Expert Systems With Applications 115 (2019) 346–355

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Natural disaster topic extraction in Sina microblogging based on graph analysis

Tinghuai Ma[a,b,*], YuWei Zhao[a], Honghao Zhou[a], Yuan Tian[c], Abdullah Al-Dhelaan[c], Mznah Al-Rodhaan[c]

[a] School of Computer & Software, Nanjing University of information science & Technology, Jiangsu, Nanjing 210-044, China
[b] CICAEET, Jiangsu Engineering Center of Network Monitoring, Nanjing University of information science & Technology, Nanjing 210-044, China
[c] Computer Science Department, College of Computer and Information Science, King Saud University, Riyadh 11362, Saudi Arabia

ARTICLE INFO

Article history:
Received 4 April 2018
Revised 25 June 2018
Accepted 8 August 2018
Available online 9 August 2018

Keywords:
Topic detection
Community detection
Natural disaster
Sina microblogging
Graph analysis

ABSTRACT

In this paper, we will propose a novel approach based on graph analysis which will use community structure detection algorithm to detect topics in the keywords graph of micro-blogging data. Furthermore, considering the specificity of the Sina microblogging, we propose novel keywords filtering model and graph generation algorithm to meet the dual requirements of topic detection and community detection. We validate our approach on a big natural disaster dataset from Sina micro-blog, in which about 10^3 micro-blogging posts with about 10^4 distinct feature tags. The experimental results definitely revealed the relationship between the keywords and the natural disaster topics. Our methodology is a scalable method which can adapt to the changes in the amount of data. Especially, we can get abundant information about natural disasters in the topic detection and help the government guide the rescue of disasters.

1. Introduction

In recent years, the natural disasters happen frequently in the world which caused the majority of researchers' attention, such as debris flow, floods, earthquakes, and typhoons. How to find these disasters and lock the disasters' areas at the first time has become the focus of everyone's attention. With the development of Internet and the growing popularity of various communication devices, people can no longer obtain information and exchange information only confined to the traditional media. Social networking has become the open social media services based on the new network platform (Ma et al., 2018a; Ma, Shao, Hao, & Cao, 2018b). Due to the particularly rapid development of microblogging platform, it has not only become the important means of users to explore the news events, express their views and insights, but also become the important places to disseminate hot topics (Ma et al., 2016a). Taking Sina microblogging (China) as an example, the active users has reached 100 million and the daily number of microblogging posts up to more than 300 million as of June 2017. The topic detection of microblogging will help the community and the government to find those natural disasters and emergencies which was difficult to predict in time. Above all, it will help the government to keep abreast of the network public opinion and guide the public opinion correctly.

The previous hot topic detection (Benny & Philip, 2015; Huifang, Yugang, Xiaohong, & Zhou, 2016; Yu, Zhao, Chang, & He, 2014; Zhou, Zhong, & Li, 2014) is focused on how to improve the accuracy of the algorithm, enhance the topic of expression and improve hot topics sorting rules. It usually only consider the topic detection of the text content, but ignored the topic expression. In this paper, we propose a graph based method of using community detection algorithm (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) to detect topics which can find valuable topics in the massive and messy data in the form of a keywords graph. Further, we can find extra valuable information of the disaster, such as the location, the date and so on.

One of the key challenges in microblogging data is that due to the Chinese microblogging content is short, dispersed, sparse, noisy and complicated (Huang, Yang, Mahmood, & Wang, 2012; Lo, Chiong, & Cornforth, 2017; Yang, Lin, Lin, & Liu, 2016a; Yeh, Tan, & Lee, 2016), the correlation of the different texts is weak, so the latent topic is very difficult to detect or extract (Lv et al., 2016). For the general method, the dispersion and sparseness of the text are reduced by recalculation of decentralized data or considering other factors when building topic models (Chang, Hsieh, Chen, & Hsu, 2015). The researchers utilize the auxiliary information of the

* Corresponding author.
E-mail address: thma@nuist.edu.cn (T. Ma).

https://doi.org/10.1016/j.eswa.2018.08.010
0957-4174/

text, such as considering the reprinting relationships to enhance the relevance of the text (Ma et al., 2016b). In this paper, a new keywords filtering model is proposed with the Natural Language Processing (NLP) word segmentation technique to remove the useless or redundant noise feature words. After the initial filtering, we continuously consider to select the feature words rigorously from three aspects which will make them to be the most representative keywords under both local and global conditions. We only pick those feature words with high weight to obtain a dense set of keywords. And then we use the graph generation model to transform the abstract content into the visible words graph. We utilize the relationship between the two keywords and calculate the association of each pair of keywords to obtain the kernel edges' value. To rule out the influence of the order of keywords, we use the two-by-two products to reduce the error. Finally, the community detection algorithm is proposed to divide closely related keywords into the same community which can extract the topics and show the content more intuitively. Specially, we utilize Gephi to obtain the graphical representation and we can know the distribution of each topic at a glance. Moreover, we focus on the detection and extraction of natural disaster topics. Our purpose is to find the emergency natural disasters in the big data, thereby to analysis the disasters and solve the disasters. As a result, we designed an algorithm that meet the requirement of topic extraction from the natural disaster data using the keyword extraction method mentioned above. We evaluate our proposed topic detection model on real Sina-microblogging dataset by comparing the topic detection performance. We also compared our method with other four topic detection methods in the experimental section.

The contribution of this work is threefold:

(1) We propose a novel keywords filtering model which gets representative and highly dense feature words from the complicated microblogging data. We not only considers some properties of the feature word itself, but also considers its importance in the local text as well as the global one.
(2) We utilize a graph generation algorithm transforming the abstract text data into a graph of the keywords. We also accurately calculate the value of association between two keywords to form a keywords map structure which can visualize the topics obviously.
(3) We utilize the heuristic community detection algorithm to group the disordered keyword graph in an orderly manner and proposed a topic extraction rule to describe the contents of each topics in detail which extract topic accurately and discover the relationship between topics.

The remainder of the paper is organized as follows: Section 2 reviews the related work; Section 3 describes the proposed topic detection method; Section 4 describes our experimental setup and the experimental results. We conclude the paper in Section 5.

