A Self-adaptive Sliding Window based Topic Model for Non-uniform Texts

Jin He
School of Computer and Information
HeFei University of Technology
Hefei, China, 230009
Email: jinhe@mail.hfut.edu.cn

Lei Li
School of Computer and Information
HeFei University of Technology
Hefei, China, 230009
Email: lilei@hfut.edu.cn

Xindong Wu
School of Computing and Informatics
University of Louisiana at Lafayette
Lafayette, Louisiana 70504, the USA
Email: xwu@louisiana.edu

***Abstract*—The contents generated from different data sources are usually non-uniform, such as long texts produced by news websites and short texts produced by social media. Uncovering topics over large-scale non-uniform texts becomes an important task for analyzing network data. However, the existing methods may fail to recognize the difference between long texts and short texts. To address this problem, we propose a novel topic modeling method for non-uniform text topic modeling referred to as self-adaptive sliding window based topic model (SSWTM). Specifically, in all kinds of texts, relevant words have a closer distance to each other than irrelevant words. Based on this assumption, SSWTM extracts relevant words by using a self-adaptive sliding window and models on the whole corpus. The self-adaptive sliding window can filter noisy information and change the size of window according to different text contents. Experimental results on short texts from Twitter and long texts from Chinese news articles demonstrate that our method can discover more coherent topics for non-uniform texts compared with state-of-the-art methods.**

I. Introduction

Due to its various application prospects, topic modeling has become a hotspot for research and has been widely applied to social networks analysis [1] [2], public opinion monitoring [3], information navigation [4], bursty events detection [5] and so on. With the rapid growth of social networks, social network services such as microblogging and news recommendation have produced a lot of unstructured data, especially non-uniform texts, which include both long and short texts. Therefore, discovering latent topics from non-uniform texts becomes a challenging task [6].

Non-uniform texts are prevalent on the Internet environment. The two most common types of texts are long texts like news articles and short texts like tweets. Given a set of non-uniform texts, the goal of topic modeling is to discover semantically coherent words known as topics, which can be further used to represent and summarize the content of the whole corpus [7]. Unlike most traditional texts, different lengths of texts in non-uniform corpus raise challenges for topic modeling tasks.

For topic modeling, the well-known topic models include probabilistic latent semantic analysis (PLSA) [8] and latent Dirichlet allocation (LDA) [9]. These two topic models learn document-level topics and model documents as a multinomial distribution over topics instead of words by using probabilistic methods. Each word has a randomly chosen probability according to the parameters under every topic in LDA or PLSA. Conventional topic models assume that documents are bags-of-words (BOW), which is a simplified assumption used in natural language processing and information retrieval. BOW assumption has achieved good results for analysis of text topic detection based on machine learning methods [10]. However, this assumption ignores any non-uniform structure of corpus, such as the density of topic-specific words. In short texts, because of the sparsity in the corpus, topic-specific words are closer, while in long texts, related words are farther. These conventional methods are difficult to adapt to flexible topic specific word density.

Two major strategies have been adopted to deal with how to detecting latent topics from non-uniform texts. One strategy changes analysis granularity manually before model iteration and parameter derivation. The granularity of text information analysis can be divided into a document level, sentence level, or element level. For example, Fang et al. [11] detected the document-level topic words by using weakly supervised methods over the whole corpus, and Chen et al. [12] adopted the LDA allocation for analyzing additional information about sentence boundaries in a document. Rao et al. [13] constructed a topic-level dictionary to build a word-level topic dictionary. However, these corpus modeling methods cannot overcome the weakness of sparsity within short texts and redundant information within long texts. Moreover, these methods are highly dependent on the texts, which is not effective for topic detection from non-uniform texts. The different density of topic-specific words should be recognized for non-uniform texts, such as tweets, news articles, and image captions. Therefore, extracting topics from non-uniform texts remains a challenging task for existing methods [14].

Another strategy of solving adaptively extracted topics from non-uniform texts is to aggregate additional information with self-adaptive topic extraction. One typical source of additional information is author information when analyzing research paper corpus. Rosen et al. [15] extended topic models to include authorship information, which focuses on discovering the semantical content in the corpus. Duan et al. [7] incorporated link-based importance of documents into topic modeling. Yang et al. [16] treated long texts as the auxiliary

text of short texts such as customer reviews within product brief introduction. In a sense, these methods successfully help capture the latent topics, but rely heavily on data and cannot alleviate the nonuniformity efficiently.

Unlike these approaches, in this paper, we propose an effective topic model that discovers latent topics from long and short texts by leveraging an adaptive sliding window mechanism. This method, referred to as SSWTM, is based on a simple assumption that semantically coherent words have a highly co-occurrence frequency within a certain semantical distance. Gaining insights from [17], SSWTM models the word co-occurrence to discover groups of correlated words for topic modeling explicitly. We leverage a self-adaptive sliding window to solve the problem of different topic-specific words densities. For a long text, the extraction size is relatively larger than that for a short text. Through this way, SSWTM can effectively avoid nonuniformity. Due to the self-adaptive sliding window, SSWTM can filter noise information, which not only improves the quality of topics but also saves calculation time. Therefore, our approach can learn topics over non-uniform texts effectively and efficiently.

We conduct experiments on real-world non-uniform text collections, including two normal texts, i.e. news articles and a Twitter dataset which consists of short texts, to evaluate our approach. A comparison among LDA, BTM, and SSWTM validates that our method can learn more high-quality topics on non-uniform texts.

The remainder of the paper is organized as follows. Section II reviews related work. Section III presents our correlative word pair extraction with self-adaptive windowing and the advantages of method. Experimental analysis for our model is provided in Section IV. We conclude the paper in Section V.

II. Related Work

In this section, we briefly discuss the related work from three aspects: topic detection and modeling, topic models with uniform texts, and topic models with non-uniform texts.

A. Topic detection and modeling

Due to its solid theoretical foundation, topic modeling has been widely studied to detect latent topic structures from long and short texts for much further research, such as social community detection [18], news summarization [19], sentiment classification [20], recommendation systems [21], user-based collaborative filtering [22] and so on. The main task of topic detection and modeling is to discover latent topic structures efficiently and show the topic words in a reasonable form.

Topic detection and modeling methods can be categorized into the following 3 types: (1)Topic models using matrix factorization. These topic models defined the topic modeling task as factorizing a text data matrix. The matrix represents the document-topic relationship. Matrix factorization methods include Latent Semantic Indexing (LSI) [23], which factorizes the matrix into the multiplication of three matrices, and non-negative matrix factorization (NMF) [24]. This type of topic models explain topics from the perspective of probability, but unable to recognize topics from samples that outside training set. (2)Probabilistic topic modeling. Probabilistic topic models consider each document as a mixture of a set of topics, where each of these topics has a distribution over terms [9]. The challenge of this type of topic models is to infer the hidden document-topic and topic-term distributions. The typical probabilistic topic models are PLSA and LDA. These two methods infer the hidden distributions with two different ways. LDA gives a prior Dirichlet distribution for each document-topic distribution, while PLSA assigns a topic distribution separately for each document [25]. (3)Topic modeling based on clustering. This type of topic models presents a simple hypothesis that documents which fall in one cluster contain the same topic. Huang et al. [26] adopted a single-pass clustering method with the Vector Space Model (VSM). Zhang et al. [27] proposed a text clustering based topic model by using a word vector to enrich the feature information of the texts. In a sense, topic models based on clustering can handle the ssparsity problem within short texts. However, this type of models are poorly interpreted and require a higher computational complexity.

