Hot Topic Discovery across Social Networks Based on Improved LDA Model

Chang Liu[1], RuiLin Hu[2,*]
1Chengdu Ruibei Yingte Information Technology Ltd. Company, Chengdu, China
[E-mail: liuchang923@foxmail.com]
2School of Computer and Software Engineering, Xihua University, Chengdu 610039, China
[E-mail: huruilin@stu.xhu.edu.cn]
*Corresponding author: RuiLin Hu

Received September 7, 2021; revised September 26, 2021; accepted October 11, 2021; published November 30, 2021

Abstract

With the rapid development of Internet and big data technology, various online social network platforms have been established, producing massive information every day. Hot topic discovery aims to dig out meaningful content that users commonly concern about from the massive information on the Internet. Most of the existing hot topic discovery methods focus on a single network data source, and can hardly grasp hot spots as a whole, nor meet the challenges of text sparsity and topic hotness evaluation in cross-network scenarios. This paper proposes a novel hot topic discovery method across social network based on an im-proved LDA model, which first integrates the text information from multiple social network platforms into a unified data set, then obtains the potential topic distribution in the text through the improved LDA model. Finally, it adopts a heat evaluation method based on the word frequency of topic label words to take the latent topic with the highest heat value as a hot topic. This paper obtains data from the online social networks and constructs a cross-network topic discovery data set. The experimental results demonstrate the superiority of the proposed method compared to baseline methods.

Keywords: Big data, Hot Topic Discovery, Improved LDA Model, Social Network, Topic Model.

1. Introduction

In recent years, online social network (OSN) has become an indispensable tool for people's daily communication, information acquisition and discussion of hot events. Various social network applications have been established, such as Weibo (Sina Weibo) and Douban in China, Facebook and twitter in America. These social networks produce massive information every day, making the cyberspace gradually bloated and complicated. Hot topic discovery tries to help people analyze and deal with the increasingly overloaded network information, and dig out meaningful content that users commonly concern about from the information flow generated by social network media and portal websites.

Social networks may focus on different events, for example, Toutiao and China news (China News Service) care about current events, Douban and Tieba pay more attention to interests sharing, while Tianya (Tianya forum) and QQ Zone prefer emotional communication. The existing hot topic discovery methods are mainly limited to a single social network, such as Weibo, Twitter and so on. Generally, users in the same social network have a greater chance to interact with each other, resulting in similar topics. However, users in another social network may be more concerned about other events. Hot topic discovery in multiple social networks scenario contribute to shield the differences of various data sources and effectively organize the information from multiple networks into topical information with internal relevance and aggregation. Thus, it helps to solve the problems of information redundancy, dispersion or disorganization.

In popular microblog network platforms such as Weibo and Twitter, there are obvious differences between the content released by users and traditional web texts like news, BBS and personal blog, which are mainly reflect-ed in: there are many new words, a large amount of data and short texts [1]. Traditional text processing methods are mostly based on text vectorization. When confronting a large number of short texts, these methods may suffer from too high dimension, excessive noise [2], or incapable of capturing the high-level semantics of the texts. In recent years, topic modeling has been widely used in many tasks of natural language processing [3-5]. Topic models typically cluster the semantic structure of the document set by unsupervised learning models, and extract topical information (called "latent topic" or "potential topic") by analyzing the co-occurrence among words in the text. Latent Dirichlet Allocation [5] (LDA) is a popular generative topic model which assumes that a topic is generated by the multinomial distribution of words and a document is the hybrid of multiple topics. The prior distributions of document-topic distribution and topic-word distribution are both Dirichlet distribution. LDA model can effectively identify the topic information hidden in large-scale corpus, but it still cannot completely solve the problem of text sparsity.

In order to address the above challenges, this paper proposes a novel hot topic discovery method across social networks based on the improved LDA model, which fuses the text information in multiple social network platforms into a unified dataset, and obtains the latent topic-word distribution by the improved LDA model, then extracts high-frequency topical words from the topic-word distribution, finally evaluates the hotness of topics on the basis of term-frequency. The latent topic with highest hotness value is taken as a hot topic. We crawl data from the Internet (including Weibo, Tianya and China news), and construct a dataset for cross-platform topic discovery. The experimental results show the effectiveness and superiority of our method. The main contributions of this work include below:

1. This paper proposes an improved LDA topic model to dig out latent topics, which effectively alleviates the problem of text sparsity.
2. This paper designs a topic hotness evaluation method which is suitable for multiple social networks, and it derives favorable results.

2. Related works

2.1 Topic models

Topic modeling techniques have been widely used in natural language processing (NLP) to discover latent semantic structures hidden in large-scale corpus. Deerwster et al. [3] first proposed LSA (late semantic analysis) model to transfers a document set into a lexical text matrix, and use singular value decomposition (SVD) method to establish the potential semantic space. Later, Hofmann et al. [4] improved the LSA model and proposed the probabilistic latent semantic analysis (PLSA) model, which assumed that a document contains many latent topics, and these topics are related to words. PLSA retains the feature of dimension reduction of LSA and can capture the semantic information of documents [6], but it cannot describe the dependency between documents and corpus. Blei et al. [5] improved the topic model by introducing Dirichlet distribution and proposed the Latent Dirichlet Allocation (LDA) model. However, there are no clear meaning in LDA latent topics or they may lack pertinence.

