IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 30, 2018, accepted January 8, 2019, date of publication January 31, 2019, date of current version March 12, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2893980

Hot Topic Detection Based on a Refined TF-IDF Algorithm

ZHILIANG ZHU, JIE LIANG, DEYANG LI, HAI YU, AND GUOQI LIU
Software College, Northeastern University, Shenyang 110169, China

Corresponding author: Hai Yu (yuhai@126.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61374178, Grant 61402092, and Grant 61603082, and in part by the Online Education Research Fund of MOE Research Center for Online Education, China (Qtone Education), under Grant 2016ZD306.

ABSTRACT In this paper, we propose a refined term frequency inversed document frequency (TF-IDF) algorithm called TA TF-IDF to find hot terms, based on time distribution information and user attention. We also put forward a method to generate new terms and combined terms, which are split by the Chinese word segmentation algorithm. Then, we extract hot news according to the hot terms, grouping them into K-means clusters so as to realize the detection of hot topics in news. The experimental results indicated that our method based on the refined TF-IDF algorithm can find hot topics effectively.

INDEX TERMS Feature extraction, hot topic detection, hot terms, time sensitive, user attention.

I. INTRODUCTION

According to the latest report released by the China Internet Network Information Center (CNNIC), the number of Internet users has increased to 751 million in 2017 and the Internet penetration rate is 54.3%. Among them, the number of network news users have reached 69.4%. Network news, one of the earliest Internet applications in the development of the Internet, has become an important platform for people to focus on social trends and to express their views. Compared with traditional media, such as newspapers and broadcasting, network news based on multimedia platform can offer more timely dissemination of social dynamics.

With the great attention and widespread dissemination of network news, the related hot terms will be spread rapidly, and will then become the key words for Internet users to search news of hot spots. The network hot terms have powerful influence, because they condense the inner thoughts of the contemporary people and have a strong, realistic, and critical spirit behind their simple and unconventional appearance. Therefore, with the large amount of public opinion, the network hot terms have become the basis of public opinion analysis and news hot spot discovery.

Despite a few recent advances, hot terms and hot topics detection remains challenging for at least four reasons. First, many hot words are new terms. But the existing word segmentation systems could not recognize new terms in document content. Therefore, it is very important to restore the terms that are wrongly separated by the word segmentation system. Second, TF-IDF algorithm is commonly used for text feature selection. However, for hot terms, the number of news containing hot terms is very large, leading to a lower IDF value. And the weight value could not truly reflect the importance of the hot terms. Third, there is not yet an efficient method to find hot terms from lots of news articles. Fourth, traditional hot topic detection techniques mostly use clustering algorithm to cluster the processed document content. However, the existing clustering algorithms perform poorly and are particularly susceptible to outliers. The large number of non-hot news in the news collection leads to a large number of outliers in the clustering process. Therefore, the effect of directly using the clustering algorithm on all news data sets is not satisfactory.

In our work, we propose a series of efficient methods to discover hot terms and hot topics. Firstly, we propose a Chinese new words discovery method. We define new terms as combined terms, which should not be split in word segmentation system. To restore these separated terms, we use the location information and term frequency co-occurrence of new terms to find new terms. Then, we propose a refined TF-IDF algorithm called TA TF-IDF to improve the defects of TF-IDF in calculating the weights of hot terms. To modify

The associate editor coordinating the review of this manuscript and approving it for publication was Soon Xin Ng.

the original weight value, we incorporate term's importance distribution information over time and user attention into the algorithm. Meanwhile, we define TA as heat index value, and use new feature selection algorithm and heat index value to find hot terms. Finally, to eliminate the influence of outliers on clustering algorithm, we propose a hot news filtering method, which can get better news data sets by excluding news without hot terms. On this basis, an efficient hot topic detection method is proposed.

In summary, this paper provides the following contributions:

- We propose a method to recognize new terms in network news, which is lost from the word segmentation algorithm.
- We utilize the term's time distribution information and users' attention to propose a refined TF-IDF algorithm called TA TF-IDF.
- We provide a set of coherent methods to discover hot terms in network news.
- Based on the extracted hot terms, candidate hot news is identified and then grouped into clusters that represent hot topics by using the vector space model.

This paper is organized as follows. First section introduces the background and importance of our research. Second section briefly introduces related work. Third section presents our method of detecting hot topics. Fourth section describes the experimental process and analyzes experimental results. Final section concludes the paper and proposes further research.

II. RELATED WORKS

A. TOPIC DETECTION AND TRACKING

The research on hot terms discovery originates from the topic detection and tracking (TDT) technology. TDT was first studied by scholars in 1996. Its goal was making new detection and tracking within streams of broadcast news stories [1]. Later, many research directions have arose from it, such as research in the mining and the analysis of public opinion on emergent Internet events, hot topics tracking of online public opinion, and the detection and discovery of hot events. For hot topic extraction, however, terms that appear in a large number of documents in a corpus must be identified. To do this, Bun and Ishizuka [2] proposed a different term weighting scheme TF-Proportional Document Frequency (TF-PDF), which assigns greater weights to terms that appear frequently in many documents from many channels and lower weights to those that are rarely mentioned. Although TF-PDF captures the basic concept of a hot topic, its weakness is that it does not consider variations in the popularity of a topic over time. Guo *et al.* [3] proposed a frequent pattern stream mining algorithm (i.e. FP-stream) to detect hot topics from Twitter streams. Connell *et al.* [4] proposed a method that includes two steps: bounded 1-NN for event formation and bounded agglomerative clustering for building the hierarchy. Liu *et al.* [5] developed a method based on the local expansion for topic detection (LETD), which has two major steps: (1) They first found important URLs that are most likely to describe real-life topics, (2) and then starting from these URLs, they used a local expansion method to seek out other topic-related URLs. Zheng and Li contributed a method to identify hot topics in Microblog based on the topic words, which consists of two main step. Firstly, they extracted topic words in the Microblog data according to the two factors of increasing rate of word frequency and relative word frequency from Microblog data in every time-window. Secondly, they extracted and clustered the topic words according to the similarity among them, sieving for a suitable cluster of topic words so as to describe the hot topic [6]. Lu textitet al. [7] proposed a method based on clustering analysis technique to explore health-related hot topics in online health communities. Zheng and Li [8] proposed a method that can find the hot topics in any time period based on aging theory. You *et al.* [9] used the back-propagation neural network-(BPNN) based classification algorithm to judge the hotness of topic according to its popularity, its quality as well as its message distribution over time. Li and Wei [10] proposed a method of bursty hot topic detection based on bursty feature.