2. Related works

The classic model of topic detection is proposed by Blei, Ng, and Jordan (2003) called the latent Dirichlet allocation (LDA) method, which is a Bayesian networks-based topic model widely used to identify topics from 2003. This method overcomes the shortcoming that the parameters are increased with the number of document set is increasing. Many researchers improve the LDA model according to different scenarios. Ye, Du, and Fu (2016) proposed a probabilistic generative model named Microblog Features Latent Dirichlet Allocation (MF-LDA) to extract microblog topics. They incorporate five microblog's unique features into the analysis of LDA model to improve the performance of the traditional one. Amoualian, Clausel, Gaussier, and Amini (2016) proposed two models for modeling topic and word-topic dependencies between consecutive documents in document streams. The first extension makes use of a Dirichlet distribution to balance the influence of the LDA prior (α and β) to topic and word-topic distribution of the previous document. The second extension makes use of copulas, which constitute a generic tool to model dependencies between random variables. Chen, Li, Guo, and Guo (2016) proposed a FSC-LDA model which combining the text clustering methods and feature selection methods. It can identify the number of topics adaptively, keep short micro-blog texts features better and make the result more stable. Most of the above research is based on offline data, but there are also approaches on the online data flow for topic detection and prediction. Dang, Gao, and Zhou (2016) proposed a new early detection method for emerging topics based on Dynamic Bayesian Networks in micro-blogging networks. They build a DBN-based model by the conditional dependencies between features to identify the emerging keywords and cluster the emerging keywords into emerging topics by the co-occurrence relations between keywords. Xie, Zhu, Jiang, Lim, and Wang (2016) proposed a sketch-based topic model with dimension reduction technique to achieve bursty topics from Twitter. They developed a "sketch of topic", which provides a "snapshot" of the current tweet stream and can be updated efficiently.

Although the probabilistic topic model has a wide range usage and dimensionality reduction, but it requires people to set the number of topics in advance and can't detect those new topics that haven't appear in the training data. Therefore, the unsupervised clustering method is considered to achieve automatic data analysis and topic detection which has become another focus of the researchers. Liang, Yilmaz, and Kanoulas (2016) proposed a dynamic clustering topic model method DCT for short-length streaming text. It can effectively model both the temporal nature of topics in streaming text and the sparsity problem of short text. Fitriyani and Murfi (2016) proposed a mini batch K-means method to detect topics in big datasets effectively. It is used to reduce the computational time and improve the accuracy of the algorithm. Further, although some approaches didn't use the classic clustering algorithm directly, it is developed based on the idea of clustering. Pang et al. (2015) proposed a clustering-like pattern across similarity cascades (SCs) which can truncate a similarity graph with a set of thresholds in a series of subgraphs to capture topics with maximal cliques. Then a topic-restricted similarity diffusion process is proposed to efficiently identify real topics from a large number of candidates.

In addition to the probabilistic topic model and the unsupervised clustering approaches above, more and more researchers use the graph analysis method to detect the words' association in the graph or networks. The method of graph analysis can add the topic extraction visualization and is suitable to utilize the community detection algorithm which seems more interesting and flexible (Rong, Ma, Tang, & Cao, 2018; Zhang, Ma, Cao, & Tang, 2016b). Cigarrn, ngel Castellanos, and Garca-Serrano (2016) proposed an approach based on Formal Concept Analysis (FCA), a fully unsupervised methodology to group similar content together in the matically-based topics and to organize them in the form of a concept lattice. Zhang, Wang, Cao, Wang, and Xu (2016a) proposed a hybrid relations analysis approach to integrate semantic relations and co-occurrence relations for topic detection. The approach fuses multiple relations into a term graph and detects topics from the graph using a graph analytical method. With the analysis of community detection methods, Sayyadi and Raschid (2013) proposed a graph analytical approach for topic detection. They used a KeyGraph algorithm to convert text data into a term graph based on co-occurrence relations between terms. Then they employed a community detection approach to partition the graph. Eventually, each community is regarded as a topic and terms within

the community are considered as the topic's features. Hachaj and Ogiela (2016) proposed a graph-based approach to trending topics in microblogging posts. They use the hashtags filtration model first and then use community graph generation approach to detect the most popular topics. However, these models tend to fail on the short and noisy text from social network. In this paper, we proposed a graph based topic detection and extraction method to discover the natural disasters in Sina Microblogging. We focus on the novel keywords filtering model, improved graph generation method, community detection method and topic extraction method which can improve the performance of topic detection in all directions. The difference from the method proposed by Hachaj in 2016 is that we choose different feature word selection mechanisms for Sina Microblogging while Hachaj adopts hashtags filtration, and we extend topic detection to topic extraction based on graph process while Hachaj detects topics . Further, the starting point of solving the problem is inconsistent, we want to detect those latent topics like natural disasters which may be not obvious or popular. In other words, hot topic detection in Hachaj and Ogiela (2016) is much easier to detect. Thereby, it will be great challenges to solving the short and noisy text.

3. Topic detection method

In this section, we will discuss our topic detection model (TDGA) in detail. Firstly, due to the original data is noisy and sparse, we utilize a keywords filtering model to pick out those important keywords from the feature words set. Secondly, we use the graph generation algorithm to transform the keywords into a keywords graph for the intuitive presentation. Thirdly, in order to detect the topics in the graph, we utilize the community detection algorithm to partition the keywords with the same topic into the same community. Finally, each word has its own topic label and weight. We can extract the topic words with the topic extraction algorithm. We divide the processing into the following four parts:

3.1. Keywords filtering

In traditional keywords filtering approach (Chen et al., 2016), it usually select the keywords directly depending on the word frequency or the TF-IDF value of feature words. However, the method based on word frequency will recognition some pointless and high frequency words as the keywords and the method of TF-IDF will decrease the performance of keywords filtering due to the short text of Micro-blogging. In this paper, we propose a new keywords filtering model for the selection of the feature words in short Chinese text from three aspects: 1) part of speech, 2) frequency and 3) common characteristic. We pick out those semantic and representative feature words as the keywords (Yan, Hua, & Hu, 2016) which not only exist in a blog record but also appear in several communities and be shared by many users.

1) Part of speech analysis

There are many differences between the words in the Chinese text, and the type of the part of speech are numerous and complicated. Compared with the English text, the Chinese text is more difficult to extract the useful words. We discovered that the nouns, adjectives, times and verbs can explain the natural disaster topics discussed in the blog better than other type of information. So we need to filter the lexicality of the feature words and only reserve the keyword w that belongs to the nouns, verbs, adjectives and times. Therefore, only the words that meet these four parts of speech are defined as the word w.

2) Frequency analysis

The importance of a word is inextricably linked to it's frequency, so we firstly consider the effect and influence of the frequency change on the lexicality of the feature words. The definition of the keywords with frequency analysis is shown as follow:

$$\frac{\#\sum_{1}^{N}(w \in l_i)}{\#\sum_{1}^{N} l_i} > T_1 \tag{1}$$

l_i is the unordered list of feature words of Micro-blogging i. $\#\sum_1^N(w \in l_i)$ represents the number of feature word w in unordered list of the whole blogs. $\#\sum_1^N l_i$ is the number of the whole feature words of the whole blogs. When the threshold is set too high, the result will be in the lack of topic semantics and incomplete. When the threshold is set too low, the model may not filter the useless information.