Researchers have made a series of improvements on LDA model and successfully applied these models to many different applications. For example, Ramage et al. [7] designed a supervised topic model labeled-LDA, which added clear meaning to the topic model. To remedy the defects of topic words in readability and consistency, Ma et al. [8] presented a topic model based on phrases, enhancing the semantic information of phrases via distributed representation. Zhou et al. [9] tried to solve the problem of slow processing speed of text topic clustering under the background of big data, and developed a LDA text topic clustering algorithm in stand-alone architecture. In addition, some researchers add another level to the three levels of document-topic-word. For example, Titov et al. [10] proposed a multi-granularity model to divide topics into local topics and global topics, and applied it to extract object from online user comments. Chen et al. [11] took users' social relations into account and proposed a "person-viewpoint-topic" (POT) model which could detect social groups and analyze their emotions. Iwata et al. [12] brought the time factor into topic modeling for tracking time-varying consumer buying behavior. Kurashima et al. [13] proposed a geographic topic model to analyze the location log data of multiple users, so as to recommend scenic spots. Chemudugunta et al. [14] suggested to use the model for information retrieval by matching documents at a general topic level and a specific level. Some researches combine topic model with text emotion analysis, for instance, Lin et al. [15] introduced a joint emotional topic (JST) model to analyze the emotional tendency of documents. Wang et al. [16] proposed a Lifelong Aspect-based Sentiment Topic (LAST) model to mine priori knowledge of emotions, views and their corresponding relationships from the other products. Kalaivaani et al. [17] assumed that topics generated are dependent on sentiment distributions and the words generated are conditioned on the sentiment topic pairs, then proposed to make sentiment classification based on LDA topic model. Yin et al. [18] extracted microblogs' topic with LDA model, based on which an algorithm was proposed to find key users in microblog. Kim et al. [19] extended the application of LDA and introduced generalized Dirichlet-multinomial regression (g-DMR) to

reveal the dynamic topic distributions over news articles related to COVID-19. Aiming at the problem of text sparsity in microblog social networks, Cheng et al. [18] designed a biterm topic model (BTM), which expanded the text content by defining word pairs in the text as biterms.

2.2 Hot topic discovery

Hot topic discovery aims to mine meaningful content that users commonly concerned about from the massive information on the Internet. Most researches focus on hot topic discovery in single network scenarios. Wang et al. [19] proposed the topic n-grams model, which could detect the topic and topical phrase in the text simultaneously. Inspired by collective factorization, Vaca et al. [20] managed to connect topics between different time periods. Li et al. [21] tried to discover hot news based on density clustering strategy by exploiting user's interest and topic. Liu et al. [22] proposed a hot topic detection and tracking model TDT_CC to track the heat of a topic in real time. Zhong et al. [23] detected the text topic by clustering the topic tag words, and evaluated the topic heat in combination with the internal and external characteristics of the text. Zhu et al. [24] designed a two-layer network model MSBN based on feature co-occurrence and semantic community division to detect sub-topics in microblog text. Daud et al. [25] applied hot topic detection to discover rising stars in the academic community, and proposed an algorithm HTRS-Rank based on hot topics in authors' publications.

Above mentioned methods are limited to a single social network and cannot meet the challenge of hot topic discovery in cross-network scenarios. Only a few studies consider mining topics across social networks. For example, Zhu et al. [26] proposed to integrate multiple heterogeneous information, and establish user profile from multiple perspectives in a multi-task learning framework. Wang et al. [27] tried to combine the BTM and LDA model to mitigate text sparsity across social networks, and managed to extract hot topics through clustering strategy.

3. Methods

Aiming at the problems of text sparsity and topic hotness evaluation in cross-network scenarios, this paper proposes a novel hot topic discovery method across social networks based on improved LDA model. Firstly, the text data from different social networks are preprocessed and fused to establish a unified dataset, then an improved LDA model based on semantic similarity is proposed to obtain the topic-word distribution, and a topic hotness evaluation method based on word frequency is adopted to calculate the hotness value of different topics. Finally, the topic with the highest score is selected as the hot topic in the unified networks. The overall procedure of our method is shown in **Fig. 1**.

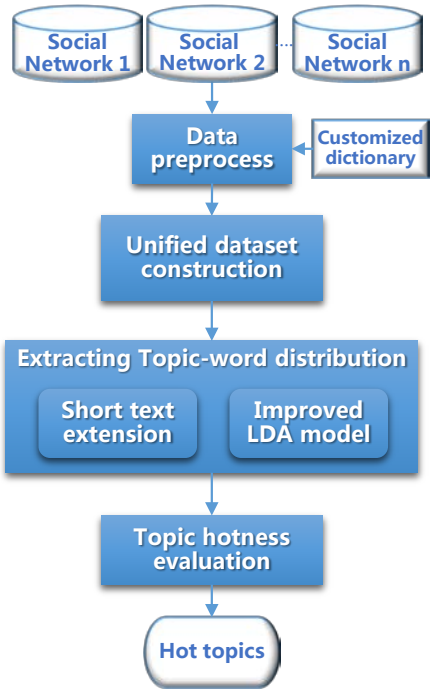

Fig. 1. Overall procedure of our method.

