.
11. R. Mehrotra, S. Sanner, and W. Buntine, "Improving lda topic models for microblogs via tweet pooling and automatic labeling," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2013, pp. 889-892.
12. M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 2010, pp. 856-864.
13. J. Li, J. Wen, and Z. Tai, "Bursty event detection from microblog: a distributed and incremental approach," *Concurrency and Computation: Practice and Experience*, Vol. 28, 2016, pp. 3115-3130.
14. G. Yang, D. Wen, and N. S. Chen, "A novel contextual topic model for multi-document summarization," *Expert Systems with Applications*, Vol. 42, 2015, pp. 1340-1352.
15. W. Xie, F. Zhu, and J. Jiang, "Topicsketch: Real-time bursty topic detection from twitter," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, 2016, pp. 2216-2229.
16. J. Huang, M. Peng, and H. Wang, "a probabilistic method for emerging topic tracking in Microblog stream," *World Wide Web*, Vol. 20, 2017, pp. 325-350.
17. X. Zhang and X. Chen, "Event detection and popularity prediction in microblogging," *Neurocomputing*, Vol. 149, 2015, pp. 1469-1480.
18. S. Petrovic, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Proceedings of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2010, pp. 181-189.
19. Y. Fang and H. Zhang, "Detecting hot topics from twitter: A multiview approach," *Journal of Information Science*, Vol. 40, 2014, pp. 578-593.
20. S. Petrovic, M. Osborne, and V. Lavrenko, "Using paraphrases for improving first story detection in news and twitter," in *Proceedings of Conference of North American Chapter of Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 338-346.
21. H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics: Real-world event identification on twitter," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011, pp. 438-441.
22. A. J. Mcminn and J. M. Jose, "Real-time entity-based event detection for twitter," in

Proceedings of International Conference of the Cross-Language Evaluation Forum for European Languages, 2015, pp. 65-77.
23. X. Dong and D. Mavroeidis, "Multiscale event detection in social media," *Data Mining and Knowledge Discovery*, Vol. 29, 2015, pp. 1374-1405.
24. M. Mathioudakis and N. Koudas, "TwitterMonitor: trend detection over the twitter stream," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2010, pp. 1155-1158.
25. J. Weng and B. S. Lee, "Event detection in twitter," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011, pp. 311-312.
26. C. Li, A. Sun, and A. Datta, "Twevent: segment-based event detection from tweets," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 155-164.
27. E. Schubert, M. Weiler, and H. P. Kriegel, "Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014, pp. 871-880.
28. H. Amiri, "Short text representation for detecting churn in microblogs," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 2566-2572.
29. H. Lu, L. Y. Xie, and N. Kang, "Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery," in *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2017, pp. 1192-1198.
30. Y. Xia, N. Tang, A. Hussain, and E. Cambria, "Discriminative bi-term topic model for headline-based social news clustering," in *Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference*, 2015, pp. 311-316.
31. C. Wang and D. M. Blei, "Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process," in *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, 2009, pp. 1982-1989.
32. T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences*, Vol. 101, 2004, pp. 5228-5235.
33. Y. W. Teh, M. I. Jordan, and M. J. Beal, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Proceedings of Advances in Neural Information Processing Systems*, 2005, pp. 1385-1392.
34. D. Mimno, H. M. Wallach, and E. Talley, "Optimizing semantic coherence in topic models," in *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 262-272.

Lei Shi (石磊) was born in 1986. He received the M.S. degree in Control Engineering from Inner Mongolia University of Science and Technology. He is now a Ph.D. candidate in Computer Science and Technology of Beijing University of Posts and Telecommunications. His research interests include social network search, data mining and cross-media search

Jun-Ping Du (杜军平) was born in 1963. She is now a Professor and Ph.D. tutor at the School of Computer Science and Technology, Beijing University of Posts and Telecommunications. Her research interests include artificial intelligence, image processing and pattern recognition.

Mei-Yu Liang (梁美玉) was born in 1985. She received her Ph.D. degree in School of Computer Science from Beijing University of Posts and Telecommunications, Beijing, China, in 2014. She ever did postdoctoral research in School of Computer Science from Beijing University of Posts and Telecommunications from 2014 to 2016. She is currently an Associate Professor in School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include image and video processing, data mining and computer vision.

Fei-Fei Kou (寇菲菲) was born in 1989. She received her M.S. degree in Computer Technology from Beijing Technology and Business University. She is now a Ph.D. candidate in Computer Science and Technology of Beijing University of Posts and Telecommunications. Her research interests include social network search, semantic analysis and semantic learning.