3) Common characteristic analysis

Since we need to find one or more keywords which have been used by different users, it is important to determine whether the feature word appears in multiple user posts. When the word is associated with multiple users, it will provide a great deal of help to the community partition. The common characteristic words are defined as follow:

$$\frac{\#l_w}{\#N} > T_2 \tag{2}$$

$\#l_w$ is the number of users that have used feature word w. $\#N$ is the number of all users. The formula represents that more users used the feature word w when the higher value of T_2 selected.

Therefore, after the filter of the feature words, we get the important keywords set. We defined our keywords set W as (3) according to 1) 2) 3):

$$W = \left(w : \exists U = (u_1, u_2, \ldots, u_N) : \forall u_i, w \in l_i \bigcap \frac{\#\sum_{1}^{N}(w \in l_i)}{\#\sum_{1}^{N} l_i} > T_1 \bigcap \frac{\#l_w}{\#N} > T_2 \right) \tag{3}$$

where:$U = (u_1, u_2, \ldots, u_N)$-is a list of indices of blogs.

The pseudo code of our algorithm is presented in Topic Detection based on Graph Analysis Algorithm(TDGA) showed in Algorithm 1.

3.2. Graph generation

In order to embody the semantic characteristics of a topic and further detect the topics with the community detection algorithm, we utilize the graph generation algorithm to transform the preprocessed keywords set W into words graph form:

$$G = (V = W, E = W \times W) \tag{4}$$

Where:

V-is a list of vertices of graph G which equals W.

E-is a list of edges of graph G which equals Cartesian product of $W \times W$.

How to determine the association between two random keywords is the kernel of graph generation algorithm. We do the following definition: an edge $e_{i,j}$ between nodes (keywords) w_i and w_j is added if the keywords co-occur in at least one document. In order to achieve the precise partition of the community in the community detection algorithm, we also need to calculate the weight of each edge. We use the conditional probability between the two nodes as the weight of its edge:

$$p(w_i|w_j) = \frac{p(w_i w_j)}{p(w_j)} = \frac{df(w_i w_j)}{df(w_j)} \tag{5}$$

Where $p(w_iw_j)$ and $df(w_iw_j)$ respectively denote the probability and the number that words w_j and w_i occur in the same blog post, while $p(w_j)$ and $df(w_j)$ respectively mean the probability and the number the words w_j appears in the micro-blog posts.

In order to ensure that the weight between the two nodes is unique and uniform, we finally define the weight between nodes as follow:

$$U(w_i, w_j) = p(w_i|w_j)p(w_j|w_i) \tag{6}$$

The value of $U(w_i, w_j)$ is between 0 and 1 which represents the tightness between the two nodes w_i and w_j. When the value is closer to 1, the relationship between the two words is stronger which seems more likely to be grouped into a community. On the contrary, they will be more likely to be divided into different communities.

3.3. Topic detection algorithm

With the above processing, we have get the strong correlation and dense keywords graph. Further, we need to detect the topics in the irregular words graph. To solve this problem, we utilize the community detection algorithm to partition the keywords with the same topic into the same community which is the core of this paper. Generally speaking, the original community detection algorithm uses the correlation between each node and the weight of each edge to divide the users in the community network into the same or different communities according to their close degree. Further, the users divided into the same community seem to have the similar hobbies or similar relationships. In this paper, we use community detection algorithm to detect the correlation between keywords. In this way, it will detect the latent topics in the noisy data set. When pairs, triplets and more keywords appear commonly together among different users, the related keywords can be divided into a community which represents as a topic. Through the above analysis, we can detect the topics in the keywords graph and extract the topics easily. It shows that the community structure is stable and the internal contact is stronger when more keywords appear closely together in different users.

In order to get the community structure in the network, we utilize the modularity optimization algorithm (Sharma & Annappa, 2017; Tsung, Ho, Chou, Lin, & Lee, 2017) to realize the partition of the communities which measures the strength of the network community structure. The formula of the modularity is shown as follow:

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_ik_j}{2m}\right)\delta(c_i, c_j) \tag{7}$$

where:

A-is adjacency matrix;

$A_{ij} = \begin{cases} 1 & \text{node i is connected to node j;} \\ 0 & \text{other.} \end{cases}$

k_i-is the degree of node i;

c_i-the community which vertex is assigned;

$\delta(c_i, c_j)$-is 1 if $c_i = c_j$ and 0 otherwise;

$m = \frac{1}{2}\sum_{ij}(A_{ij})$

The value of the modularity depends on the community distribution of nodes in the network which can be used to quantitatively measure the quality of the network community. The closer the value is equals to 1, the strength of the community structure is stronger, that is the better partition of communities. So the optimal network community can be obtained by maximizing the module Q.

In this paper, the heuristic algorithm is used to optimize the objective function, and the value of the module is utilized to judge the performance of the community. The main process of the algorithm is as follow:

1) Given a weighted network with N vertices, assign each vertex to a different community in the network. Each vertices is an independent community.
2) For each vertex i, consider the neighborhood node j of i to evaluate the value growth of the modularity of the community where j is located if i is placed in. Then put i into the j's community with the largest growth of the modularity.
3) Create a new network consisting of vertices in the community through steps 1) and 2).
4) Think of the divided community as a new vertex in the new network.
5) The weights between the new vertices are composed of the weights of the vertices in the corresponding communities, and the weights of the vertices in the community form the self-connection of the corresponding vertices.
6) Step up to the end of the algorithm until the module degree is no longer increased.

3.4. Topic extraction algorithm

With the end of the community detection algorithm, each keyword belongs to a certain topic. In other words, each topic has a set of keywords which can describe the topic in detailed. In order to make the topic more concise and legible, each topic can be characterized by top 10 keywords with the highest weight of being related to the certain topic. With the analysis in Section 3.1, we defined the score $S(z_{ri})$ representing the importance of the keyword w_i in the exposition of the rth $(1 < r < K)$ topic. The formula of $S(z_{ri})$ is defined as follow which is a transformation of formula (3):

$$S(z_{ri}) = \frac{\#\sum_{1}^{N}(w \in l_i)}{\#\sum_{1}^{N} l_i} \times \frac{\#l_w}{\#N} \tag{8}$$

We calculate the score for each of the keywords in each topic and sort keywords in descending order. Then, we only select the top 10 keywords as the final representation of each topic. Due to our experimental data is related to the natural disaster, we can find what, where, when and how the natural disaster happened. Once the information of natural disasters has been identified, the government can make decisions quickly and minimize the damage.

4. Experiment

The prototype of the proposed method was implemented mainly in JAVA SE 1.8. The data in this experiment is stored in mySQL and the database contains the relationship between the blogs and the feature words. The community structure detection and visualization is carried out under Gephi 0.91. We are concerned with the Chinese text data on Sina Miroblogging. People can publish any opinions, events and feelings about recent news or events, personal encounters, emotional expressions and so on the Sina platform. All we have to do is to crab the blog posts they post and analysis those data. So it can be imagined that the noise data in these data is very large which increased the difficulty of topic detection. In this paper, we focus on our experimental data on natural disasters. Due to the wide variety of natural disasters, we only focus on earthquakes, typhoons, floods, debris flow and heavy rain. The data was downloaded by crawler application written in python through the application of Sina microblogging API interface. We follow the microblogging account like the global weather, news, government, other authoritative bloggers and ordinary users and crawl the blog posts considering the microblogging user id and the content from June 2017 to August 2017. Due to the number of data collection restrictions of microblogging API in-

Table 1
The summary of data present in the two datasets.