B. Topic models with uniform texts

Based on the above description, probabilistic topic models have better interpretability and expansibility. The most common type of texts for probabilistic topic models is uniform text, such as news articles and scientific literature. The length of these texts not dramatically changed, and the topic distribution in the texts is also uniform. Uniform texts are usually long and are generally presented as a set of documents with rich semantic information for topic detection and modeling. Many different methods have been proposed and evaluated by various evaluations to learn latent topic structures for uniform texts. Especially, LDA has been widely used in various topic detection tasks due to its better generalization ability and extensibility [28]. Compared to the unigram model [29], which is based on the assumption that each document is generated by one topic, probabilistic topic models like LDA are less limited to model a huge collection of normal texts because of the more complex assumption which is adopted by probabilistic topic models that each text is generated from multiple topics [30].

In recent years, topic models with uniform texts have been extended with many complicated variants and extensions on the standard LDA model from sdifferent points of views. For instance, for the fake of adaptivity of the number of topics while dealing with uniform texts, which leads to serious limitation - -hidden variables cannot inform hidden structuresa explicitly. Yee Whye et al. [31] solved this inherent problem by using a Bayesian nonparametric method, which sets priors on the infinite dimensional space of probabilistic distributions. Moody et al. [32] proposed a Dirichlet topic model that learns dense word vectors based on LDA modeling to improve the interpretability at the document level. Although the existing topic models with uniform texts can capture the topical relationship over text corpus, all these models only deal with normal long

texts and perform poorly in terms of time performance when the dictionary is too large.

C. Topic models with non-uniform texts

Along with extensive studies on social networks, social-related texts with non-uniform characteristics are being used in various situations such as long texts like news articles and short texts like the review comments both at the same websites. Compared to conventional long texts, the word co-occurring information in short texts is sparse because of the shorter text length. To overcome this inherent lack and improve the effectiveness of topic models, many short text topic models have been proposed. One idea is to apply topic models with uniform texts directly to short texts [1]. Because of the limited text information in a single short text, this method is not effective. The most common method is to aggregate pseudo-long text for short text before inference. For example, Yuan et al. [33] proposed a pseudo-document based topic model that aggregates related pseudo-documents implicitly for short texts against data sparsity.

Meanwhile, whether it is direct application or pseudo-long text aggregation, these approaches were not designed for the situation that long and short texts are mixed. Qiang et al. [34] proposed a topic model for heterogeneous texts to alleviate this problem with an assumption that each long text deals with multiple topics and each short text only contains one topic. But this assumption is too simple to capture the word co-occurrence information. Especially, the auxiliary information like pseudo-documents that [34] leverages is only effective when the pseudo-long texts are closely related to the short texts. This requirement is not always met.

In this paper, we focus on topic learning with non-uniform texts that has not been explored before. Especially, we are inspired by the Biterm Topic Model (BTM) [17] which models over the whole corpus rather than individed documents separately. Motivated by BTM, it is reasonable to alleviate the sparsity problem of short texts through a self-adaptive sliding window on the whole corpus based on word co-occurring information within the collection of non-uniform texts. Our experimental results will demonstrate that our method can perform better than other baseline methods with non-uniform texts.

III. Self-adaptive sliding window for non-uniform texts

The preceding analysis shows that the existing topic models can not cope with the nonuniformity with the scenario that long and short texts appear at the same time because of the lack of versatility and applicability. For example, uniform text topic models, such as PLSA and LDA, consider a document to be composed of multiple topics and each topic consists of multiple related words. On this basis, various kinds of extensions have been proposed by incorporating external information like authorship [35], and purchase information [36]. For short text topic models, whether directly applied to conventional normal texts [1] or aggregated pseudo-long texts [33], in dealing with long texts, the topics are not comprehensive.

All of the previously mentioned topic models can only deal with a single type of texts. In the scenario of non-uniform texts, our SSWTM below can model their topical relationship effectively by using a self-adaptive sliding window. In this section, we describe the main idea of our SSWTM approach by presenting the self-adaptive sliding window, the topic modeling process, and details of the algorithm.

A. Self-adaptive word pair extraction

Before presenting the details of our method, we firstly show the main idea of the self-adaptive word pair extraction process over non-uniform texts.

Inspired by BTM, which alleviates the sparsity problem of short texts by learning the topics at the corpus level [17], we adopt the definition of disordered word pairs, which keep some co-occurrence information. From the semantic point of view, a disordered word pair with a certain word co-occurrence frequency has a better similarity and relevance to show the topical relationship. Although the co-occurring patterns of word pairs perform better on short texts, the performance on modeling long texts is lacking due to the much more sufficient content and noise information. Therefore, using a fixed sliding window size to extract related word pairs will be overly dependent on local co-occurrence information which reduces the quality of topics especially in the case of long texts. Therefore, our work is based on the assumption that the more similar words are the closer the semantic distance to each other in non-uniform texts. We adopt a self-adaptive sliding window, which has a dynamic size of the window to extract more related words and filter some noise.

The “word pair” in this paper denotes an unordered word co-occurrence pattern in non-uniform texts. According to the text length and density in the current time slice, the size of the sliding window, that is, the granularity of the unordered word pairs is given. This process alleviates the nonuniformity in different text lengths and reduces the complexity of model computation. For example, when a long text has a larger proportion in the non-uniform corpus with a smaller sliding window size, we can not get multiple topics due to the much more local information. The topics will be repeatable and have much more limitation to find the global knowledge due to the longer distance among related words in long texts. Therefore, using the self-adaptive extraction mechanism will promote model flexibility.

B. Local topic modeling and global topic modeling

In order to preserve the temporal characteristics with different topic densities, we need to analyze the importance of the topic modeling process. The topic modeling process could be regarded as a choice between local topic modeling and global topic modeling.

Local topic modeling learns topics from a document collection divided by timestamps firstly, then obtains topic words with time characteristics, while global topic modeling learns

topics over the whole document collection firstly, and then divides topic words based on the time characteristics of documents. [37] showed that local topic modeling can better discover the emergence of new topics and the extinction of old topics, and can find the relationship among topics. On the contrary, the global topic modeling focuses on the accuracy of topics obtained.

We adopt local topic modeling because our corpus is about real events with time characteristics. We cut the corpus according to timestamps in advance, which is a preprocessing for our model.