3.1 Unified dataset construction

In cross-network scenarios, social networks often have various data formats and organizations. Microblog information usually includes user name, publishing time, source and content (as shown in **Fig. 2**). News and blog websites usually include title, publishing time, source and content (as shown in **Fig. 3-4**). In order to construct a unified dataset, we extract the content part of data records from different sources to form an independent document. After text preprocessing such as removing repeated text, punctuation, stop words and conversion of different Chinese characters (Traditional Chinese words are transferred to simplified ones), word segmentation and part-of-speech tagging are carried out, and then the documents that are too short are discarded, while too long documents are cut-off. In order to improve the accuracy of Chinese word segmentation, we have manually constructed a small customized dictionary, which contains some popular new words, such as "thumbs-up", "hot-search", "new-era", etc. Finally, the documents from multiple sources are merged into a unified document collection.

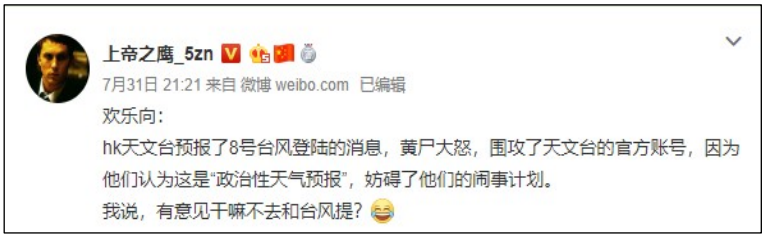

Fig. 5. A sample record from Weibo.

Fig. 6. A sample record from China news.

Fig. 7. A sample record from Tianya.

3.2 Extracting Topic-word distribution

To address the problem of text sparsity in social networks, we combine the short text expansion method based on semantic similarity with LDA model to form an improved LDA topic model, which is used to obtain the topic-word distribution of documents in social media.

3.2.1 Short text extension based on semantic similarity

The core problem to be solved in short text expansion is how to ensure the diversity and semantic consistency of newly added words. Traditional methods simply combine existing words in the original text [18] to form pairs of new words. However, the new words lack diversity and ignore the semantic relationship between words. In recent years, the rapid progress of word embedding technology provides us with new ideas. Word embedding model maps words in natural language into real vector space, so that semantically similar words have similar vector representation. In this paper, CBOW [28] model is borrowed to realize the vector representation of words in the text, and then any new word semantically closest to the original text are chosen by comparing the word vector similarity. Formally, for any short text $\mathcal{T}$ with a length of $\mathcal{S}(\mathcal{S} \geq 2)$ and a minimum text length of $\mathcal{N}(\mathcal{N} > \mathcal{S})$, the number of words to be expanded is $\mathcal{N} - \mathcal{S}$. First, two similar words w_i and w_j are randomly selected from $\mathcal{T}$ so that the similarity $sim(\vec{w}_i, \vec{w}_j)$ of their correspondent word vector $\vec{w}_i$ and $\vec{w}_j$ is greater than a threshold $\tau_1 \in [0,1]$. This paper takes cosine similarity as a measure of word vector similarity as:

$$sim(\vec{w}_i, \vec{w}_j) = \frac{\vec{w}_i \cdot \vec{w}_j}{|\vec{w}_i||\vec{w}_j|} \quad (1)$$

Then select a new word w_k from vocabulary $\mathcal{V}$, and make sure the similarity between its corresponding word vector $\vec{w}_k$ and the former chosen word vector $\vec{w}_i$ and $\vec{w}_j$ is the largest, denoted as:

$$arg\max_{w_k \in \mathcal{V}} sim\left(\frac{\vec{w}_i + \vec{w}_j}{2}, \vec{w}_k\right), w_k \notin T \tag{2}$$

To further ensure the semantic consistency between the new word w_k and the original text, we set another threshold $\tau_2 \in [0,1]$ to filter out any words with lower similarities, as:

$$sim\left(\frac{\vec{w}_i + \vec{w}_j}{2}, \vec{w}_k\right) > \tau_2 \tag{3}$$

Finally, w_k is added to the original text $\mathcal{T}$, and then repeat the above selection process to obtain all the words to be added. It is worth noting that the similarity threshold τ_1 and τ_2 jointly determine the semantic relevance between the new words and the original text. The greater its value, the closer the semantics between the new words and the original text. Therefore, appropriate similarity threshold setting can ensure the consistency between the extended text and the original text, which helps to improve the accuracy of topic modeling.