In general, the existing approaches for topic detection can be divided into two categories: The first category includes machine-learning-based methods. These methods are focused on exploiting clustering techniques for topic and event detections. The second category contains the content-based methods, in which group web pages with similar contents as topics using techniques such as natural language processing. However, research on directly finding hot spots based on text feature selection has not been performed.

B. TEXT FEATURE SELECTION

To analyze the mass of news data more effectively, it is a prerequisite to use the appropriate feature information to represent the content of a piece of news. Therefore, feature selection is particularly important for the text information. Moreover, feature selection is an important data-preprocessing technique in bioinformatics and signal processing. Some new feature selection methods, such as multi-objective particle swarm optimization approach for cost-based feature selection in classification [11], and a return-cost-based binary firefly algorithm [12], are effective to solve feature selection problems. At present, scholars at home and abroad have carried out a significant number of research and exploration in text feature selection and keyword extraction. In general, keyword extraction methods can be divided into three main categories: statistical-based methods, machine-learning-based methods, and semantic-based methods.

The machine-learning method is based on a large amount of corpus information, based on the use of keywords to build the model, through training to obtain the corresponding parameters, and finally using these training parameters and models to extract keywords from the text. Zhang *et al.* [13] proposed a keyword extraction method based on the support

vector machine model, which requires the use of the corpus with keywords to train the text content.

Semantic-based methods analyze text from the relationship between words and the meaning of the word itself. The meaning of a word depends on the emotional tendencies of the text, context information, and text category information. Liu *et al.* [14] proposed a key phrase extraction algorithm based on lexical chains, by constructing lexical chains to express multiple narrative clues of the article. Subsequently, they selected strong chains that have rich topic information to express the main message of the article. Finally, they selected phases from the strong chains that can fully express the strong chain from different sides as key phrases.

The statistic-based method refers to the use of statistical information in the document to guide the extraction of keywords. Considering that network news generally has a certain structure, many scholars have improved the algorithm. Yuan *et al.* [15] utilized a variety of features including statistical features, location features, and part of speech features to evaluate the weight of candidate keywords. Some scholars have also improved the algorithm from the perspective of news text category [16], [17]. Xu *et al.* [17] were convinced that the news from each of the categories have some proper nouns that appear frequently in the document, but are not meaningful. They proposed a refined TF-IDF algorithm based on channel distribution information between the IDF values in multiple channels. However, it is not possible to generate hot terms directly from these improved algorithms.

III. MATERIALS AND METHODS

We have already observed that simply considering the term frequency and the inverse document frequency is insufficient for hot topic detection. Since terms or words are the basic elements of any news report, changes in the content of reports will be reflected by variations in a term's usage. Moreover, a topic is composed of many related news reports, changes in a topic's popularity are accompanied by variant usage of key terms or "hot terms." Therefore, we propose a novel approach for recognizing hot terms in order to accurately identify hot topics, which has four major parts:

- We first propose a method to recognize new words, which is separated by the word segmentation algorithm;
- We then exploit the term's time distribution information and users' attention to propose an algorithm to improve the IDF value;
- Next, we present a set of methods on finding hot terms;
- Based on the hot terms, we can get the filtered candidate hot news, and put them into clusters to get hot topics.

A. A METHOD TO RECOGNIZE NEW TERMS IN NETWORK NEWS

Network news, as an Internet media, spread the latest important events daily, resulting in an endless stream of new words. However, as the foundation of Chinese natural language processing, the word segmentation system for Chinese words could not accurately identify these new terms. These new terms always represent important meanings, and even the public's attitude towards hot events. Therefore, it is necessary to rebuild such words to obtain a more meaningful and complete expression of words. For example, "circle of friends" is a term in Chinese, which appears multiple times in the text. However, the Chinese word segmentation system splits it into two terms: "friend" and "circle". Obviously, a combination of words expresses a more abundant meaning. Further, we will lose the term that could represent significant meanings for network news after segmentation processing. Undoubtedly, this will directly affect the results of discovering hot terms of network news. The purpose of new word discovery is to restore the hot terms with actual public opinion. After much observation, we found that these split terms generally have the following characteristics:

- Terms are near each other and appear both simultaneously and consistently.
- Very few other words match with split terms.

Based on the findings above, an algorithm for generating new terms is proposed as follows:

Algorithm 1 An Algorithm for Generating New Terms

Input: W, a set of Chinese work segmentation results;
Output: NW, a new set of Chinese words;

```
for each w_i ∈ W do
    for each w_j ∈ (W − {w_i}) do
        while w_i.freq == w_j.freq do
            for m ← 0 to w_i.indexArray.size do
                while |w_i.indexArray[m] − w_j.indexArray[m]| == w_i.length or w_j.length do
                    count ← count + 1
                end while
            end for
            while count == w_i.freq do
                exchange w_i and w_j
            end while
            w_j ← w_i + w_j;
            w_i ←null;
        end while
    end for
end for
```

B. A NEW CHARACTERISTIC WORD WEIGHING ALGORITHM

The TF-IDF algorithm is widely used for feature selection in text information processing. It is primarily composed of two aspects [18]: (1) term frequency(TF), representing the frequency of occurrence of a feature term in the text set; (2) inversed document frequency (IDF), is a measurement of the general importance of a term, which is offset by the frequency a term appears in the data set [19], [20]. The importance of characteristic words in text concentration will enhance with the enhancement of the word frequency in a

text, but it will be inversely proportional to the word frequency in the whole text concentration [21], which doesn't accord with the purpose of this paper. In this paper, time distribution information and the user attention are considered as two influential factors in hot terms extracting.

1) THE EFFECT OF TIME DISTRIBUTION INFORMATION

Obviously, the frequency that hot terms are used in hot topics vary over time and each hot term has its own life cycle. Using volcanoes as an example, the period when the hot terms are active is like a volcanic eruption. We also call it burst pattern, which has been proved as a useful sensor of real life topics or events [22]. Therefore, we need to identify the active period of hot terms. According to statistical information, the number of news articles where the term T appears has a regular pattern with time distribution. For the event of "Liu Guoliang is not in charge of coaching", the relation between the time and number of occurrences of the news including the term "Liu Guoliang" is shown in Fig.1:

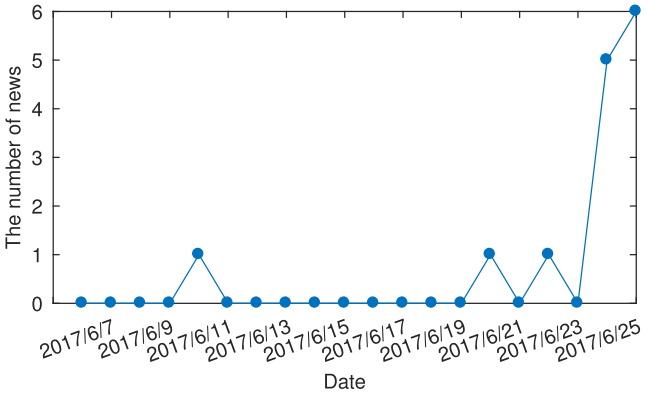

FIGURE 1. The number of news articles that contains the term daily.