C. Model description

Different from the majority of the generative topic models which simulate the process of document generation, SSWTM is based on the entire relevant word co-occurring patterns on non-uniform texts. Through this document level modeling with word pairs, SSWTM improves the way of discovering relevant word pairs, enhances the accuracy of topics, and retains time characteristics. In this subsection, we present the details of the model, including preprocessing, local topic modeling self-adaptive word pairs extraction.

Given a non-uniform text collection, as described in Section III-B, the collection divided by time characteristics contains the length information for each document, which will be used to produce the size of a self-adaptive sliding window. Then we obtain W unique words over the whole corpus by one scanning process with the self-adaptive sliding window that is controlled by a hyperparameter γ.

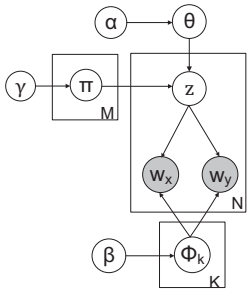


Fig. 1. The generative process of SSWTM. Each node in the graph denotes a random variable. A plate denotes an iteration of the model within it. The bottom right corner of a plate shows the number of iterative processes.

After this preprocessing, we get a corpus with M documents, expressed as $D = \{d_1, d_2, ..., d_M\}$. Suppose D contains N word pairs, referred as $B = \{b_1, b_2, ..., b_N\}$. The graphical representation of SSWTM is shown in Fig. 1. Let $z \in [1, K]$ be a topic indicator variable. Then the topics could be represented as a K-dimension multinomial distribution $\Theta = \{\theta_1, \theta_2, ..., \theta_K\}$ over W unique words, referred as $P(z)$. $\theta_z = P(z = k)$ expresses the k-th topic probabilistic distribution over W unique words with $\Theta = \{\theta_k\}_{k=1}^{K}$. We let the topic-word probabilistic distribution be $P(w|z)$ with a $K \times W$ matrix Φ where the k-th row represents the k-th topic probabilistic distribution over W unique words with $\phi_{k,w}$, $P(w|z = k)$ and $\sum_{w=1}^{W} \phi_{k,w} = 1$. Moreover, given the length of a word sequence in a document collection, referred as WN_d, the distribution of the size of a self-adaptive sliding window WL_d could be represented as $\Pi = \{\pi_1, \pi_2, ..., \pi_M\}$.

Formally, we present the generative process of SSWTM in Algorithm 1. Here, to obtain the distribution Θ, we use Dirichlet prior hyperparameters α and β, respectively, and use the hyperparameter γ as a prior for distribution Π. The first step is to perform preprocessing, on Line 1 of Algorithm 1. Steps 2 to 3 correspond to the processes of generating topic distribution Θ and topic-word distribution Π, as shown on Lines 2 to 5 of Algorithm 1. Step 4 corresponds to the process of obtaining the size of the self-adaptive sliding window, on Line 7. As stated at the beginning of this section, for each document, we use a self-adaptive sliding window to extract word pairs. After preprocessing, a document can be represented as a word sequence $d_i = \{w_1, w_2, ..., w_i, w_j, ..., w_N\}$, where w_i differs from w_j. Therefore, the size of each sliding window WL_d can be calculated with the number of the unique words in the current document, referred as WN_d. This process is described as follows:

$$WL_d = \gamma * WN_d \tag{1}$$

This self-adaptive sliding window still introduces noise words, which is a widespread problem with window-based methods. We introduce a unique word threshold Fre_w, referred as $Fre_w = n_w/W$ where n_w denotes the number of the word appearance and W denotes the number of unique words in the whole corpus. This threshold is to control the extraction of word pairs and to determine whether to adopt each word in a word pair. Step 5 draws word pairs according to the size of the window, on Line 8 of Algorithm 1. By using this threshold Fre_w for limiting the word frequency, we can reduce the time and space consumption of the model in the iteration process. Step 6 draws the topic distribution over word pairs. On Lines 10 and 13, we firstly draw a topic distribution θ_z, and then draw the topic-word distribution ϕ for each word in a word pair.

Following the above procedure, we can write the probabilistic distribution of a word pair b_i conditioned on Θ, Φ, and Π in a self-adaptive sliding window as follows:

$$\begin{aligned} P(b|\theta, \phi, \pi) &= \sum_{k=1}^{K} P(w_i, w_j, z = k|\theta, \phi, \pi) \\ &= \sum_{k=1}^{K} P(z = k|\theta_k, \pi_k) P(w_i|z = k, \phi_{k,w_i}) \cdot \\ &\quad P(w_j|z = k, \phi_{k,w_j}) \\ &= \sum_{k=1}^{K} \theta_k \phi_{k,w_i} \phi_{k,w_j} \pi_k \end{aligned} \tag{2}$$

Given the hyperparameters α, β, and γ, we can obtain the probability of b_i by integrating over Θ, Φ, and Π:

$$P(b_i|\alpha, \beta, \gamma) = \int \int \sum_{k=1}^{K} \theta_k \phi_{k,w_i} \phi_{k,w_j} \pi_k \mathrm{d}\Theta \mathrm{d}\Phi \mathrm{d}\Pi \tag{3}$$

Taking the product of the probability of a single word pair, we obtain the likelihood of the whole non-uniform text collection:

$$P(B|\alpha,\beta,\gamma)=\prod_{i=1}^{N}\int\int\sum_{k=1}^{K}\theta_k\phi_{k,w_i}\phi_{k,w_j}\pi_k \mathrm{d}\Theta\mathrm{d}\Phi\mathrm{d}\Pi \quad (4)$$

Algorithm 1 Self-adaptive sliding window based topic modeling for non-uniform texts

Input: number of topics K, α, β, γ, and document collection
Output: Θ, Φ, and Π

Divide the collection of documents by timestamps D = {d_1,d_2,...,d_M};
Draw $\theta\sim$ Dirichlet(α);
for each topic $k\in$[1, K] **do**
    draw $\phi_k \sim$ Dirichlet(β);
for each document $d\in$[1, M] **do**
    WL_d = $WN_d * \gamma$ in Eq. 1
    draw word pairs ($w_{i,1}$, $w_{i,2}$) according to WL_d;
for each word pair $b_i \in B$ **do**
    draw z_i ∼Multinomial(θ);
    draw $w_{i,1}$, $w_{i,2}$ ∼Multinomial(ϕ_z)

D. Parameter estimation

To estimate α, β, and γ, which control the topic distribution Θ, the topic-word distribution Φ, and the self-adaptive sliding window size Π respectively, we compute posterior probability $P(z|B)$, that is, the conditional probability of topics under the obtained word pairs.