3.2.2 Improved LDA model.

Topic model is a kind of statistical model that clusters the latent semantic structure of documents in an unsupervised manner. This paper adopts LDA model for topic discovery across social networks. The model assumes that a document is generated by the polynomial distribution of multiple topics, in which each topic is generated by the polynomial distribution of all words in the vocabulary, and the prior distribution of topic-word distribution and document-topic distribution are both Dirichlet distribution. Let $\mathbf{w}$ be a document composed of several words, and $\mathbf{z}$ be a set of topics. Notation $\boldsymbol{\theta}$ denotes a collection of document-topic distributed in the corpus, and $\boldsymbol{\phi}$ represents a collection of topic-word distributed for all topics, ϕ_z denotes the topic-word distribution of topic $z \in \mathbf{z}$, then the LDA model can generally be expressed as a joint conditional probability distribution as:

$$p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\alpha, \beta) = p(\boldsymbol{\phi}|\beta)p(\boldsymbol{\theta}|\alpha)p(\mathbf{z}|\boldsymbol{\theta})p(\mathbf{w}|\phi_z) \tag{4}$$

where α and β are the model hyper-parameters, which represent the a priori preference on topic distribution and word distribution, respectively.

Posts, blogs and articles published by users in social media can be regarded as a set of documents. Through the text expansion method introduced in Section 3.1, documents with too short length $\mathbf{w}_i$ can be expand to the required length of document $\widetilde{\mathbf{w}}_i$. All documents in the network form a corpus $\mathcal{C} = \{\widetilde{\mathbf{w}}_1, \widetilde{\mathbf{w}}_2, \ldots\}$. We uses LDA model to obtain the topic-word distribution of social media text, and estimates the parameters through Gibbs sampling [29]. Topic-word distribution ϕ_z corresponding to latent topic z is computed by:

$$\phi_z^{(w_i)} = \frac{\mathrm{TW}_z^{w_i} + \beta}{\sum_{i=1}^{i=|\mathcal{V}|} \mathrm{TW}_z^{(w_i)} + |\mathcal{V}|\beta} \tag{5}$$

where $\mathrm{TW} \in \mathbb{N}^{|\mathbf{z}| \times |\mathcal{V}|}$ represents the counting matrix of the words assigning to all topics. Notation $\mathrm{TW}_z^{w_i}$ denotes the count of the word w_i assigned to topic z, and $|\mathcal{V}|$ represents the length of the word vocabulary. The word distribution probability vector $\vec{\boldsymbol{p}}_z = \left[\hat{\phi}_z^{(w_i)}\right]_{i=1}^{i=|\mathcal{V}|}$ of topic z can be obtained by parameter estimation. This vector represents the word distribution of latent topics z over all words in the vocabulary $\mathcal{V}$.

3.3 Topic hotness evaluation

At present, there is no widely accepted hotness evaluation metrics for hot topics in social networks. Data sources are diverse and data structures are different from each other in cross-network scenarios. Traditional evaluation methods based on user behavior (such as "like", "comment" and "forwards") rely on user interaction information, so it is difficult to apply to multiple social networks occasion. On this condition, the most direct way to evaluate hotness is by the frequency of key words in the text. As the latent topics discovered by topic models has no clear meaning, the most frequent words in a topic are generally used to describe the topic. Generally, we take the number of C entity words with the highest probability of occurrence in topic z as the tag set $l_z = \{t_1, t_2, \dots, t_i, \dots, t_C\}$ of this topic, where t_i represents the i-th tag in the tag set. Entity words are that with practical meaning (such as nouns, verbs, adjectives, etc.), which can act as sentence components separately. The corresponding function words (such as prepositions, conjunctions, etc.) do not contain practical meaning.

This paper presents a hotness evaluation method based on topic tag frequency. Intuitively, the popularity of a topic is directly proportional to the total number of counts that its tags appear in all documents on the network, and inversely proportional to the total number of documents in a social network and the total number of words in a document. Formally, given a group of social networks $\mathcal{G} = \{G_1, G_2, \dots, G_i, \dots\}$, each social network contains a series of documents $G_i = \{doc_i^1, doc_i^2, \dots, doc_i^j, \dots\}$. The total number of documents in G_i is denoted as M_i. The total number of words in doc_i^j is represented as N_i^j. For a latent topic $z \in \mathbf{z}$ in $\mathcal{G}$, we first find the topic tag set l_z from its corresponding topic-word distribution ϕ_z, and then count the number of times $cnt_{t_k}^{doc_i^j}$ for each tag t_k in the tag set l_z occurred in each document of network G_i, and finally the hotness value h_z of this topic is obtained by weighted summation, as:

$$h_z = \sum_{G_i \in \mathcal{G}} \frac{1}{M_i} \sum_{doc_i^j \in G_i} \frac{1}{N_i^j} \sum_{t_k \in l_z} cnt_{t_k}^{doc_i^j} \tag{6}$$

The topic with the highest hotness value is taken as the hot topic in the unified network. This hotness evaluation method based on topic tag frequency does not depend on any information other than the text itself. Thus, it can be widely used for topic hotness evaluation in multiple social network scenarios.