As shown in Fig.1, when the term becomes hot, a sharp increase in the number of news articles containing it will occur. The original IDF value could not reflect the importance of the term when it was a hot word. Instead, the value is lower than its real importance value. Therefore, the IDF value needs to be adjusted according to the time distribution information. Subsequently, the refined IDF value can reflect the realistic importance value of the hot terms.

We select TW consecutive days' news data as a data collection D, that is, the date the term T belongs to is the TW^{th} day. Further, the news data set of the i^{th} day is expressed as OD_i. Assume that each hot term has an active period of AD days (and TW is an integer multiple of AD). Divide the data collection D by every AD days, and each AD days' data as a subset, then collection D is represented as $N_{TW} = \{D_i | 1 \leq i \leq TW/AD\}$, dataset $D_i = OD_{AD*(i-1)+1} + OD_{AD*(i-1)+2} + \cdots + OD_{AD*i}$. We then calculate the IDF values on each data set D_i; these subset IDF values are represented as $IDF_{TW} = \{IDF_i | 1 \leq i \leq TW/AD\}$. To explore the relationship between time and the IDF value of the hot terms, data from 2017.6.13–2017.6.24 is shown in Fig.2.

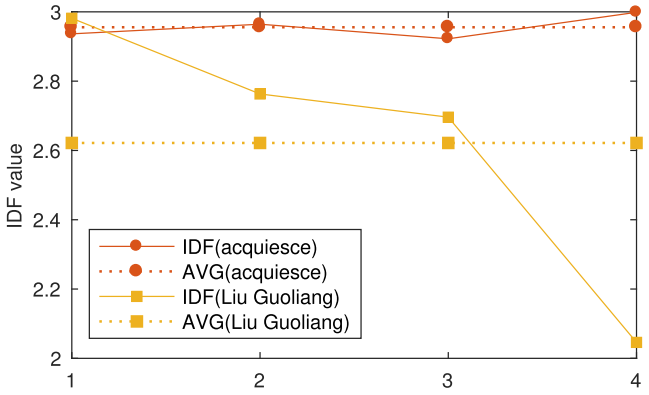

FIGURE 2. The IDF value distribution diagram of each time period.

As shown in Fig.2, the IDF value of the word "acquiesce" was generally higher than the value of "Liu Guoliang". This indicates that "acquiesce" is more important than "Liu Guoliang". However, "Liu Guoliang" is a hot word, whereas the word "acquiesce" is meaningless; therefore, the term "Liu Guoliang" is more important than "acquiesce". Obviously, the IDF values distribution of the term "acquiesce" did not change much. The IDF values distribution of the term "Liu Guoliang" fluctuates, and when the term became a hot word, the IDF value decreased significantly.

The standard deviation is a measure used to quantify the amount of variation or dispersion of a set of data values. Based on these findings, the standard deviation is used to describe the information for a term. Given a set of the subset IDF_{TW} values of term T, the standard deviation is expressed as follows:

$$SD_{IDF}(T) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(IDF_i - \mu)^2} \tag{1}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} IDF_i, \quad n = TW/AD \tag{2}$$

For the terms presented in Fig.2, the rates are shown in Table 1. It can be seen from the table that the importance increases along with its standard deviation value.

TABLE 1. The standard deviation of IDF value for terms in Fig.2.

Term	SD_{IDF}
Liu Guoliang	0.348702125
acquiesce	0.029212786

As shown in Table 1, an obvious disparity occurs between the values of the two terms. However, it is difficult to set a clear boundary to distinguish the standard deviation of hot terms and the meaningless words. Moreover, the standard deviation itself does not have time sensitivity, that is, it could not determine whether the term T is a hot word in the current

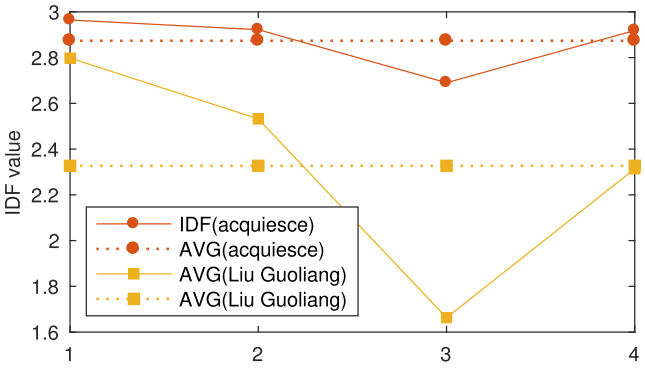

FIGURE 3. The IDF value distribution of each time period.

TABLE 2. The standard deviation of IDF value for terms in Fig.3.

Term	SD_{IDF}
Liu Guoliang	0.419880586
acquiesce	0.10752768

time period. To better explain this scenario, we present Fig.3, where the data is from 2017.6.17 to 2017.6.28 (and "Liu Guoliang" is not a hot word on 2017.6.28). The standard deviation of the subset IDF_{TW} values are shown in Table 2.

By comparing the data of Tables 1 and 2, we can see that the standard deviation of "Liu Guoliang" is greater than the word "acquiesce". However, the standard deviation is still large when the word "Liu Guoliang" is not a hot word. Therefore, only the standard deviation could not identify the hot terms at a specific time.

To obtain the difference in IDF values between a particular time period in which the TW^{th} day belongs to and the other ones, we made a comparison between the IDF_{TW} value sets, in which one includes the particular time period and the other does not. Without a particular time period, the subset IDF value is represented as $IDF'_{TW} = \{IDF_i | 1 \le i \le TW/AD - 1\}$, and its standard deviation value is expressed as follows:

$$SD'_{IDF}(T) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}(IDF_i - \mu')^2} \tag{3}$$

$$\mu' = \frac{1}{n-1}\sum_{i=1}^{n-1} IDF_i, \quad n = TW/AD \tag{4}$$

The ratio, dividing $SD_{IDF}(T)$ by $SD'_{IDF}(T)$, is used to identify the hot word at the correct time period, and is expressed as follows:

$$adjust_{Time}(T) = \frac{SD_{IDF}(T)}{SD'_{IDF}(T)} \tag{5}$$

To make the value of multiple terms more meaningful when compared, the $adjust_{Time}$ value is normalized.