In order to reduce the computational cost, we use an approximate inference method. The most common approaches are sampling and variational inference. In this paper, we adopted Gibbs sampling and Markov chain to obtain samples randomly and compute the conditional probability through all posterior values except for z_i. Therefore, the total conditional probability is represented as follows:

$$\begin{aligned}
P(z_i=k|z_{\neg i},B) &= \frac{P(z_i=k,b_i|z_{\neg i},B_{\neg i},\alpha,\beta,\gamma)}{P(b_i|z_{\neg i},B_{\neg i},\alpha,\beta,\gamma)} \\
&= \frac{P(z,B|\alpha,\beta,\gamma)}{P(z_{\neg i},B|\alpha,\beta,\gamma)} \\
&= \frac{P(z,B|\alpha,\beta,\gamma)}{P(z_{\neg i},B_{\neg i}|\alpha,\beta,\gamma)P(b_i|\alpha,\beta,\gamma)} \\
&\propto \frac{P(z_i=k,b_i|\alpha,\beta,\gamma)}{P(z_{\neg i},B_{\neg i}|\alpha,\beta,\gamma))} \\
&\propto (n_{\neg i,k}+\alpha)(n_{\neg i,w_x|k}+\beta+1)\cdot \\
&\quad \frac{(n_{\neg i,w_y|k}+\beta)\gamma}{\sum_B(n_{\neg i|k}+\beta+1)(n_{\neg i|k}+\beta)}
\end{aligned} \quad (5)$$

which $z_{\neg i}$ denotes topic distribution over all word pairs except b_i, $B\neg i$ denotes all word pairs except b_i, $n_{\neg i,k}$ denotes the number of times that all words assigned to topic *k* except word pairs b_i, $n_{\neg i,w_x|k}$ is the number of times word w_x is assigned to topic *k* except word pairs b_i, B is the number of word pairs of all documents, and W denotes the number of unique words over all documents.

By integrating each random variable in Equation 5, we can get the marginal probability distribution for $\theta_k, \phi_{k,w}$, and π_k. For $\theta_k, D = \{d_1, d_2, ..., d_M\}$, its marginal probability distribution could be obtained by integrating over $\phi_{k,w}$ and π_k:

$$P(\theta_k|\cdot)\propto\prod_{k=1}^{K}(\theta_{k,B}^{(d)})^{n_{k,B}^{(d)}+\alpha} \quad (6)$$

where $n_{k,B}^{(d)}$ denotes the number of times word pairs are assigned to topic *k* in document *d*.

For $\phi_{k,w}$, $w\in\{1,2,...,W\}$, its marginal probability distribution is:

$$P(\phi_{k,w}|\cdot)\propto\prod_{k=1}^{K}\prod_{w=1}^{W}(\phi_{k|w,B}^{(d)})^{n_{k|w,B}^{(d)}+\beta} \quad (7)$$

where $n_{k|w,B}^{(d)}$ denotes the number of times word W is assigned to topic k in document *d*.

For π_k, its marginal probability distribution is:

$$p(\pi_k|\cdot)\propto\prod_{k=1}^{K}(\pi_{k,B}|(d))^{n_{k,B}^{(d)}+\gamma} \quad (8)$$

where $n_{k,B}^{(d)}$ denotes the number of times word pairs are assigned to topic k in document d, which reflects the length of document d.

According to Bayesian rules and conjugate properties of Dirichlet, these counts are used to estimate distributions θ, ϕ and π as follows after convergence:

$$\theta_k=\frac{n_k+\alpha}{N_B+K\alpha} \quad (9)$$

$$\phi_{k,w}=\frac{n_{w|k}+\beta}{n_{\cdot|k}+W\beta} \quad (10)$$

$$\pi_k=\frac{n_k\gamma}{N_B+K\gamma} \quad (11)$$

E. Comparison with text topic models

For better understanding the idea and generative process of SSWTM, we compare it with two typical text topic models, i.e., LDA [9] and BTM [17] in this section. Fig. 2 shows a graphical representation of these two models. Latent Dirichlet Allocation (LDA) is a typical topic model proposed by Blei et al. with *bag-of-words* assumption. Based on bag-of-words, LDA assumes that a document consists of multiple topics from a perspective of document generation. First, LDA randomly samples a topic distribution θ_d according to Dirichlet random sampling controlled by α, then gets the words in a document according to the topic-word distribution $z_{d,w}$ controlled by β, and then randomly samples the words are assigned to the topic by using β_k. The generative process of LDA is shown in Fig. 2(1). It shows that each word is assigned to a topic and the topic which assigned to the word is produced by the corresponding topic probability.

The generation process of LDA in accordance with the generation process of a document, but it is highly dependent on the document structure, and the assigned topic of a word depends on other word assignments in the same document. For short texts which lack co-occurrence features and text lengths,

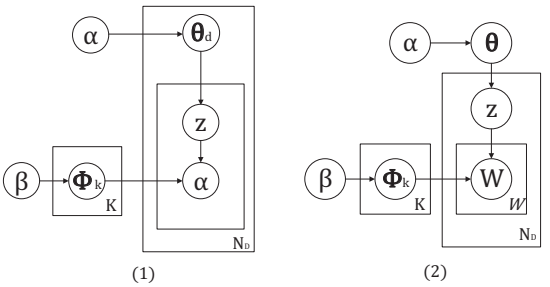


Fig. 2. Graphical representation of (1) LDA, and (2) BTM. Each node in each graph denotes a random variable. A plate denotes an iteration of the model within it. The bottom right corner of a plate shows an iteration process.

LDA could not accurately reflect the latent topic structure from the sparse texts. Therefore, we adopt the idea that modeling the whole corpus rather than each document directly. For normal long texts, the order of the words can not be reflected because of *bags-of-words*, so that we can not understand topic changes in different time slices.

Biterm Topic Model (BTM), illustrated in Fig. 2(2), models on the whole corpus based on the mixture of unigrams and LDA, but with non-uniform texts its performance is not good especially in the situation that normal long texts occupy a large proportion. As a result of the iterative calculation and modeling of all word pairs in the sliding window, too much noise data is introduced, which leads to the decrease of topic quality and the efficiency of BTM.

In a sum, the major trouble of LDA and BTM lies in the lacks of flexibility. Our approach alleviates the sparsity problem within short texts and filter more reasonable words assigned to topics by using a self-adaptive sliding window. Meanwhile, by filtering noise words, SSWTM improves topic quality and the efficiency of iterative computing.

IV. Experimental analysis

A. Databases

In the experiments, we use three datasets to evaluate text modeling effectiveness of LDA, BTM and our approach over non-uniform texts: (1) *the South Sea of China* (*SSC*) news dataset crawled from a popular Chinese social news website[1]; (2) *AlphaGo* news dataset crawled from a Chinese social news website[1] according to keyword "AlphaGo"; (3) Tweets about Obama President Election, referred as "Obama", a typical collection of short texts used by [38] that provides 112155 tweets sampled from November 2008 to November 2009. Besides its content, each tweet includes a timestamp. Details of these datasets are listed in Table I. On the *Obama* dataset, we divide the dataset by month based on the timestamp of each tweet. Therefore, we can do local topic modeling on the texts with same time characteristics. In Fig. 3 and Fig. 4, we plot the document length distribution of *SSC* and *AlphaGo* datasets. We can see that *AlphaGo* has a larger document length and a larger number of documents. This explains more time consumption on *AlphaGo* than *SSC* for each text modeling method.