4. Experiments

4.1 Datasets

This paper collects text data from three social networks (Weibo, Tianya and China news) by web crawlers. The time interval is one month from July 1 to July 31, 2019. In the same time period, the number of documents in different social networks varies greatly, which reflects the different popularity of different social networks. Among them, Weibo has the largest amount of data and the widest popularity, followed by China news and Tianya. In addition, there are significant differences in the text length of a single document in the three networks. We discard documents with a text length of less than 6 words, and cut-off the long documents, so that the document length of Weibo, Tianya and China news does not exceed 250, 500 and 1000 words,

respectively. The statistical information of dataset is shown in **Table 1**.

Table 2. Statistics of the datasets.

Social Networks	# of Documents	Min. # of words	Max. # of words	Ave. # of words	Ave. # of words Extended
Weibo	67 405	6	250	30.5	32.9
Tianya	4 352	10	500	101.9	102.1
China news	27,733	15	1000	363.9	-

4.2 Experimental settings

4.2.1 Evaluation metric.

Inspired by the work of Wang et al. [27], this paper uses the average JS divergence to measure the performances of topic discovery methods. Generally, the larger the JS divergence between any two distributions in a group of topics **z**, the higher discriminative ability between topics. This result in better performance of the topic discovery model. For topic-word distribution ϕ_{z_1} and ϕ_{z_2} of any two topics $z_1 \in \mathbf{z}$ and $z_2 \in \mathbf{z}$, their JS divergence is calculated by:

$$JS(\phi_{z_1}||\phi_{z_2}) = \frac{1}{2}KL\left(\phi_{z_1}||\frac{\phi_{z_1}+\phi_{z_2}}{2}\right) + \frac{1}{2}KL\left(\phi_{z_2}||\frac{\phi_{z_1}+\phi_{z_2}}{2}\right) \quad (7)$$

where $KL(\phi_i||\phi_j) = \sum_{x\in v}\phi_i(x)\, log\frac{\phi_i(x)}{\phi_j(x)}$ represents the KL divergence between the two distributions. Average the JS distance between any two distributions in a group of topics **z** is obtained by:

$$JS_{ave}(\boldsymbol{z}) = \frac{1}{|\boldsymbol{z}| \times (|\boldsymbol{z}|-1)} \sum_{i\in\boldsymbol{z}, j\in\boldsymbol{z}, i\neq j} JS(\phi_i||\phi_j) \quad (8)$$

4.2.2 Baseline methods.

We select the following baseline topic discovery methods to access the performance of the proposed method:

- PLSA [4]: is a statistical model used to analyze the co-occurrence relationship between topics and words in documents. It aims to learns the low-dimensional vector representation of variables through the dependency between observed variables and hidden variables.
- LDA [5]: it is a probabilistic generative model, which assumes that a topic is represented by the multinomial distribution of words and a document is represented by the multinomial distribution of topics. The prior distributions of word distribution and topic distribution are both Dirichlet distribution.
- BTM [18]: is a topic model that uses biterms for text enhancement based on LDA model. Pairs of words in the text are defined as new biterms.

4.2.3 Parameter settings.

As the data obtained from social networks is full of noise, the data is firstly preprocessed,

including removing duplicate text, punctuation and stop words, using zhconv[1] to convert traditional Chinese characters into simplified ones, and using Jieba[2] word segmentation toolkit for word segmentation and part-of-speech tagging. In order to improve the accuracy of Chinese word segmentation, we have manually constructed a small customized dictionary, which contains some network new words, such as "thumbs-up", "hot-search", "new-era", etc.

In the short text expansion stage, all the documents of the three networks in the dataset are merged into a corpus, and the word vectors are pre-trained with the word2vec toolkit of gensim[3]. The word vector dimension is set to be 200. The similarity thresholds are set to $\tau_1 = 0.3, \tau_2 = 0.7$, and then expand the text with length less than 15 in Weibo and Tianya to 15 words, while the data of China news remains unchanged.

In the topic detection stage, the LDA toolkit of scikit-learn[4] is used to discover the topic distributions, and the words with very low frequency (occurs in less than 5 documents) or very high frequency (occurs in more than 50% of documents) in the dataset are filtered out. Number of latent topics, LDA model hyper-parameters α, β are set to 10, 0.1 and 0.1 respectively, and set the number of tags in the topic tag set to $C = 10$.

4.3 Experimental results and analysis

4.3.1 The effects of topic discovery.

In order to evaluate the performance of the topic discovery method proposed in this paper in cross-network scenario, we firstly analyze the performance of topic division on the three social networks. **Fig. 8** shows the average JS divergence value of each method. Compared with the baseline topic models, our proposed method achieves the highest average JS divergence score, it can effectively distinguish the texts of different topics. In addition, compared with the BTM model which also adopts short text expansion algorithm, the JS divergence score of our method is 3.1% (i.e. 0.031) higher, which suggests that the short text expansion method based on semantic similarity has better performance.