Algorithm 2 The Algorithm to Compute $Adjust_{Time}$ Value

Input: term T, news data collection D, TW, AD;
Output: $Adjust_{Time}$ value of term T;

$numOfTimePeriod \leftarrow TW/AD$
for $i \leftarrow 0$ to $numOfTimePeriod$ **do**
  $sum_{IDF} \leftarrow sum_{IDF} + IDF_i$
end for
$\mu \leftarrow sum_{IDF}/numOfTimePeriod$
for $i \leftarrow 0$ to $numOfTimePeriod$ **do**
  $sumOfSD_{IDF} \leftarrow sumOfSD_{IDF} + (IDF_i - \mu)^2$
end for
$SD_{IDF} \leftarrow \sqrt{sumOfSD_{IDF}/numOfTimePeriod}$
for $i \leftarrow 0$ to $numOfTimePeriod$ **do**
  $sumOfSD'_{IDF} \leftarrow sumOfSD'_{IDF} + (IDF_i - \mu')^2$
end for
$SD'_{IDF} \leftarrow \sqrt{sumOfSD'_{IDF}/(numOfTimePeriod - 1)}$
$adjust_{Time} \leftarrow SD_{IDF}/SD'_{IDF}$
$Adjust_{Time} \leftarrow getNormalization(adjust_{Time})$
return $Adjust_{Time}$

The normalized value is expressed as follows:

$$A\mathit{djust}_{Time}(T) = \begin{cases} adjust_{Time}(T), & 0 \le adjust_{Time}(T) \le 1 \\ \dfrac{\log(adjust_{Time}(T))}{\log(limit\_Value)} & \\ *(\beta - 1) + 1, & 1 < adjust_{Time}(T) < limit\_Value \\ \beta, & adjust_{Time}(T) \ge limit\_Value \end{cases} \tag{6}$$

The algorithm of $Adjust_{Time}$ is shown as follows:

For a hot word, the $Adjust_{Time}$ value could reflect time sensitivity. Among the subset IDF values, if the last one is far away from the mean, the term will be assigned with a much higher $Adjust_{Time}$ value.

2) THE EFFECT OF USER ATTENTION

From the sociological point of view, the emergence of hot topics is closely related to the choice of group behavior and public attention [23]. Therefore, it is necessary to consider the influence of users' attention in hot terms research. The user attention to news is reflected in many aspects. The most obvious aspect is the number of news hits and the number of user participation, such as news comments. Many people click on a piece of news that they are not really interested in, or click on news based on their speculative title that does not match the content. In contrast, the amount of user participation, which can express the user's real emotional inclination, is more suitable to reflect the user's attention to the news. Further, the number of news hits and the number of news comments are positively related such that the number of user participation on the news is selected to measure user attention to the news. News attention is expressed as follows:

$$Attention(News) = Participants \tag{7}$$

For a term, the higher the user attention of related news, the more likely the item becomes a hot word. Therefore, the user attention of term T is expressed as the average of the user attention of all the news containing this term. Assuming that the number of news articles containing the term T is n, the news dataset is $D_T = \{d_1, d_2, \ldots, d_n\}$. Therefore, the user attention of the term T is expressed as follows:

$$Attention_{AVG}(T) = \sum_{i=1}^{n} Attention_i / n \tag{8}$$

where $Attention_i$ is the user attention of news d_i.

The $Attention_{AVG}$ value and the original IDF value are not in an order of magnitude, so it is difficult to adjust the original value directly. The $Attention_{AVG}$ value is normalized as follows:

$$Adjust_{Attention}(T) = \begin{cases} 0, & Attention_{AVG}(T) = 0 \\ \dfrac{\log(Attention_{AVG}(T))}{\log(limit\_Value)} * \beta, & 0 < Attention_{AVG}(T) \\ & < limit\_Value \\ \beta, & Attention_{AVG}(T) \\ & \geq limit\_Value \end{cases} \tag{9}$$

3) THE TA TF-IDF ALGORITHM

Based on the research work above, the $Adjust_{Time}$ value and the $Adjust_{Attention}$ value are positively related to the importance of hot terms; therefore, the adjustment method is expressed as follows:

$$Adjust_{IDF}(T) = \alpha * Adjust_{Time}(T) + (1 - \alpha) * Adjust_{Attention}(T) \tag{10}$$

where α is an adjustment parameter used to adjust the influence of the distribution of time and the distribution of user attention on the importance of the terms. This adjustment is done to prevent the numerical value of one aspect from being too large to submerge the numerical value on the other. Accordingly, the improved IDF value is expressed as follows:

$$TA - IDF(T) = Adjust_{IDF}(T) * IDF(T) \tag{11}$$

Further, the refined $TF - IDF$ algorithm is expressed as follows:

$$TA\ TF - IDF(T) = TF(T) * TA - IDF(T) \tag{12}$$

C. HOT TERMS RECOGNITION WITH TA TF-IDF

In this part, the TA TF-IDF algorithm is used to find hot terms from the daily network news. To be more specific, the method has three major steps:

1) The improved algorithm is used to calculate the weight value of terms.
2) For a piece of news d_i, take the weight of top p terms as key words to represent the news, and choose them as candidate hot terms, represented as $F(d_i) = \{T_{i1}, T_{i2}, \ldots, T_{ip}\}$. Then, keywords for all news in one day is expressed as $FD = \{F(d_1), F(d_2), \ldots, F(d_n)\}$, where n is the number of news articles in a day.
3) All these candidate hot terms are sorted by the $Adjust_{IDF}$ value, the first k is the final hot terms, that is, $HW = \{T_1, T_2, \ldots, T_k\}$.

The specific procedures of experiments are shown in Fig.4.

D. EXTRACTING HOT TOPICS BASED ON HOT TERMS

Based on the extracted hot terms, related hot news can be filtered from news corpus. And then, we use the vector space model to represent all these filtered news, and group these news into KMeans clusters to get hot topics.