[1]http://www.sina.com.cn

TABLE I
Basic Information of Datasets

Dataset	SSC	AlphaGo	Obama
#Unique Words	3970	44260	144071
Average Text Length	548.45	62628.81	85948
#Avg Word Pairs in BTM	7567.32	887849.00	139.73
#Avg Word Pairs in SSWTM	1816.05	466012.67	98.02

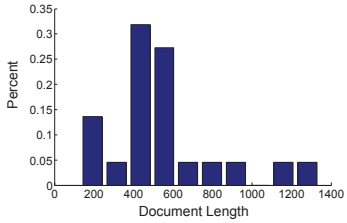


Fig. 3. The document length distribution of SSC dataset

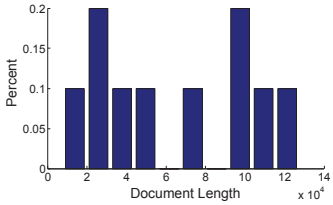


Fig. 4. The document length distribution of AlphaGo dataset

Before modeling over these datasets, we remove meaningless words such as stop words, website addresses, low-frequency words, and punctuations to reduce noisy data within the whole corpus. Then we divide the corpus by timestamps to get a local corpus collection, and also a global corpus that is not cut by timestamps.

B. Experimental settings and evaluation

We compare the proposed approach with LDA and BTM. The reason why we choose these two methods as baselines lies in (1) baseline LDA is a typical topic model to deal with normal long texts; and (2) baseline BTM is a novel topic model to deal with the sparsity problem of short texts and models over the whole corpus rather than each document. After preprocessing, we can validate whether our method is more effective and more accurate than other methods through the evaluation. In the experiments, topics are referred as a collection of relevant words. For LDA, we use the open-source implementation package LDAGibbsSampling[2], and for BTM, we use the open-source implementation BTM[3].

The parameters α, β, and γ were determined via grid search on the smallest dataset (SSC) in our experiments with each parameter in $\{0,\ 0.01,\ 0.1,\ 1,\ 10\}$ to find the best ones. We finally set $\alpha = 50/K$, where K denotes the number of topics, $\beta = 0.01$ and $\gamma = 0.09$ on *i.e. SSC*, $\gamma = 0.0001$ on *AlphaGo* and *Obama*. We set K from 5 to 20 respectively. In all methods, Gibbs sampling was run for 1000 on *SSC* and 100 iterations on *AlphaGo* and *Obama* considering these two datasets are too large. The results of the experiments are given in Section IV-C.

[2]https://github.com/yangliuy/LDAGibbsSampling
[3]http://github.com/xiaohuiyan/BTM

TABLE II

TOPICS EXTRACTED FROM **SSC** DATASET WITH $K = 10$. THE FIRST ROW LISTS THE TOP-20 WORDS WITH HIGHEST PROBABILITIES, WHILE THE SECOND ROW LISTS NON-TOP WORDS RANKED FROM 191 TO 200

LDA	BTM	SSWTM
China/America/arbitration/about/ Japan/foreign minister/claim/ the Philippines/dispute/Sinapore/ Australia/case/Russia/report/result/ but/strategy/Chinese/statement/country	China/the South Sea of China/America/ artitration/about/**ASEAN**/Japan/ foreign minister/claim/the Philippines/ dispute/Singapore/Australia/case/Ruassia/ report/result/Chiese/country/region	China/the South Sea of China/America/ atbitration/about/**ASEAN**/Japan/ foreign minister/claim/the Philippines/ dispute/Sinapore/Australia/case/Russia/ report/but/result/strategy/Chinese/**25**
means/direction/Tsingtao/destroyer/ commander/**enviro**/patrol boat/Yazhou Liu/ocean/environmen/not only/National Defence University/plan/revolution/time/ capacity/formation/cautious/some/true	*revolution*/**change**/America and Japan/ 2016/further/deploy/intention/international community/edition/policy/Chian-US/these/ profit/viewpoint/Beijing/guided missile/Reed Tablemount/Security Bill/hyperpower/army	**Chinese army**/the Philippines/futher/ deploy/Asia-Pacific/premier/international law/declaration/2016/activity/South China Sea/intention/illegality/**China-US**/ and/edition/policy/world/exchange/**Korea**

TABLE III

TOPICS EXTRACTED FROM **ALPHAGO** DATASET WITH $K = 10$. THE FIRST ROW LISTS THE TOP-20 WORDS WITH HIGHEST PROBABILITIES, WHILE THE SECOND ROW LISTS NON-TOP WORDS RANKED FROM 191 TO 200

LDA	BTM	SSWTM
AI/alphago/human/Go/Lee Se-dol/ machine/robot/China/Goole/development/ **intelligence**/technique/competition/domain/ **chess player**/company/**chess**/investment/ man vs.machine	AI/alphago/human/Go/Lee Se-dol/ robot/China/Google/development/data/ intelligence/technique/domain/**chess player**/**chess**/company/future/investment/ problem/capacity	AI/alphago/human/Go/Lee Se-dol/ machine/robot/China/Google/development/ intelligence/**competition**/technique/domain/ chess player/chess/company/future/ investment/problem
victory/Hong Kong/automobile/player/ black/neural network/**Ma**/boost/premier/ **Chen**/vr/Zuchkerberg/thanks/China Japan and Korea/Wan Gang/**Yu**/EU/**Qiao**/**Li**/ **Heng**	hedging/the public/manager/three countries/rise/in favor of/university/earnings/ Kunlun/individual share/plate/kinetic energy/China-US/productivity/share/capital/ this/stabilization/China Japan and Korea/ Go piece	simplify administration/Russia/cross-Straits/**pension**/region/divergence/danger/ peasant/Wanwei/three counties/meeting/ the public/minister/Kunlun/quiz/ kinetic energy/China-US/productivity/ share/achivement

TABLE IV

TOPICS EXTRACTED FROM **OBAMA** DATASET WITH $K = 10$. THE FIRST ROW LISTS THE TOP-20 WORDS WITH HIGHEST PROBABILITIES, WHILE THE SECOND ROW LISTS NON-TOP WORDS RANKED FROM 191 TO 200

LDA	BTM	SSWTM
obama/president/tcot/barack/nobel/ peace/prize/health/news/care/speech/ house/michelle/video/white/plan/people/ administration/today/don	obama/president/tcot/barack/nobel/ peace/health/prize/news/care/speech/ house/michelle/video/white/plan/people/ administration/ today/don	obama/president/tcot/barack/nobel/ peace/health/prize/news/care/speech/ house/michelle/video/white/plan/people/ administration/ today/don
cario/**notre**/Ghana/**joker**/iranian/ iranelection/**fly**/chavez/judge/sotomayor/ jersey/hood/police/supreme/muslim/ results/arts/inauguration/hbo/certificate	qualify/supreme/**low**/Mashable/deserve/ court/**congrate**/mortgage/Iraq/beer/ **awarded**/**awarded**/opinion/birth/H1N1/ Wilson/travel/HIV/Olympics/approval	Rio/interrupts/IOC/audio/bid/**religioncrazy**/ **low**/2016/AIDS/pitch/moon/Olympics/ deserve/election/congrats/mortgage/**crimes**/ NJ/**marijuana**