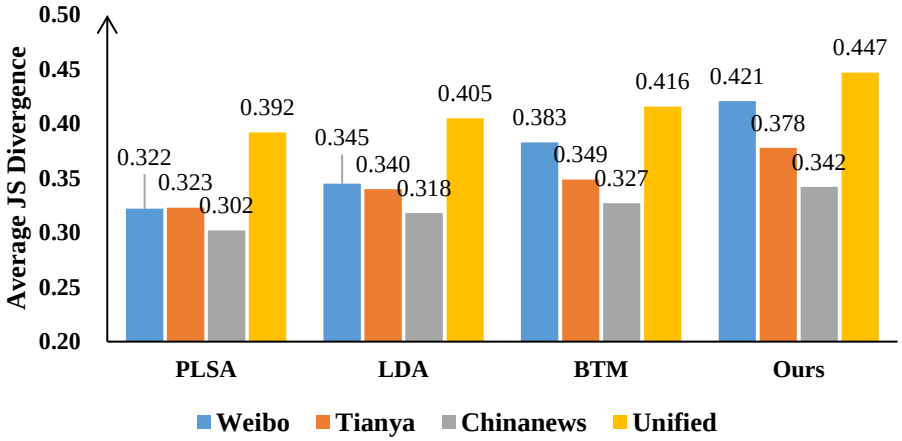

Fig. 9. Results of comparison methods.

1 https://pypi.org/project/zhconv/
2 https://pypi.org/project/jieba/
3 https://radimrehurek.com/gensim/
4 https://scikit-learn.org/

4.3.2 Parameter Sensitivity analysis

This part evaluates the parameter sensitivity of the proposed model on three main parameters: the number of latent topics, and LDA's hyper-parameterα and β. Starting with the default settings listed in Section 4.2.3, each time we only change the value of one parameter, while others remain unchanged. **Fig. 10** shows the influence of the number of potential topics. It can be seen that the average JS divergence value of the two models shows a similar upward trend with the increase of the number of latent topics. When it exceeds 10, the model performances change relatively smaller, which means that the more the number of latent topics, the better topic discrimination can be obtained by the LDA-based topic models, and the performances gradually become stable.

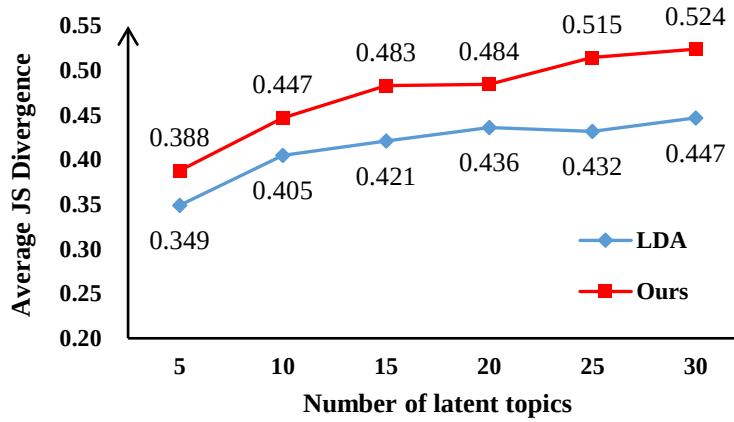

Fig. 11. Model performances on unified dataset w.r.t the number of latent topics.

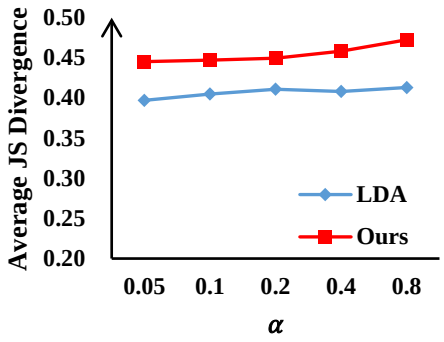

Fig. 12. Model performances w.r.t α.

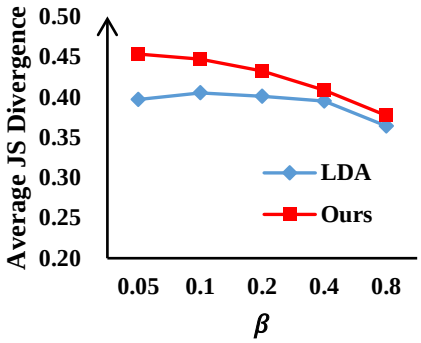

Fig. 13. Model performances w.r.t β.

We can also observe from **Fig. 14** and **Fig. 15** that our proposed approach constantly maintains stable performances with regard to fluctuations of parameter α and β, which proves that our model is robust to hyper-parameter tuning.

4.3.3 Topic heat evaluation effect.

Using the topic hotness evaluation method proposed in this paper, the hotness values of each topic in the dataset are calculated and ranked according to their hotness value. **Table 3** shows the top ten topics and their tag words (original texts are in Chinse, we translated into English hereafter).