The vector space model is widely used to compare the similarity of two documents in the field of TDT, particularly in New Event Detection and Topic Tracking. However, its limitations are obvious when we set a higher precision or recall rate for TDT work. To get more accurate similarities of documents, we should provide more features to represent a document, that will lead to a high dimensional sparse matrix. Unfortunately, it will take up more space and cost more time to compute cosine similarity of two documents. On the contrary, using less features will make it hard to compare the similarity of documents when there are insufficient keywords in the documents being compared.

The shortcomings of the simple vector space model approach suggest that there is a need for target documents represented by more accurate features. Our method can overcome this shortcoming based on two measures. First, news grouped into clusters are filtered based on extracted hot terms. As the number of news decreases, the number of features in the vector space model will also decreases. Second, hot terms have strong representativeness for news content. As the weight of hot terms based on TA TF-IDF increase, hot terms are more likely to be retained, and we can use fewer features to represent news content. Therefore, we can construct a lower dimensional matrix to model the news.

And besides, there is a certain number of news that has nothing to do with hot topics in original news corpus. When we group all news into clusters, we will get a bad clustering results because of the number of outliers. Applying our method to extract hot news, we will get a better clustering results based on the filtered news.

IV. RESULTS AND DISCUSSION

A. EXPERIMENTAL DATASETS

In terms of data collection, NetEase news from August 2017 to November 2017 were collected by a crawler, a pure text collection that contains about 20000 news articles. Each piece of news consists of its title, news content, release time, and the number of user participants. Among them, the number of participants includes the number of news comments, the number of praise points, and the number of objections on comments. For each piece of news, the NLPIR segmentation system provided by the Chinese Academy of Sciences is used to obtain the results of word segmentation and word frequency statistics.

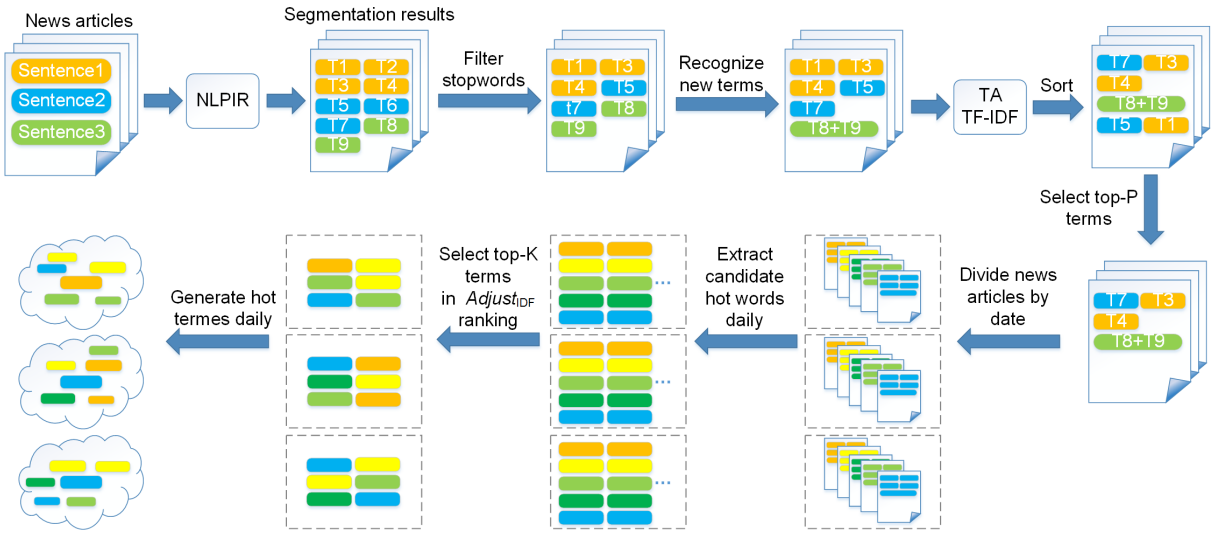

FIGURE 4. The process of generating hot terms.

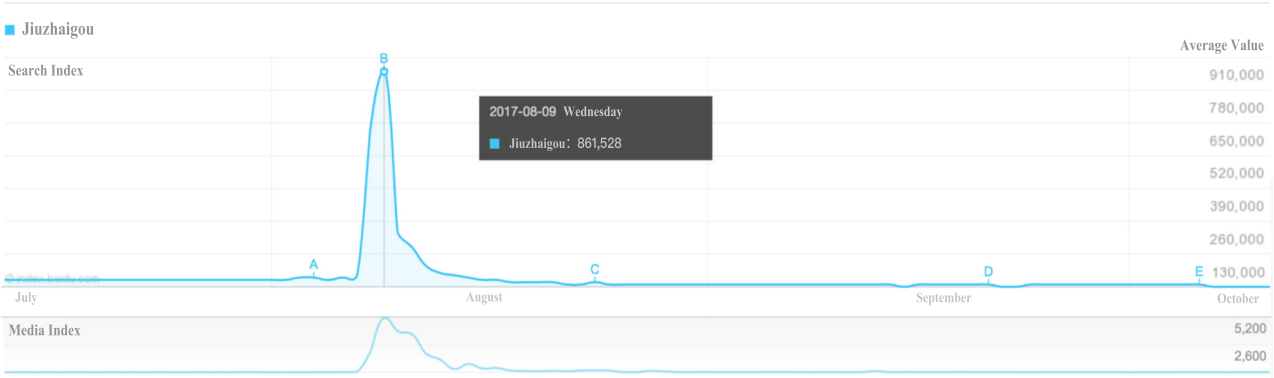

FIGURE 5. The Baidu Index for hot term "Jiuzhaigou".

B. EXPERIMENTAL VERIFICATION METHOD

The Baidu index is a data sharing platform based on Baidu's massive Internet users' behavior data, in which the search index and media index are included. Search index, based on the search volume of Internet users in Baidu, shows the scientific analysis results of the weighted sum of search frequency in the Baidu web search for each word. The media index shows the number of news articles related to keywords, which are reported by the major Internet media and included in Baidu news.

In conclusion, the media index reflects the media's attention on keywords, while the search index reflects the users' attention on the keywords from the user's perspective. Therefore, in this experiment, these two indices are combined to verify whether the experimental result is a hot word and then the experimental accuracy is calculated. For example, Fig 5 shows the item "Jiuzhaigou" is a hot word on August 9.