The *SSC* and *AlphaGo* datasets are representative of normal long texts, and the Twitter dataset *Obama* is a typical collection of short texts. We aim to evaluate the quality of the topics extracted in the situation that two types of texts exist at the same time. We evaluate the quality of topics through topic relevance, that is, the more frequent a word co-occurrence pattern shows on a topic, the higher relevance the topic has. We adopt an external topic evaluation *Pointwise Mutual Information* (PMI) [39], which measures the coherence of topics based on the highest probability of the first T words through pointwise mutual information. PMI evaluates each topic model by using an external text source, i.e., Wikipedia for English corpus *Obama* and Baike[4] for Chinese corpora *SSC* and *AlphaGo*.

Given the number of topics *K*, we choose the highest probability of the first *T* words in each topic, PMI between each word pair is given as follows:

$$\text{PMI}(w_x, w_y) = \log \frac{p(w_x, w_y)}{p(w_x)p(w_y)} \quad x, y \in \{1, 2, ..., T\} \quad (12)$$

where $x \neq y$. Given each PMI(w_x, w_y) of a topic *k*, the PMI of the topic is represented as the median of all PMI(w_x, w_y). In this paper, we use another equation to measure PMI of topic *k* as follows:

$$\text{PMI}(k) = \frac{1}{N(N-1)} \sum_{1 \leq x < y \leq T} \text{PMI}(w_x, w_y) \quad (13)$$

where $P(w_x, w_y)$ is the probability of a co-occurring word pair (w_x, w_y), and $P(w_x)$ is the probability of the word w_x estimated empirically from the external datasets. To evaluate the quality of topics from every method, we set *T* to 5, 10, 20, respectively with $K \in \{5, 10, 15, 20, 25, 30\}$. After computing PMI for each topic and normalizing PMI, we measure the quality of topics with the average value.

C. Experimental results and analysis

To evaluate the quality of all topics in LDA, BTM, and SSWTM, we firstly perform visual comparisons among them and then use PMI to compare topic coherence of the three methods.

1) Qualitative comparison: In SSWTM, K topics are regarded as a probabilistic distribution over W unique words. The number of topics chosen for LDA, BTM, and SSWTM is the same, $K = 10$. We collect top 20 words in each topic into

[4]http://baike.baidu.com

TABLE V
PMI OF LDA, BTM AND SSWTM

Number of Topics		K = 10			K = 15			K = 20			K = 25		
Dataset	Method	Top5	Top10	Top20	Top5	Top10	Top20	Top5	Top10	Top20	Top5	Top10	Top20
SSC	LDA	29.63%	39.38%	36.57%	24.65%	42.52%	37.15%	30.14%	41.20%	31.08%	26.30%	40.06%	33.20%
	BTM	52.26%	78.53%	80.71%	40.58%	80.03%	81.10%	52.63%	72.76%	80.18%	52.53%	78.15%	79.20%
	SSWTM	**64.22%**	**89.46%**	**92.04%**	**60.03%**	**87.29%**	**88.32%**	**57.92%**	**81.55%**	**83.34%**	**59.43%**	**79.45%**	**85.01%**
AlphaGo	LDA	14.19%	23.13%	24.69%	17.59%	18.93%	24.72%	12.74%	19.61%	28.19%	15.48%	25.22%	26.22%
	BTM	34.33%	53.97%	66.07%	35.42%	56.38%	63.69%	39.88%	56.65%	65.06%	40.65%	60.19%	67.98%
	SSWTM	**38.88%**	**55.99%**	**69.83%**	**35.81%**	**57.46%**	**68.39%**	**47.95%**	**61.56%**	**68.35%**	**42.11%**	**65.38%**	**74.68%**
Obama	LDA	25.80%	34.62%	43.28%	26.84%	32.51%	44.07%	22.92%	32.93%	40.89%	23.13%	32.12%	35.59%
	BTM	34.08%	66.82%	85.06%	33.81%	69.32%	86.74%	36.87%	69.61%	90.60%	35.22%	71.08%	91.67%
	SSWTM	**38.78%**	**83.89%**	**98.09%**	**42.62%**	**75.36%**	**98.92%**	**40.11%**	**77.67%**	**96.15%**	**42.74%**	**78.36%**	**95.61%**

a topical word set for each method individually. Table II shows the topics extracted by LDA, BTM, and SSWTM individually. For a topic with high quality, the top words should reflect the main effects according to the topic while the non-top words should be coherent to the top words as much as possible. In Table II, the first row lists the top 20 words with highest probabilities over the *SSC* dataset, while the second row lists non-top words ranked from 191 to 200. All Chinese words for display have been translated into English.

Table II shows the extracted words for topic "the South Sea of China". According to the top-20 words, we can analyse the quality of the extracted topic words. We find that "China", "the South Sea of China", "arbitration" and some country names are shown in the first row for each method. However, comparing BTM and SSWTM to LDA, we find that LDA misses some important information like "ASEAN". Comparing BTM to SSWTM, we find that the focuses of them are different. BTM focuses on the depth of the topics, such as "country" and "region" which could reflect the scope of the event, while SSWTM focuses on the temporal causal relationship, that is, the comprehensiveness of the topic, like "25". The second row in Table II shows the non-top words for each method ranked from 191 to 200 which could be referred as the coherence of the topics. The higher the non-top words related to the topics, the higher quality the topics have. LDA covers a smaller topic range and has some noisy words, such as "not only" and "enviro". BTM also has the problem of duplication of topics, such as "revolution" and "change", and has too much military-related content. SSWTM covers the military-related and international situation and other aspects of the event, such as "Chinese army", "China-US" and "Korea".

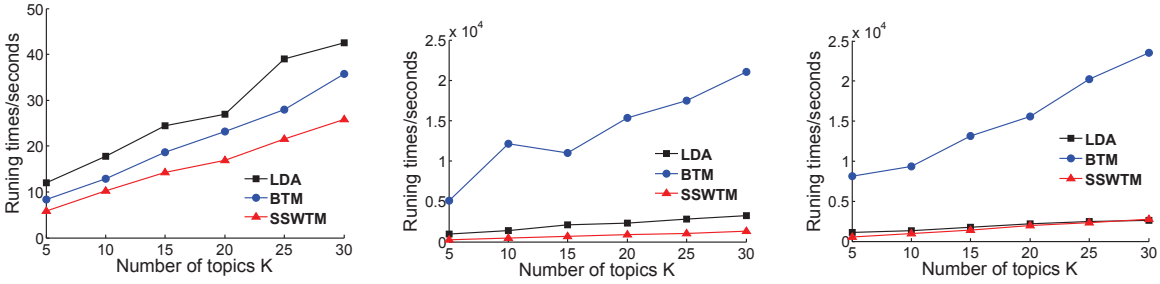
Table III shows the topics with "AlphaGo". In the first row of Table III, the comparison among LDA, BTM, and SSWTM shows that each method has a different focus, such as LDA and BTM are concerned about the event itself while SSWTM is concerned about the nature of event, such as competitive events. In the second row, the noisy words in LDA, such as "Ma" and "Chen", which represent simple Chinese surnames, show low quality of the topics. The comparison between BTM and SSWTM shows that SSWTM can learn comprehensive content of the topic event, including finance like "share", politics like "Russia" and "region", and people's livelihood like "pension".