Dataset name	Blogs	Feature words	Distinct feature words	Keywords
Natural disaster	3595	53,960	9984	1936
Typhoon	1460	22,900	4871	1471

Table 2
An example of NLP technique.

	Content
original	Rainstorm warning, everybody pay attention to safety.
NLP	Rainstorm /n warning /vn , /w everybody /rr pay attention to /v safety /an ./w

Table 3
The meaning of the acronyms.

The part of speech categories	Speech abbreviation
Noun	n (nr ns nt nz)
Time	t (tg)
Verb	v (vd vn vx vi vl vg)
Adjective	a (ad an ag al)
Pronoun	r (rr rz ry)
Punctuation	w

terface and the strict review of the user application for the developer platform, we finally collected 3595 records related to natural disaster data which including earthquakes, typhoons, floods, debris flow and heavy rain. In order to facilitate the discussion of experimental results, we also collected 1460 data related to the natural disaster typhoons. From this period of time we crawled, there are two strong typhoons which are harmful to our country called Haitang and Hato. All the data we collected is in Chinese. In order to facilitate the elaboration and analysis, the data in the table was translated into English and the data in the figure was not changed. In the next step, we will use these data independently to test the experimental performance. The overall experimental data is shown in Table 1.

To deal with the complicated data, we utilize NLPIR technology to preprocess data. NLPIR (Yang, Jin, & Chen, 2016b) is a powerful word segmentation tool, it can not only divided text content into the semantic feature words and phrases, but also be able to mark part of speech of each feature word which will be contributed to the selection of characteristic words. With the help of NLPIR, we transfer each micro-blogging data into the following form and treat each word as a feature word. Remove the duplicate words and we get distinct feature words. Here is an example as Table 2.

The ordinary data contains the content of the data (Rainstorm warning, everybody pay attention to safety.). Then we utilize the NLP technique for word segmentation like the third line in the Table 2. Each word has its own part of speech and the meaning of the abbreviation can be found in the Table 3.

To evaluate the proposed topic detection method, we compared our method with Graph Analytical Approach (GAA) (Sayyadi & Raschid, 2013), Modified Latent Dirichlet Allocation (MF-LDA) (Ye et al., 2016) and UWTD (Pang et al., 2015). The method of MF-LDA is the representation of probability model, it considers the Chinese Microblogs' features and calculate the weight of feature words from five aspect to improve the traditional LDA topic model. We choose a clustering like pattern methods called (UWTD) to be the representation of the clustering approaches. This method investigate methods from the perspective of similarity diffusion and proposed a clustering like pattern across similarity cascades to calculate the similarity between each pair of words and cluster them in several groups to generating multi-granularity topics. The method of GAA is similar with our method which also focuses on the graph model. It utilizes a collection of documents as a keyword co-occurrence graph. However, it only picks those keywords with high frequency and then uses an off-the-shelf algorithm based on the betweenness centrality metric for finding the edges between communities in a network which is completely different from ours. In addition, they do not have further research on topic words extraction.

In our experiments we want to examine:

1. explore the magnitude of T_1 values to the effects of experimental results.
2. the discussion on the scalability of the method.
3. the performance and accuracy of detecting topics.

We have set the threshold of parameter T_2 with 0.005 (Hachaj & Ogiela, 2016) : T_2 parameter guarantees that a keyword has to appear in 0.5% of whole analyzed population. This is not less, because as we can see in Table 1 the number of the distinct feature word is only approximately 5 smaller than overall number of feature word. That means the single type of keyword might appear rarely and with our approach it would be possible to remain those that not only appeared several times in tweets but also in tweets of different users.

We applied our method to the entire data set and randomly selected 50%, 25%, and 10% of the data for the experimental comparison. In the paper, we use the criterion of the modularity to measure the performance of community partition and use NMI (Normalized Mutual information) to compare the similarity of two community partition graphs (lizan, Liugang, & Yang, 2013; Shang et al., 2013). Moreover, we use the value of precision to evaluate the accuracy of topic detection. The definition is as follows:

Definition 1 (Modularity:)**.** It is also known as modular measures and is a common method of measuring the strength of a network community structure. The value of modularity can be found in Section 3.3 of Eq. (7).

Definition 2 (Normalized Mutual Information)**.** It is a basic concept in information theory and is often used to describe the statistical relevance between two systems, or the same information is contained between the two systems. In this paper, the criterion of the normalized Mutual Information can be used to measure the similarity of the community partition results of the same data with different volumes. When the volume of the data becomes smaller, the number of nodes in the community becomes smaller, too. We delete the non-existent nodes in the comparison graph from the original graph and make the calculation. The value of NMI is formulated as:

$$NMI(a,b) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij} log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{C_a} N_{i.} log(\frac{N_{i.}}{N}) + \sum_{j=1}^{C_b} N_{.j} log(\frac{N_{.j}}{N})} \quad (9)$$

where C_a is the original number of community (we use the result of community partition based on 100% data as the original number of community), C_b denotes the number of found community. The matrix N represents the confusion matrix, where N_{ij} is simply the number of nodes in the original community i that appear in the detected community j. $N_{i.}$ and $N_{.j}$ are the sum over row i and column j of confusion matrix respectively. N is the number of nodes. If the obtained partition perfectly matches the original one, its NMI-value takes the maximum of 1. While, if the found

Table 4
Modularity of TDGA and GAA.

	T_1	0.005	0.01	0.025	0.05
TDGA (this work)	Modularity	0.467	0.643	0.279	0.266
GAA	Modularity		0.344		

Table 5
NMI of natural disaster dataset with TDGA and GAA.

% N of dataset	100%	50%	25%	10%
TDGA (T_1=0.01)	1.000	0.994	0.990	0.473
GAA	1.000	0.675	0.562	0.327

partition is entirely independent of the original partition, NMI = 0 corresponds to the situation that the entire network is found to be one community.