C. THE limit_Value OF $Adjust_{Time}$ AND $Adjust_{Attention}$

The range of $Adjust_{Attention}$ values is wide: the lowest is 0, and the highest is approximately 500 000. To observe more intuitively the distribution of the terms' $Attention_{AVG}$ value, Fig 6 shows the distribution of the terms' $Attention_{AVG}$ value from August 5, 2017 to September 6, 2017. Terms whose $Attention_{AVG}$ value is larger than 100 000 are not included in Fig 6, because the number of such terms is few.

Fig.6 shows that terms' attention value concentrated mostly in between 0 to 20000. According to the statistics, terms in this range accounted for about 95% of the terms. Therefore, we assign 20 000 to *limit_Value* in Eq 10. Similarly, we assign 1000 to *limit_Value* in Eq 7.

D. THE DISCUSSION OF EXPERIMENTS IN DISCOVERING HOT TERMS

Based on the work above, we conduct a detailed experimental work.

1) IMPACT OF $Adjust_{Time}$, $Adjust_{Attention}$ AND $Adjust_{IDF}$ ON ACCURACY

In our system, we propose to utilize two different yet relevant types of information to generate hot terms: time distribution information and user attention. Prior studies in the field often

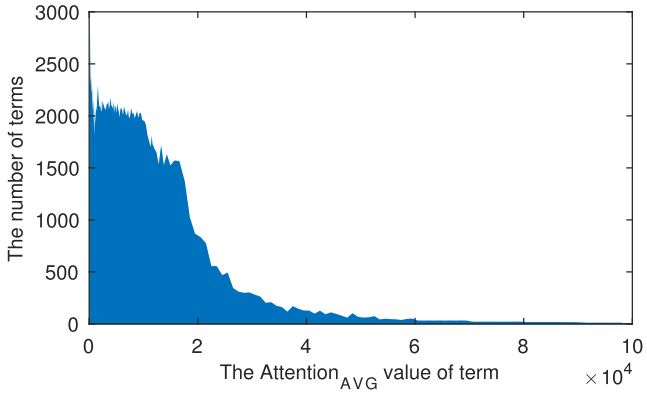

FIGURE 6. The distribution of the terms' $Attention_{AVG}$ value.

use attention to detect hot events. However, based on our analysis in Section 3, hot terms are related to time distribution information and user attention. To evaluate the performance of our hybrid approach, we test the effect of only one factor to the results. Specifically, we randomly picked 20 days' worth of news from our news dataset, and generate hot terms (top 5, top 7, and top 10) on different factors. For comparison, we compute the averaged accuracy of the results. Fig.7 depicts the result.

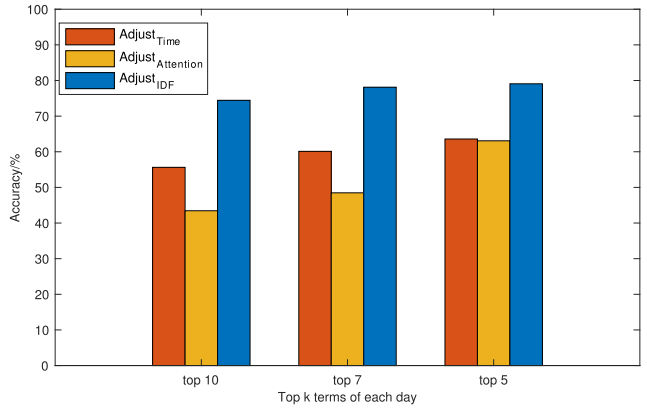

FIGURE 7. Impact of $Adjust_{Time}$, $Adjust_{Attention}$ and $Adjust_{IDF}$ on accuracy.

From the comparison, we observed that: (i) our hybrid approach that considers both aspects tends to perform best, (ii) the generated hot terms based purely on one single aspect cannot outperform the system based on the combinations. The straightforward reason is that more intrinsic properties of the hot terms can be revealed as we take more aspects into account, and (iii) for one factor's performance, the results based on time distribution information perform better than the one with user attention.

2) IMPACT OF THE α VALUE IN EQ 11 ON ACCURACY

To find the best combination state of the time distribution information and the distribution of user attention by expressing the importance of the terms (computing terms' refined IDF value) and selecting hot terms from candidate terms, we use an adjusted value of α to improve the performance of these two factors, which is expressed in Eq 11. To find an optimal value for α, we adopt the same experimental setup with the procedure above. As we change the value of α by traverse from 0 to 1 in a 0.1 step, the averaged accuracy of generating hot terms is observed. The tuning result is shown in Fig.8. Fig.8 shows that for the top 5, 7, and 10 selected hot terms, when α is 0.4, the accuracy reaches the peak. That is, the user attention has a little more influence on the selected top words results. However, the difference is small between the ratios of these two factors. Therefore, we believe that the contributions of these two factors are equivalent.

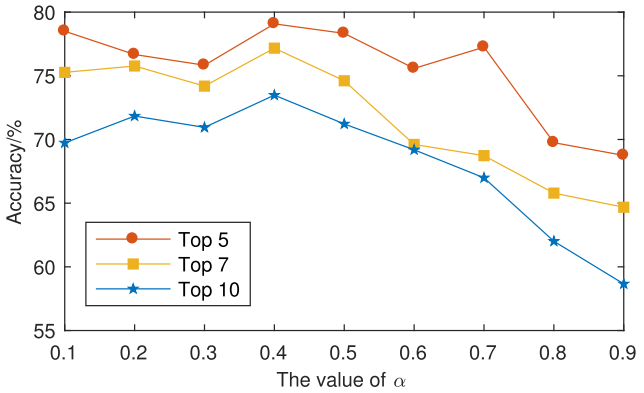

FIGURE 8. Impact of the α value in Eq 11 on accuracy.

3) IMPACT OF THE β VALUE ON ACCURACY

In our system, the $Adjust_{IDF}$ value is used to adjust the original IDF value, such that its value is related to the terms' original IDF value and the importance of the terms in the news. Based on our analysis in Section 3, we assume that the range of value for $Adjust_{IDF}$ is 0 to β. Prior related studies use up to 10 terms to represent a piece of news in feature selection; hence, we believe that the top 10 terms in each piece of news can potentially become hot terms. For each piece of news, when terms are arranged in the ascending order according to the original IDF value, we found that the original IDF value of the first term is about 10 times at most of the 10^{th} value. That is, hot terms at the bottom of the original IDF value multiplied by 10 can improve their ranking. Therefore, we assume that the value of β is between 2 to 10. To find the optimal value for β, a series of experiments has been performed to compare the efficiency. The tuning result is shown in Fig.9.