In the first row of Table IV, it shows that LDA, BTM, and SSWTM have similar results with topic "Obama". However, in the second row, which can reflect the coherence of topic words, shows that the topics learned by LDA include some noise words, such as "notre", "joker" and "fly". And non-top words learned by BTM and SSWTM have several same words. But the result in BTM has overlapping parts, such as "award" and "awarded" which point to the same event. The result in SSWTM includes more related words to topic "Obama", such as "religioncrazy", "crimes" and "marijuana", showing that SSWTM is much more competent to discover different aspects of events related to the topic, such as religion, international politics, and social security.

Through the above visual comparisons, our method has better flexibility when dealing with non-uniform texts, and can learn more accurate and more comprehensive topics due to the self-adaptive sliding window which is better adapted to alternated text length to discover topics, and take into account the corpus size and word co-occurrence patterns.

2) Quantitative comparison: In order to quantitatively analyze the experimental results, we use PMI to evaluate the quality of topics discovered by each method and show the results in Table V. As described in Section IV-B, PMI reflects the degree of relevance among topics and so is a good indicator to evaluate topic models. We express PMI of each method as a percentage in Table V after normalization. We set the number of topics K from 5 to 30 and list results of each method with the number of topics from 10 to 25 due to space limitation. The number of words in each topic T is from 5 to 20.

First, we can see that SSWTM performs significantly better than LDA on the three datasets. BTM also learns more relevant topics than LDA, but the improvement is less than SSWTM. Especially, SSWTM achieves PMI 92.04% on the *SSC* dataset which means the topics learned by SSWTM are much more relevant than LDA, which achieves 36.57%, and BTM, which achieves 80.71%. Secondly, we can find that LDA, BTM, and SSWTM all perform best on the *Obama* dataset, and their performance on *AlphaGo* is relatively poor. For LDA, this is because the lack of discovering word co-occurrence information in short texts, but the large number of short texts like *AlphaGo* can alleviate this problem. For BTM and SSWTM, the results on *Obama* are best because of the stronger word co-occurrence information in short texts and limited word co-occurrence information in long texts like *SSC*. However, the huge number of short texts reduces the quality of topics on *AlphaGo* because of the huge number of word pairs it produces in the iteration process. Our method SSWTM



(a) Time cost on SSC dataset (b) Time cost on AlphaGo dataset (c) Time cost on Obama dataset

Fig. 5. Time cost comparison on (a) SSC dataset, (b) AlphaGo dataset and (c) Obama dataset with K from 5 to 30 and $T = 20$, respectively.

incorporates the word co-occurrence information by modeling on the whole corpus with temporal information so that we can filter some of the low-quality word pairs to improve PMI than BTM and save some of the time cost.

3) Topic evolving: After analyzing topic coherence and topic quality, we also try to investigate on topic evolving according to timestamps. Fig. 6 illustrates the topics of the *Obama* dataset from Nov 2008 to Oct 2009 with $K = 10$. The reason we choose this dataset is that the topics over *Obama* are closely related to the real event during this time sequence, which makes it easy for us to verify the results. The higher probability the topic has, the more important the topic is. We present the highest probability of five topic words below the curve. We can find that these words are all related to "obama", "barack", and "president", which are coherent to the dataset. And the curve reaches its peak in November 2008 and December 2008. During this period, the election ended, and Obama elaborated his economic revitalization plan, which has been widely concerned.

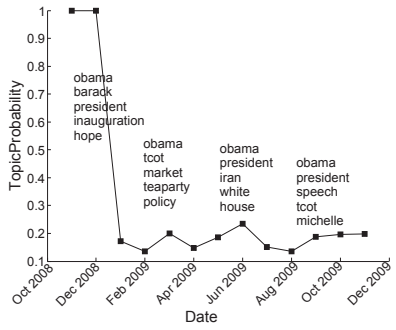


Fig. 6. Topic evolving over *Obama* dataset according to timestamps from Nov 2008 to Nov 2009

4) Efficiency comparison: To demonstrate the efficiency of LDA, BTM, and SSWTM, we plot the time cost of these three methods in Fig. 5 on the *SSC* , *AlphaGo* and *Obama* datasets with K from 5 to 30, respectively. Fig. 5(a) shows the time cost on these three datasets. We can find that on normal long texts, LDA costs more time than BTM and SSWTM. Through Fig. 5(b) and Fig. 5(c), we can see BTM costs much more time than LDA and SSWTM, which means BTM sacrificesefficiency to improve the quality of topics. Especially when the number of topics is 30, BTM costs much more time, while the time cost of LDA and SSWTM increases steadily with the increase in the number of topics. Comparison between Fig. 5(a) and Fig. 5(b) shows that when dealing with long texts, the efficiency of each method is related to the size of the corpus. When the size of the corpus is small (i.e., SSC), through a certain number of iteration process, the time consumption difference among these three methods is not very large. When the corpus is over a certain scale, the time consumption will be huge if BTM is adopted which extracts too many word pairs with noise.

V. Conclusion

This paper has attacked a problem of learning topics over non-uniform texts that consist of both long and short texts, which can be learned through neither normal long text models nor short text models. Hence, we presented a self-adaptive sliding window based topic model, referred to as SSWTM, by modeling the whole corpus which includes both long and short texts through jointly considering the text length and word co-occurrence information. The main idea is based on an assumption that the more similar the semantics of a word pair, the closer distance the word pair has. Experimental results on non-uniform text datasets, which include two Chinese normal long texts datasets and a short text dataset crawled from Twitter, have demonstrated that our method outperforms the baseline methods LDA and BTM. Besides, our method SSWTM is more efficient than LDA and BTM on both long and short texts. All these benefits make SSWTM a better choice for non-uniform text analysis and topic modeling applications, such as social network analysis, public opinion monitoring, and information navigation.

This paper has focused on non-uniform texts and learning latent topics over non-uniform texts effectively and efficiently. There are perspectives for improvment. For the improvement of our algorithm, we would like to evaluate our method on more datasets including both long and short texts and find possible evolving relationships among topics. For applications, we would like to use high-quality topics for event prediction rather than the characterization of an event that has occurred.