4.1. The effects of T_1

In this section we will discuss the issues raised in the previous section. In our graphical visualization results, the nodes having the same color share the same community. First, we use the data of the typhoon to test the effect of T_1 with our method and we set the values of T_1 to 0.005, 0.01, 0.025 and 0.05 respectively, as shown in Fig. 1(a)–(d) and Table 4. From the first four figures in Fig. 1, we can see that with the value of T_1 increasing, the number of keywords in the graph gradually decreases. When T_1 is 0.005 in Fig. 1(a), most of the noise data and irrelevant information is remained, making the representation of the topic becomes obscure and can't accurately extract each topic. It contains different four topics. Due to each topic has more than hundreds keywords, we can't extract the content of the topics accurately. At the same time, the value of the modularity in Table 4 decreases compared with T_1=0.01 which means the effect of community partition is worse. We can obviously find out that when T_1 is 0.01, the topic information expression is the most representative one and the value of the modularity in Table 4 is the highest. The community detection algorithm accurately divides typhoon data into two topics, one is called Typhoon Haitang and the other one is called Typhoon Hato in Fig. 1(b), so the performance of the community partition is also optimal. There is an interesting phenomenon when T_1 is 0.025. As we can see in the Fig. 1(c), the dataset is divided into two groups. The big one merges two topics into one and we can hardly define the content of the topic. The other one is about the influence of Typhoon. The value of the modularity is only higher than the worst one. When the T_1 value is 0.05, due to the text information is too concise and a large number of related information is filtered, the expression of the topics becomes incomplete and fragmented. It can only detect the topic of Typhoon Haitang, and make the same topic Typhoon Hato into Typhoon and Rainstorm in Fig. 1(d). At the same time, the value of the modularity decreases. Therefore, we can draw the conclusion from the discussion of the above experiment that when T_1 set value of 0.01, the detection of topics achieves the best performance and the best effect of community partition.

As a comparison, we use the same typhoon data to test the similar method of GAA and the result is showed in Fig. 1(e) and Table 4. From Table 4 we can see, the modularity of GAA is 0.334 which just happens to form a community. It means the performance of community partitioning is not ideal. Further, we can find that the GAA method partition the data into three communities in Fig. 1(e). The community with red nodes is about typhoon Haitang. The green one is associated with the impact of the typhoon and the yellow one is about the emergency rescue. From the above analysis, we can discover that GAA only detect the topic of typhoon Haitang and missed the topic of typhoon Hato. In addition, it discovered two more topics which belong to the topic of typhoon. These phenomena are powerful enough to show that performance of topic detection by GAA is poor. The possible reason of this phenomenon is that the parameters in GAA are fixed and it may not be suitable for this experiment. As a conclusion, our method achieves better performance of the complex dataset.

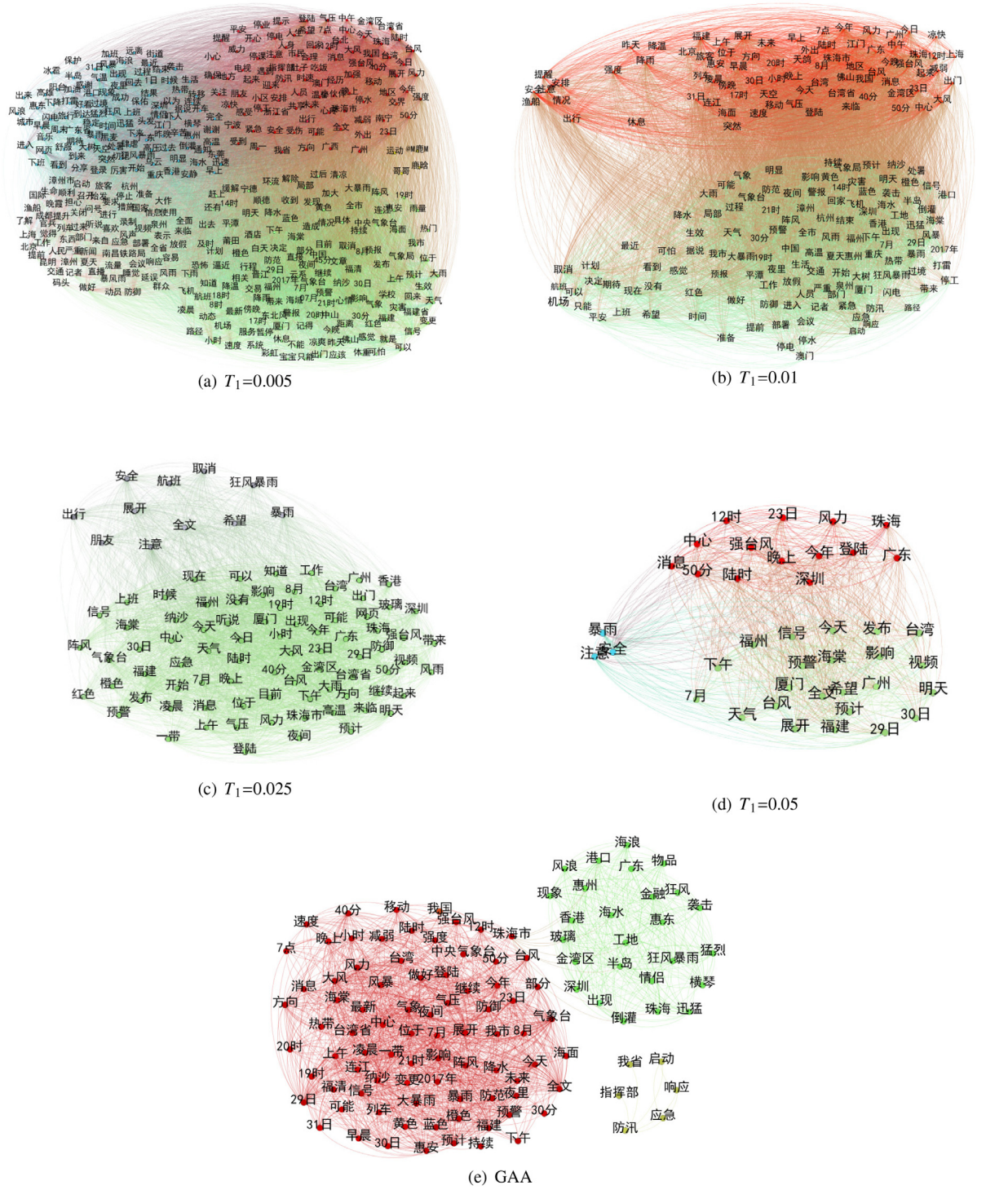
4.2. The scalability of the method

Next, we tested the scalability of the proposed model. Since the experiment is optimal when $T_1 = 0.01$, the experiment is carried out at $T_1 = 0.01$. We utilize the whole natural disaster data and randomly selected 50%, 25% and 10% of the whole natural disaster data for topic detection. The results of the experimental performance are judged by the value of the NMI and the structure of topic detection. We compare the experimental results with different volumes of data to 100% dataset results in Fig. 2 and Table 5. We can see that the three graphs with 100%, 50% and 25% data are similar in the results of community partition which contains six different topics in Fig. 2 (a)–(c). The boundaries of each community partition are clear and the topic distribution of each graph is similar. However, the experimental result with the 10% dataset in Fig. 2(d) is poor that the boundary of community partition is disappearing in the community with pink nodes and it is difficult to distinguish those different topics. We can only extract the topic that the earthquake happened in Si Chuan. From Table 5, we can clearly see that the value of NMI only changes 0.01% with the amount of data is more than 25%. In the meantime, when data changes to 10%, there is a substantial decline of the NMI value which means the community partition changes. The possible reason for this phenomenon is that due to the reduction of the amount of data, most of the feature words are not taken into account, the correlation between the feature words is weakened which resulting in the unsatisfactory experimental results. However, the experimental results are excellent when the data set is bigger than 10% which represents the nice scalability of the model.