Table 4. Hot topics and their tags in unified network

Hot Topics	Topic tags
Top1	Funny, video, super topic, friends, netizens, like, know, fans, husband, see
Top2	Enterprise, development, China, economy, service, work, problem, innovation, country, market
Top3	United States, occurrence, report, police, security, rescue, Hong Kong, China, personnel, reporter
Top4	China, UK, USA, Japan, Iran, country, research, technology, earthquake, report
Top5	China, culture, development, activities, world, cooperation, history, national, international, exchange
Top6	Garbage, city, development, classification, construction, reporter, project, industry, tourism, work
Top7	Children, students, work, journalists, schools, the elderly, teachers, discovery, men, parents
Top8	Health, hospital, patient, problem, reporter, discovery, need, use, treatment, doctor
Top9	Competition, China, player, champion, weather, region, high temperature, world championships, finals, appearance
Top10	Company, market, reporter, growth, year-on-year, price, display, case, information, crime

From the above hot topic tags, we can identify some hot topics in China in July 2019, including online funny videos, China's economic and enterprise development, US security reports, etc.

In order to verify the effectiveness of our method in cross-network scenarios, we conducted experiments on each single network. The results of hot topics are shown in **Table 5-6** The topics mentioned in the results of unified networks are underlined. It can be seen that the hot topics of each social network are part of the hot topics of unified networks. Meanwhile, the hot topics of each social network also include topics not mentioned in our previous results. This is because these topics are hot topics in current social networks, but they cannot be regarded as hot topics across multiple networks. In addition, Weibo and China news contain more hot topics, while Tianya contains less. This may be because hot topics in social networks are usually related to major events in reality. Information from Weibo and China news is usually related to these events, and texts from Tianya is typically related to daily life. Compared with daily life, social events are more likely to become a hot topic.

Table 7. Hot topics in Weibo

Hot Topics	Topic tags
Top1	Super topic, funny, Wang Junkai, video, health, netizens, like, Xiao Zhan, Wang Yibo, fans
Top2	Children, work, garbage, classification, China, netizens, mothers, know, daughters, teachers
Top3	Man, mobile phone, video, discovery, China, company work, release, occurrence, police

Table 8. Hot topics in Tianya

Hot Topics	Topic tags
Top1	Friends, husbands, cheating, know, children, deal with, men, like, things, life
Top2	China, Japan, United States, South Korea, Huawei, country, company, economy, world, enterprise
Top3	United States, Iran, Britain, Russia, Trump, report, country, Syria, occurrence, President

Table 9. Hot topics in Chinanews

Hot Topics	Topic tags
Top1	China, problems, work, education, development, country, United States, society, students, activities
Top2	Enterprise, economy, construction, industry, market, service, growth, technology, development, institution
Top3	Garbage, discovery, classification, occurrence, life, children, work, time, hospital, site

5. Conclusion

This paper studies the problem of hot topic discovery in cross-network scenarios, and proposes a hot topic discovery method based on the improved LDA model. The LDA model is improved by the text expansion method based on semantic similarity, which effectively alleviates the problem of text sparsity. The model exploits topic tag frequency to evaluate the topic hotness. Our method is verified and evaluated on three social networks: Weibo, Tianya and China news. Experimental results demonstrate that the proposed method can effectively distinguish the texts of different topics, and achieves better performance over baseline methods. In addition, the experiment also suggests that the hot topics of each social network are part of the unified social network, and the hot topics in online social networks are usually closely related to major events in real society. Through hot topic mining of multiple online social networks, we may also analyze the popular views in the network, and know well about the hot topics in different regions and the evolution of hot topics.

References

[1] H. Lu, Y. Lou, B. Jin, and M. Xu, "What is discussed about covid-19: A multi-modal framework for analyzing microblogs from sina weibo without human labeling," *Computers, Materials and Continua*, vol. 64, no. 3, pp. 1453-1471, 2020. Article (CrossRef Link)

[2] X. Fern, and C. Brodley, "Cluster ensembles for high dimensional clustering: an empirical study," *Journal of Machine Learning Research*, vol. 22, 01/01, 2004.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391-407, 1990. Article (CrossRef Link)