As depicted, for the top 5, 7, and 10 selected hot terms, there is very little difference between the maximum and the minimum in accuracy. The rate differences are 2.41%, 3.43%, and 2.57%, respectively. Therefore, the change in the β value has no significant influence on the accuracy.

The experimental results reflected that using the improved algorithm to calculate the weight value of all news feature words for one day takes about 431 seconds, including 5393 feature words from 299 news articles.

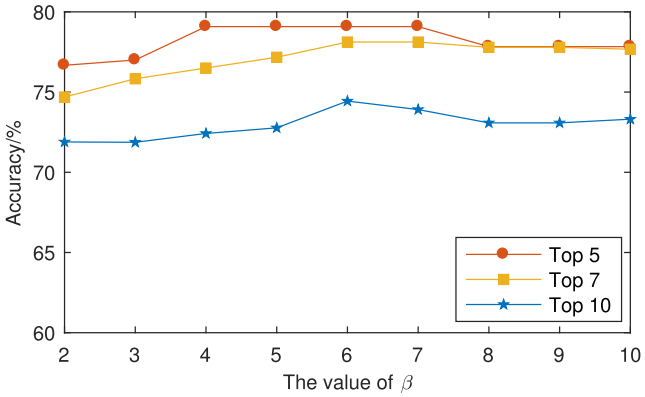

FIGURE 9. Impact of the β value on accuracy.

4) IMPACT OF TOP p TERMS OF EACH NEWS ON ACCURACY

Prior related studies often use top p terms to represent a piece of news in feature selection, where the value of p is 3-10. Our goal is to generate hot terms, therefore we assume that the value of p is 1 to 10 when getting candidate terms from each news. As we change the value of p, the averaged accuracy of generating hot terms is observed. The tuning result is shown in Fig.10.

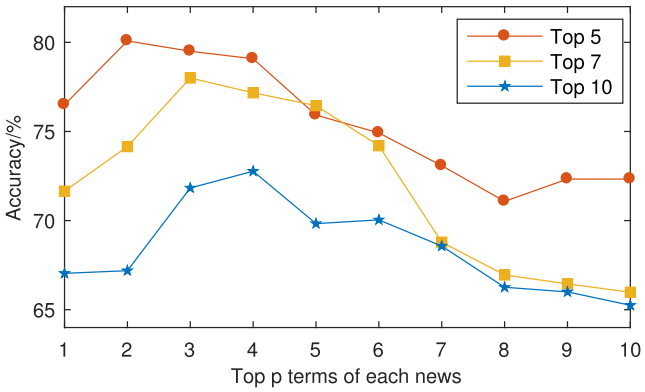

FIGURE 10. Impact of top p terms of each news on accuracy.

Fig.10 shows that for the top 5, 7, and 10 selected hot terms, the overall accuracy is first increased and then decreased along with the increasing of p. When the value of p is 4, the accuracy drops significantly. When p is between 2 to 4, the accuracy performs well. However, this value may change with different word segmentation algorithms and different data sets. We need to adjust it according to the actual conditions.

In conclusion, we randomly picked 20 days' worth of news from our news dataset, and the average accuracy is 78.36%.

E. THE RESULTS OF EXPERIMENTS IN DISCOVERING HOT TERMS

We randomly picked several days' results of hot terms to draw a word cloud picture. Five main steps are included:

1) Select top k terms in $Adjust_{IDF}$ ranking for each day, represented as $HW = \{T_i|0 < i \leq k\}$, where i represents the ranking of term T_i, and the value of k is 10 in our experiment.
2) For terms in HW, assign weighting coefficient for each term based on its ranking. For term T_i, its weighting coefficient is $W(T_i) = 1 - i * 0.1$.
3) For each term, multiplying its $Adjust_{IDF}$ value by its weighting coefficient is its *Score* value, $Score(T_i) = Adjust_{IDF}(T_i) * W(T_i)$.
4) For all terms, eliminate duplicated terms and the new *Score* value of the repeated term is the average of all its old *Score* values. Meanwhile, record the repeated number of times for each term, represented as *Repeat_Times*. That is, *Repeat_Times* value is the number of times term T appears minus 1. For instance, the *Score* value of the term T is 1 if it appears twice.
5) For all terms, sort them by score value. Then, give the *size* value to each term in the use of word cloud. For each term, enlarge its *size* value according to its *Repeat_Times*.

After the steps above have been performed, we draw a piece of Word Cloud figure to display the detected hot terms, and the results are shown in Fig.11.

FIGURE 11. Word cloud.

From the results in Fig.11, we conclude that our generated hot terms represent some social hot topics during a time period. A total of 12 topics are explicitly exposed in Table 3:

Table 5 illustrates the hot topics related to generated hot terms. Some hot spots of society such as "a magnitude 7.0 earthquake struck Jiuzhaigou County," "college students were deeply trapped in pyramid selling organization," and "the anti-corruption campaign is sweeping the country" in August 2017 have caused widely concern in China. Therefore, the hot terms generated by our method are accurate and effective. Users can search the generated hot terms to view the web page to explore more about hot events.

F. THE RESULTS OF HOT TOPIC DETECTION

Reference [6] contributes a method, which aims to identify hot topics in Microblog based on the topic words. And we proposed a TA TF-IDF algorithm, which aims to identify

TABLE 3. The related hot topics about the generated hot terms.

No	hot terms	Hot Topic
1	-Li Wenxing -Zhongxiang City -Lin Huarong	College students were deeply trapped in pyramid selling organization.
2	-Terminal High Altitude Area Defense	South Korea deployed Terminal High Altitude Area Defense System.
3	-Jiuzhaigou -Paradise hotel -Red Cross Society of Sichuan -Seismological Bureau -Nuorilang Waterfall -Jinghe County	A magnitude 7.0 earthquake struck Jiuzhaigou County, Sichuan Province.
4	-Chen Lan -Girl -Li Bingxin -Nanjing	A girl was obscene in public in the waiting room at Nanjing Railway Station.
5	-Han Chunyu	The experiments of Han Chunyu's paper could not be reproduced.
6	-Yulin City -Cesarean section -Suide County -Pregnant women -Labour room -Natural childbirth	A pregnant woman jumped from hospital buildings to her death.
7	-Bengbu County	Case of girl's disappearance in Bengbu.
8	-Heshun County	Coal mine landslide accident in Heshun County.
9	-black bear -Badaling Zoo -Badaling	A black bear burt people in Badaling Zoo.
10	-Zhu Xiaozhen	The babysitter killed the female employer and her children.
11	-Leng Mengmei	Chinese students studying abroad was killed abroad.
12	-Xie Yang -Xie Chao	The anti-corruption campaign is sweeping the country.