Fig. 3 and Table 5 show the community partition results with different volumes of data computed by GAA. It is not difficult to

Algorithm 1 Topic Detection based on Graph Analysis Algorithm.

Input: Original blogs($b_1, b_2, \ldots, b_k$)
Output: Keywords set $W(w_1, w_2, \ldots w_m)$
Topic zr of keywords set($z_{r1}, z_{r2}, \ldots, z_{ri}$)
data preprocessing with NLPIR technology
for b_1 to b_k **do**
  for each feature word ($w_1, w_2, \ldots, w_n$) **do**
    if w_l satisfy formula 3 **then**
      add feature word w_l to keywords set W;
    end if
  end for
end for
get keywords set;
for w_1 to w_m **do**
  if $p(w_i|w_j) > 0$ **then**
    connect node w_i and node w_j;
  end if
end for
get keywords graph;
use the heuristic algorithm to detect the topics in the keywords graph
get topic z_r of keywords set($z_{r1}, z_{r2}, \ldots, z_{ri}$)
4 compute the score of each keyword in topic z_r
use the scores to sort the keywords in descending order
pick out top 10 keywords as the representing of the topic z_r



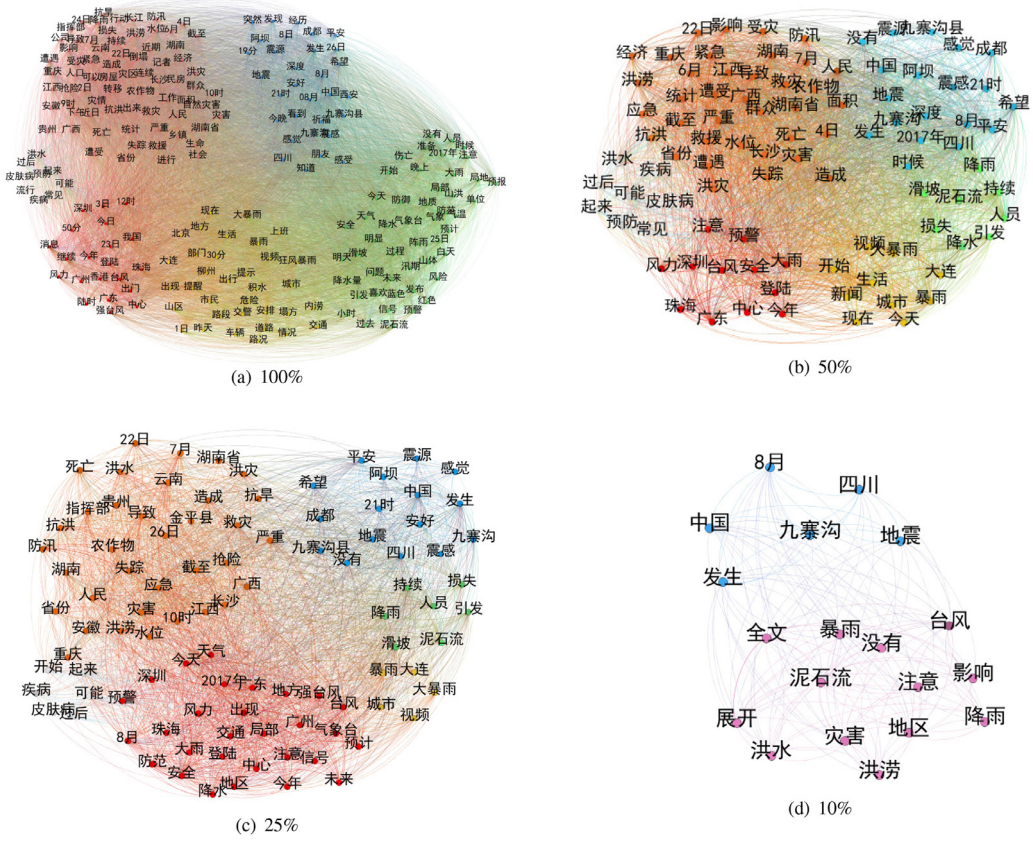
(a) T_1=0.005 (b) T_1=0.01

(c) T_1=0.025 (d) T_1=0.05

(e) GAA

Fig. 1. Results of topic detection by TDGA with (a)T_1=0.005, (b)T_1=0.01, (c)T_1=0.025, (d)T_1=0.05 and (e)GAA with the typhoon dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

find that the results of the community partition are changed with different rates of data from Fig. 3. There are two topics with orange and blue nodes being detected in the first three figures and only one topic with blue nodes being detected in all of these four figures. As soon as the volume of data set changes, the result of the topic detection changes. It is more intuitive to see from Table 5 that the value of NMI is changed about 32.5% with 50% data compared by 100% data. When the volume of data comes to 25% and 10%, the value of NMI changed about 50% and even more. Compare with our method, the performance of our methods are significantly ahead of GAA and extremely stable above 25%. Although there is a large decline at 10%, there is also a weak advantage better than GAA. Therefore, the scalability of GAA is worse than our method. In general, it can be concluded that the proposed



(a) 100% (b) 50% (c) 25% (d) 10%

Fig. 2. Results of topic detection with randomly selecting (a)100%, (b)50% (c)25%, (d)10% of natural disaster dataset by TDGA with T_1=0.01. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