Acknowledgment

This research is supported by the National Key Researh and Development Program of China(No. 2016YFB1000900), the National Basic Research Program of China(973 Program)(No. 2013CB329604), the National Natural Science Foundation of China(No. 61229301), and the Program for Changjiang Scholars ans Innovative Research Team in University(PCSIRT) of the Ministry of Education of China(No. IRT13059).

REFERENCES

[1] Y. Wang, E. Agichtein, and M. Benzi, "Tm-lda:efficient online modeling of latent topic transitions in social media," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 123–131.

[2] N. Phan, J. Ebrahimi, D. Kil, B. Piniewski, and D. Dou, "Topic-aware physical activity propagation in a health social network," *IEEE Intelligent Systems*, vol. 31, no. 1, pp. 5–14, 2016.

[3] B. Ma, N. Zhang, G. Liu, L. Li, and H. Yuan, "Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach," *Information Processing & Management*, vol. 52, no. 3, pp. 430–445, 2016.

[4] Y. C. Pan, C. C. Wang, Y. C. Hsieh, T. H. Lee, Y. S. Lee, Y. S. Fu, Y. T. Huang, and L. S. Lee, "A multi-modal dialogue system for information navigation and retrieval across spoken document archives with topic hierarchies," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005, pp. 375–380.

[5] J. Li, J. Wen, Z. Tai, R. Zhang, and W. Yu, "Bursty event detection from microblog: a distributed and incremental approach," *Concurrency & Computation Practice & Experience*, vol. 28, no. 11, pp. 3115–3130, 2016.

[6] O. Jin, N. N. Liu, K. Zhao, Y. Yu, and Q. Yang, "Transferring topical knowledge from auxiliary long texts for short text clustering," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28*, 2011, pp. 775–784.

[7] D. Duan, Y. Li, R. Li, R. Zhang, X. Gu, and K. Wen, "Limtopic: A framework of incorporating link based importance into topic modeling," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 10, pp. 2493–2506, 2014.

[8] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1*, 1999, pp. 289–296.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[10] M. Lan, C. L. Tan, H. B. Low, and S. Y. Sung, "A comprehensive comparative study on term weighting schemes for text categorization with support vector machines," in *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005 - Special interest tracks and posters*, 2005, pp. 1032–1033.

[11] L. Fang, M. Huang, and X. Zhu, "Exploring weakly supervised latent sentiment explanations for aspect-level review analysis," in *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, 2013, pp. 1057–1066.

[12] R. C. Chen, R. Swanson, and A. S. Gordon, "An adaptation of topic modeling to sentences," *Computing Research Repository*, vol. abs/1607.05818, 2016.

[13] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news," *World Wide Web*, vol. 17, no. 4, pp. 723–742, 2014.

[14] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models: #twitter trends detection topic model online," in *COLING*, 2012, pp. 1519–1534.

[15] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers, "Learning author-topic models from text corpora," *Acm Transactions on Information Systems*, vol. 28, no. 1, p. 4, 2010.

[16] Y. Yang, F. Wang, F. Jiang, S. Jin, and J. Xu, "A topic model for hierarchical documents," in *IEEE First International Conference on Data Science in Cyberspace, DSC 2016, Changsha, China, June 13-16*, 2016, pp. 118–126.

[17] X. Cheng, Y. Lan, J. Guo, and X. Yan, "BTM: topic modeling over short texts," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 12, pp. 2928–2941, 2014.

[18] Z. Zhao, S. Feng, Q. Wang, J. Z. Huang, G. J. Williams, and J. Fan, "Topic oriented community detection through social objects and link analysis in social networks," *Knowledge-Based Systems*, vol. 26, pp. 164–173, 2012.

[19] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-document summarization using sentence-based topic models," in *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore, Short Papers*, 2009, pp. 297–300.

[20] X. Fu, L. Guo, Y. Guo, and Z. Wang, "Multi-aspect sentiment analysis for chinese online social reviews based on topic modeling and hownet lexicon," *Knowledge-Based Systems*, vol. 37, no. 2, pp. 186–195, 2013.

[21] V. Subramaniyaswamy and S. Chenthur Pandian, "Effective tag recommendation system based on topic ontology using *Wikipedia* and wordnet," *International Journal of Intelligent Systems*, vol. 27, no. 12, pp. 1034–1048, 2012.

[22] S. Wu, W. Guo, S. Xu, Y. Huang, L. Wang, and T. Tan, "Coupled topic model for collaborative filtering with user-generated content," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 6, pp. 908–920, 2016.

[23] S. T. Dumais, "Latent semantic analysis," *Annual review of information science and technology*, vol. 38, no. 1, pp. 188–230, 2004.

[24] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, 2000, pp. 556–562.

[25] A. Elbagoury, R. Ibrahim, A. K. Farahat, M. S. Kamel, and F. Karray, "Exemplar-based topic detection in twitter streams," in *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pp. 610–613.

[26] B. Huang, Y. Yang, A. Mahmood, and H. Wang, "Microblog topic detection based on LDA model and single-pass clustering," in *Rough Sets and Current Trends in Computing - 8th International Conference, RSCTC 2012, Chengdu, China, August 17-20, 2012. Proceedings*, 2012, pp. 166–171.

[27] H. Zhang and G. Zhong, "Improving short text classification by learning vector representations of both words and hidden topics," *Knowledge-Based Systems*, vol. 102, pp. 76–86, 2016.

[28] H. N. Tran and A. Takasu, "Partitioning algorithms for improving efficiency of topic modeling parallelization," *Computing Research Repository*, vol. abs/1510.04317, pp. 315–320, 2015.

[29] K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2/3, pp. 103–134, 2000.

[30] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[31] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[32] C. E. Moody, "Mixing dirichlet topic models and word embeddings to make lda2vec," *Computing Research Repository*, vol. abs/1605.02019, 2016.

[33] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong, "Topic modeling of short texts: A pseudo-document view," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17*, 2016, pp. 2105–2114.

[34] J. P. Qiang, P. Chen, W. Ding, T. Wang, F. Xie, and X. D. Wu, "Topic discovery from heterogeneous texts," in *28th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2016, San Jose, CA, USA, November 6-8*, 2016, pp. 196–203.

[35] N. Kawamae, "Author interest topic model," in *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23*, 2010, pp. 887–888.

[36] T. Iwata and H. Sawada, "Topic model for analyzing purchase data with price information," *Data Mining and Knowledge Discovery*, vol. 26, no. 3, pp. 559–573, 2013.

[37] J. Zhang and F. Li, "LDA topic evolution based on global and local modeling," *Journal of ShangHai JiaoTong University*, vol. 46, no. 11, pp. 1753–1758, 2012.

[38] Z. Wang, L. Shou, K. Chen, and G. Chen, "On summarization and timeline generation for evolutionary tweet streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1301–1315, 2015.

[39] D. Newman, S. Karimi, and L. Cavedon, "External evaluation of topic models," *Australasian Document Computing Symposium*, pp. 1–8, 2011.