[4] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence*, pp. 289-296, 1999.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993-1022, 2003.
[6] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 1–2, pp. 177–196, 2001.
[7] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pp. 248-256, 2009.
[8] J. Ma, J. Cheng, L. Zhang, L. Zhou, and B. Chen, "A phrase topic model based on distributed representation," *Computers, Materials and Continua*, vol. 64, no. 1, pp. 455-469, 2020. Article (CrossRef Link)
[9] Z. Zhou, J. Qin, X. Xiang, Y. Tan, Q. Liu, and N. N. Xiong, "News text topic clustering optimized method based on TF-iDF algorithm on spark," *Computers, Materials and Continua*, vol. 62, no. 1, pp. 217-231, 2020.
[10] I. Titov, and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proc. of International Conference on World Wide Web*, pp. 111-120, 2008. Article (CrossRef Link)
[11] H. Chen, M. Wang, H. Yin, W. Chen, X. Li, and T. Chen, "People opinion topic model: Opinion based user clustering in social networks," in *Proc. of International Conference on World Wide Web*, pp. 1353-1359, 2017. Article (CrossRef Link)
[12] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior," in *Proc. of IJCAI International Joint Conference on Artificial Intelligence*, pp. 1427-1432, 2009.
[13] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura, "Geo topic model: Joint modeling of user's activity area and interests for location recommendation," in *Proc. of ACM International Conference on Web Search and Data Mining*, pp. 375-384, 2013. Article (CrossRef Link)
[14] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," *Advances in Neural Information Processing Systems*, pp. 241-248, 2007.
[15] C. Lin, and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proc. of International Conference on Information and Knowledge Management*, pp. 375-384, 2009. Article (CrossRef Link)
[16] S. Wang, Z. Chen, and B. Liu, "Mining aspect-specific opinion using a holistic lifelong topic model," in *Proc. of International Conference on World Wide Web*, pp. 167-176, 2016. Article (CrossRef Link)
[17] P. C. D. Kalaivaani, and R. Thangarajan, "Enhancing the Classification Accuracy in Sentiment Analysis with Computational Intelligence Using Joint Sentiment Topic Detection with MEDLDA," *Intelligent Automation and Soft Computing*, vol. 26, no. 1, pp. 71-79, 2020.
[18] M. Yin, X. Liu, G. He, J. Chen, Z. Tang, and B. Zhao, "A Method of Finding Hidden Key Users Based on Transfer Entropy in Microblog Network," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 8, pp. 3187-3200, 2020. Article (CrossRef Link)
[19] J. H. Kim, M. H. Park, Y. Kim, D. Nan, and F. Travieso, "Relation Between News Topics and Variations in Pharmaceutical Indices During COVID-19 Using a Generalized Dirichlet-Multinomial Regression (g-DMR) Model," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 5, pp. 1630-1648, 2021. Article (CrossRef Link)
[20] X. Cheng, X. Yan, Y. Lan, and J. Guo, "BTM: Topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928-2941, 2014. Article (CrossRef Link)
[21] X. Wang, A. McCallum, and X. Wei, "Topical N-grams: Phrase and topic discovery, with an application to information retrieval," in *Proc. of IEEE International Conference on Data Mining, ICDM*, pp. 697-702, 2007. Article (CrossRef Link)
[22] C. Vaca, A. Mantrach, A. Jaimes, and M. Saerens, "A time-based collective factorization for topic discovery and monitoring in news," in *Proc. of International Conference on World Wide Web*, pp. 527-538, 2014. Article (CrossRef Link)

[23] J. Li, and X. Ma, "Research on hot news discovery model based on user interest and topic discovery," *Cluster Computing*, vol. 22, no. 4, pp. 8483-8491, 2019. Article (CrossRef Link)
[24] Z. H. Liu, G. L. Hu, T. H. Zhou, and L. Wang, "TDT_CC: A Hot Topic Detection and Tracking Algorithm Based on Chain of Causes," in *Proc. of International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 27-34, 2018. Article (CrossRef Link)
[25] M. Zhong, "Hot Topic Discovery in Online Community using Topic Labels and Hot Features," *Technical Gazette*, vol. 26, no. 4, pp. 1068-1075, 2019. Article (CrossRef Link)
[26] G. L. Zhu, Z. Z. Pan, Q. Y. Wang, S. X. Zhang, and K. C. Li, "Building multi-subtopic Bi-level network for micro-blog hot topic based on feature Co-Occurrence and semantic community division," *Journal of Network and Computer Applications*, vol. 170, pp. 10, Nov, 2020. Article (CrossRef Link)
[27] A. Daud, F. Abbas, T. Amjad, A. A. Alshdadi, and J. S. Alowibdi, "Finding rising stars through hot topics detection," *Future Generation Computer Systems-the International Journal of Escience*, vol. 115, pp. 798-813, 2021. Article (CrossRef Link)
[28] D. Zhu, Y. Wang, C. You, J. Qiu, N. Cao, C. Gong, G. Yang, and H. M. Zhou, "MMLUP: Multi-source & Multi-task learning for user profiles in social network," *Computers, Materials and Continua*, vol. 61, no. 3, pp. 1105-1115, 2019. Article (CrossRef Link)
[29] X. Wang, B. Zhang, and F. Chang, "Hot Topic Community Discovery on Cross Social Networks," *Future Internet*, vol. 11, no. 3, pp. 60, 2019. Article (CrossRef Link)
[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proc. of The 1st International Conference on Learning Representations*, 2013.
[31] 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. Article (CrossRef Link)

Chang Liu received a M.S. from the Zhengzhou University, and a B.S. in Software Engineering from Zhengzhou University of Light Industry. He is an Research Associate of Chengdu Ruibei Yingte Information Technology Ltd. Company, Chengdu, China. His current research interest includes artificial intelligence, big data, hot topic detection, emotion recognize, and pose estimation.

Ruilin Hu received a B.S. degree in measurement and control technology and instrumentation program from the Shanghai University of Electric Power, China. He is currently pursuing the master's degree with Xihua University, China. His research areas include software engineering and data mining.