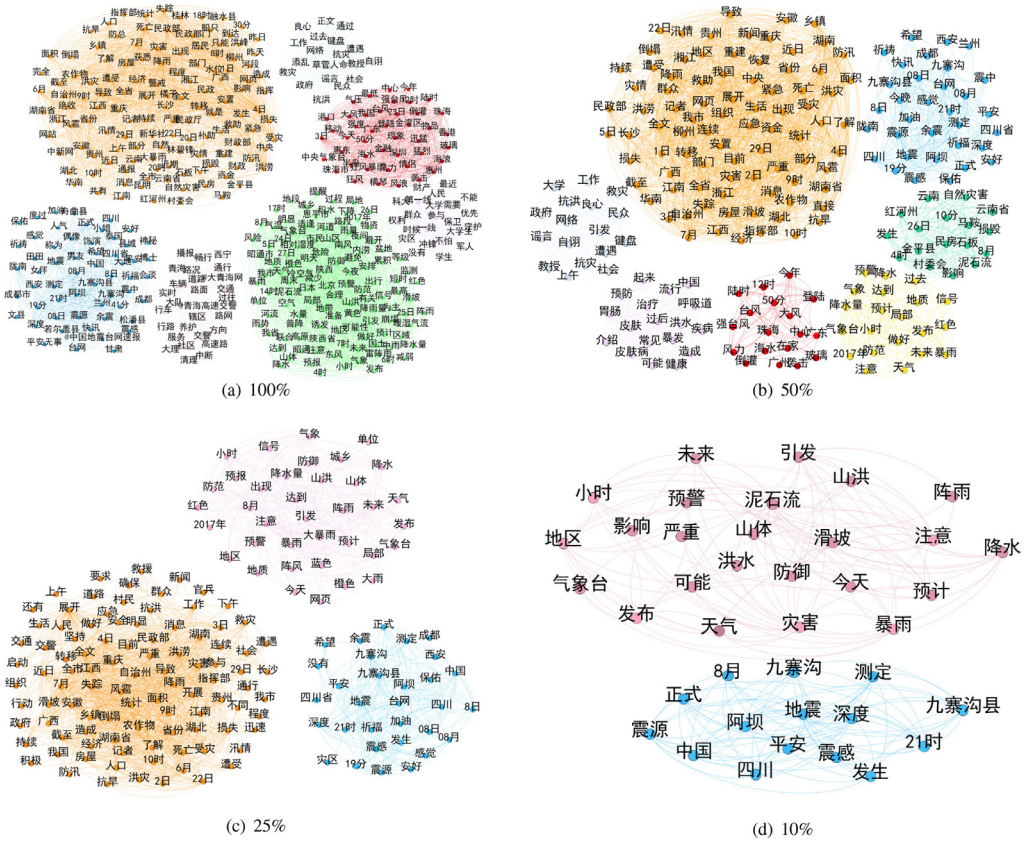
Table 6
The evaluation of different methods.

Method	Precision	Recall	F_1
MF-LDA	0.807	0.772	0.789
UWTD	0.843	0.735	0.785
GAA	0.851	0.694	0.764
TDGA	0.916	0.848	0.880

method has better scalability and applicability with different volumes of the data.

4.3. The accuracy and performance of detecting topics

We utilize the whole natural disaster dataset on the detection of the topics and the extraction of topics as shown in Fig. 2(a), Tables 6 and 7. Firstly, we discuss the accuracy of the four different methods' results. We use recall, precision and F_1 score to measure the quality of the experiment. Table 6 shows that all four methods can be used to detect topics in our experiment dataset effectively and that the two methods with graph analysis (GAA and TDGA) is better than MF-LDA and UWTD in precision. What's more, our method has not only the highest precision and a higher recall and F_1 score which demonstrates the effectiveness of our proposed approach. The possible reason for the low recall of GAA is that the algorithm pruning lots of feature words with the process of finding Maximum Clique. Secondly, we discuss the topic detection performance of our method. From Fig. 2(a), we can clearly find the whole data is divided into five big communities and one small community. The first big cluster was connected with floods which marked with orange nodes (tags: flood, flooding, collapse, etc.). We can also know that with the floods happening, the houses and the crops are destructed. Many people missed and died which brought great disaster to the country and the society; The second with earthquake which marked with blue nodes (earthquake, Jiuzhaigou, sense of shake, etc.). At the same time, the people of the whole country are praying for the peace for Jiuzhaigou; Next was debris flow marked with green nodes (debris flow, landslide, etc.). What's more, we can infer that the cause of the debris flow is the continuous rainfall. Fortunately, there is no casualties; The fourth with heavy rain which marked with yellow nodes (rainstorm, waterlogging, heavy rain, etc.). The heavy rain hampered traffic and caused landslides and waterlogging in Dalian and Beijing; The last one is typhoon which marked in red nodes (typhoon, landed, wind power, etc.). There is an interesting topic in the gray nodes (common, skin disaster, prevention, etc.) which related with the disease and prevention after flooding. In this topic, we can find that after the disaster people will pay attention to health information. It is a signal for the hospital to prepare the related drugs and examinations. Although this topic is not in our expectation, it has little effect on the final results of our experiment. In summary, the proposed method is able to detect these five topics related with natural disasters unsupervised and accurately which will provide a great help for the follow-up disaster relief and post-disaster reconstruction. From the result of topic extraction showed in Table 7, we can understand the topics easily. The first five topics are talking about different natural disasters and the last one is concluding the skin disaster or epidemic after the flood. Through the analysis above, the method proposed in this paper can not only effectively detect the topics related to natural disasters in microblogging, but can also extract the topics concisely which is out of state-of-the-art.



(a) 100% (b) 50% (c) 25% (d) 10%

Fig. 3. Results of topic detection with randomly selecting (a)100%, (b)50%, (c)25%, (d)10% of natural disaster dataset by GAA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
The results of topic extraction.

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
24th	August	25th	Beijing	today	flood
9 o'clock	8th	partial	30 o'clock	50 o'clock	after
continuous	tonight	rain	rainstorm	this year	prevention
flood	China	debris flow	heavy rain	3rd	skin disaster
anti-flood	Jiuzhaigou	observatory	road	typhoon	possible
Hunan	Chengdu	precaution	appear	landing	common
Guangxi	Aba	future	waterlogging	HongKong	epidemic
crops	earthquake	geology	traffic	Shenzhen	disaster
resulting in	blessing	night	police	wind	public
affected	peaceful	risk	remind	center	attention

5. Conclusion

In this article, we redefined the criteria for the extraction of keywords in the short Chinese text considering the factors such as word frequency, part of speech, contextual relationship and so on. From the topic extraction results, the novel keywords filter model achieved good results. Further, we can find that the proposed method has good robustness and stability by changing the size of the data in the course of experiment. So that the topics can be detected normally even though the amount of data is small or the information is missing. Based on our experimental results of the unsupervised community partition and topic detection with the community detection algorithm, our approach extracts different topics accurately and efficiently. Therefore, when natural disasters occur, we can accurately detect the occurrence of natural disasters by collecting relevant information in microblogging and guide the public to the activities related to disaster relief. Moreover, we compared our method with other three methods in the topic detection, the value of precision, the scalability of the model and other aspect. The experiments show that our method is better than GAA in all aspects and has the best accuracy compared with MF-LDA, UWTD and GAA which demonstrating the superiority of our method.

In recent years, natural disasters happened frequently which has brought a great threat to the human and society. Due to most of the natural disasters usually happen suddenly, how to make the early detection of such topics seize the time nodes or be able to achieve predictions will be the focus of our research. The analysis or predictions of this type however is not a straightforward task, we want it to be a goal of our future research.

Acknowledgments

This work was supported in part by National Science Foundation of China (No.61572259, 61173143), Special Public Sector Research Program of China (No. GYHY201506080), and was also supported by PAPD. The authors extend their appreciation to the

Deanship of Scientific Research at King Saud University for funding this work through research group no. RGP-VPP-264.

References

Amoualian, H., Clausel, M., Gaussier, E., & Amini, M. R. (2016). Streaming-lda: A copula-based approach to modeling topic dependencies in document streams. In *ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 695–704).

Benny, A., & Philip, M. (2015). Keyword based tweet extraction and detection of related topics. *Procedia Computer Science, 46*, 364–371.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research Archive, 3*, 993–1022.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics, 2008*(10), 155–168.

Chang, Y. C., Hsieh, Y. L., Chen, C. C., & Hsu, W. L. (2015). A semantic frame-based intelligent agent for topic detection. *Soft Computing, 21*(2), 1–11.

Chen, Y., Li, W., Guo, W., & Guo, K. (2016). Popular topic detection in chinese micro-blog based on the modified lda model. In *Web information system and application conference* (pp. 37–42).

Cigarrn, J., ngel Castellanos, & Garca-Serrano, A. (2016). A step forward for topic detection in twitter: An fca-based approach. *Expert Systems with Applications, 57*(C), 21–36.

Dang, Q., Gao, F., & Zhou, Y. (2016). *Early detection method for emerging topics based on dynamic bayesian networks in micro-blogging networks*. Pergamon Press, Inc..

Fitriyani, S. R., & Murfi, H. (2016). The k-means with mini batch algorithm for topics detection on online news. *International conference on information and communication technology*.

Hachaj, T., & Ogiela, M. R. (2016). Clustering of trending topics in microblogging posts: A graph-based approach. *Future Generation Computer Systems*.

Huang, B., Yang, Y., Mahmood, A., & Wang, H. (2012). Microblog topic detection based on lda model and single-pass clustering. In *International conference on rough sets and current trends in computing* (pp. 166–171).

Huifang, M. A., Yugang, J. I., Xiaohong, L. I., & Zhou, R. (2016). Hot topic discovering algorithm for microblog based on discrete particle swarm optimization. *Computer Engineering*.

Liang, S., Yilmaz, E., & Kanoulas, E. (2016). Dynamic clustering of streaming short documents. In *The ACM SIGKDD international conference* (pp. 995–1004).

lizan, Liugang, & Yang (2013). A genetic algorithm for community detection in complex networks. *Journal of Central South University, 20*(5), 1269–1276.

Lo, S. L., Chiong, R., & Cornforth, D. (2017). An unsupervised multilingual approach for online social media topic identification. *Expert Systems with Applications, 81*.

Lv, Y., Ma, T., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2016). An efficient and scalable density-based clustering algorithm for datasets with complex structures. *Neurocomputing, 171*, 9–22. doi:10.1016/j.neucom.2015.05.109.

Ma, T., Jia, J., Xue, Y., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2018a). Protection of location privacy for moving *k*nn queries in social networks. *Applied Soft Computing, 66*, 525–532. doi:10.1016/j.asoc.2017.08.027.

Ma, T., Rong, H., Ying, C., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2016a). Detect structural-connected communities based on BSCHEF in C-DBLP. *Concurrency and Computation: Practice and Experience, 28*(2), 311–330. doi:10.1002/cpe.3437.

Ma, T., Shao, W., Hao, Y., & Cao, J. (2018b). Graph classification based on graph set reconstruction and graph kernel feature reduction. *Neurocomputing, 296*, 33–45. doi:10.1016/j.neucom.2018.03.029.

Ma, T., Wang, Y., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A., & Al-Rodhaan, M. (2016b). LED: A fast overlapping communities detection algorithm based on structural clustering. *Neurocomputing, 207*, 488–500. doi:10.1016/j.neucom.2016.05.020.

Pang, J., Jia, F., Zhang, C., Zhang, W., Huang, Q., & Yin, B. (2015). Unsupervised web topic detection using a ranked clustering-like pattern across similarity cascades. *IEEE Transactions on Multimedia, 17*(6), 843–853.

Rong, H., Ma, T., Tang, M., & Cao, J. (2018). A novel subgraph k^+-isomorphism method in social network based on graph similarity detection. *Soft Computing, 22*(8), 2583–2601.

Sayyadi, H., & Raschid, L. (2013). *A graph analytical approach for topic detection*. ACM.

Shang, Ronghua, Bai, Jing, Jiao, Licheng, et al. (2013). Community detection based on modularity and an improved genetic;algorithm. *Physica A Statistical Mechanics & Its Applications, 392*(5), 1215–1231.

Sharma, J., & Annappa, B. (2017). Community detection using meta-heuristic approach: Bat algorithm variants. In *Ninth international conference on contemporary computing* (pp. 1–7).

Tsung, C. K., Ho, H., Chou, S., Lin, J., & Lee, S. (2017). A spectral clustering approach based on modularity maximization for community detection problem. In *Computer symposium* (pp. 12–17).

Xie, W., Zhu, F., Jiang, J., Lim, E. P., & Wang, K. (2016). Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge & Data Engineering, 28*(8), 2216–2229.

Yan, D., Hua, E., & Hu, B. (2016). An improved single-pass algorithm for chinese microblog topic detection and tracking. *IEEE international congress on big data*.

Yang, L., Lin, H., Lin, Y., & Liu, S. (2016a). Detection and extraction of hot topics on chinese microblogs. *Cognitive Computation, 8*(4), 1–10.

Yang, X., Jin, P., & Chen, X. (2016b). The construction of a kind of chat corpus in chinese word segmentation. In *IEEE / Wic / ACM international conference on web intelligence and intelligent agent technology* (pp. 168–172).

Ye, Y., Du, Y., & Fu, X. (2016). Hot topic extraction based on chinese microblog's features topic model. In *IEEE international conference on cloud computing and big data analysis* (pp. 348–353).

Yeh, J. F., Tan, Y. S., & Lee, C. H. (2016). Topic detection and tracking for conversational content by using conceptual dynamic latent dirichlet allocation. *Neurocomputing, 216*, 310–318.

Yu, R. G., Zhao, M. K., Chang, P., & He, M. W. (2014). Online hot topic detection from web news archive in short terms. In *International conference on fuzzy systems and knowledge discovery* (pp. 919–923).

Zhang, C., Wang, H., Cao, L., Wang, W., & Xu, F. (2016a). A hybrid termcterm relations analysis approach for topic detection. *Knowledge-Based Systems, 93*, 109–120.

Zhang, Y., Ma, T., Cao, J., & Tang, M. (2016b). K-anonymisation of social network by vertex and edge modification. *IJES, 8*(2/3), 206–216. doi:10.1504/IJES.2016.076114.

Zhou, E., Zhong, N., & Li, Y. (2014). Extracting news blog hot topics based on the w2t methodology. *World Wide Web-internet and Web Information Systems, 17*(3), 377–404.