TABLE 4. Experimental results of different methods.

Method	Accurate
Method based on TA TF-IDF	84%
Method in Ref. [6]	72%

TABLE 5. Experimental results of different methods.

Date	TA TF-IDF	Method in Ref. [6]	Total number of news
August 04, 2017	42	106	281
August 05, 2017	22	138	248
August 08, 2017	41	301	373
August 10, 2017	31	196	332
August 15, 2017	22	183	299
August 22, 2017	14	240	297
September 05, 2017	21	69	226

hot topics based on the hot terms. The two are similar in purpose and have certain comparability. Therefore, we select the literature [6] as a comparison algorithm.

According to the hot topics provided by the news, we compare our method with that used in [6]. This paper contributes a method, which aims to identify hot topics based on the topic words. This method, through dividing the time-window, extracts topic words according to the two factors of increasing rate of word frequency and relative word frequency from data in every time-window. When dealing with the increasing rate of word frequency, it thought that when a word appears frequently at a certain length of time, and its frequency is increasing obviously compared with the one in the former time-window; or it is obviously decreasing compared with the one in the future time-window, it means, to a certain extent, this word has relation with some fairly new topic of the news. And the length of the time-window is k hours. It has two weakness. Firstly, the length of the time-window is too small, the result in such time period has some contingency, and it

could only identify the hot terms when them appear or disappear. Secondly, our method just need former information, and it can identify the hot terms in real time. However, the method in [6] not only need former information but also need future information, it could only identify the hot terms in the past. Therefore, it has poor applicability and effectiveness.

We use the feature selection method in [6] to replace our TA TF-IDF algorithm to compute the weight value of terms in order to extract hot terms. Results of hot topic detection are presented in Table 4, and a detailed description of the number of filtered news is given in Table 5. It shows the number of news grouped into clusters by using two methods. In the table, we can find that our method can effectively filter non-hot news, while the topic terms based on the method in [6] are inaccurate for hot topic detection. Table 4 shows that all two methods can be used to find hot topics in our experiment dataset effectively and that our method has a higher accuracy rate.

V. CONCLUSIONS

In this paper, an expeditious and efficient hot terms and topics detection method is proposed. We discover new words by location information and word frequency co-occurrence. And the TA TF-IDF is proposed to overcome the shortcomings of traditional methods in explaining the importance of hot terms by time distribution information and user attention. Then, based on the refined TF-IDF and TA, we proposed a novel hot terms detection method, which can find hot terms rapidly. And experimental results showed that the average accuracy of our hot terms finding method can reach 78.36%. Finally, we proposed a hot topics detection method based on the filtering method with hot terms. In addition, to test the performance of the whole system, an existing approach in [6] was implemented for comparison. An extensive evaluation demonstrated that the efficiency of our system is better.

For the future work, we plan to explore hot topics from various perspectives to improve the universality of our approach. This is because that our current approach requires continuous news data to deal with the timeliness and the data requirements is a bit of stringent.

DATA AVAILABILITY

The experimental data used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report final report," in *Proc. Darpa Broadcast News Transcription Understand. Workshop*, 1998, pp. 194–218. [Online]. Available: http://repository.cmu.edu/compsci/341

[2] K. K. Bun and M. Ishizuka, "Topic extraction from news archive using TF∗PDF algorithm," in *Proc. 3rd Int. Conf. Web Inf. Syst. Eng.*, 2002, pp. 73–82.

[3] J. Guo, P. Zhang, J. Tan, and L. Guo, "Mining hot topics from twitter streams," *Proc. Comput. Sci.*, vol. 9, no. 11, pp. 2008–2011, 2012.

[4] M. Connell, A. Feng, G. Kumaran, H. Raghavan, C. Shah, and J. Allan, "UMass at TDT 2004," in *Proc. Topic Detection Tracking Workshop Rep.*, 2004, pp. 109–115.

[5] H. Liu, J. He, Y. Gu, H. Xiong, and X. Du, "Detecting and tracking topics and events from Web search logs," *ACM Trans. Inf. Syst.*, vol. 30, no. 4, 2012, Art. no. 21. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2382438.2382440

[6] J. Zheng and Y. Li, "A hot topic detection method for chinese microblog based on topic words," in *Proc. 2nd Int. Conf. Inf. Technol. Electron. Commerce*, 2014, pp. 262–266.

[7] Y. Lu, P. Zhang, J. Liu, J. Li, and S. Deng, "Health-related hot topic detection in online communities using text clustering," *PLoS ONE*, vol. 8, no. 2, p. e56221, 2013.

[8] D. Zheng and F. Li, "Hot topic detection on BBS using aging theory," in *Proc. Int. Conf. Web Inf. Syst. Mining*, 2009, pp. 129–138.

[9] L. You, Y. Du, J. Ge, X. Huang, and L. Wu, "BBS based hot topic retrieval using back-propagation neural network," in *Proc. Int. Conf. Natural Lang. Process.*, 2004, pp. 139–148.

[10] H. Li and J. Wei, "Netnews bursty hot topic detection based on bursty features," in *Proc. Int. Conf. E-Bus. E-Government*, 2010, pp. 1437–1440.

[11] Y. Zhang, D.-W. Gong, and J. Cheng, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 1, pp. 64–75, Jan./Feb. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7243331/

[12] Y. Zhang, X. Song, and D. W. Gong, "A return-cost-based binary firefly algorithm for feature selection," *Inf. Sci.*, vols. 418–419, no. 1, pp. 561–574, 2017.

[13] K. Zhang, H. Xu, J. Tang, and J. Li, "Keyword extraction using support vector machine," in *Proc. Int. Conf. Web-Age Inf. Manage.*, 2006, pp. 85–96.

[14] M. Liu, X.-L. Wang, and Y.-C. Liu, "Research of key-phrase extraction based on lexical chain," *Chin. J. Comput.*, vol. 33, no. 7, pp. 1246–1255, 2010.

[15] J. Yuan, X. Mao, "Keyword extraction from chinese news Web pages based on multi-features," *Comput. Eng. Appl.*, vol. 50, no. 19, pp. 222–226, 2014.

[16] Q. Kuang and X. Xu, "Improvement and application of TF-IDF method based on text classification," in *Proc. Int. Conf. Internet Technol. Appl.*, 2010, pp. 1–4.

[17] M. Xu, L. He, and X. Lin, "A refined TF-IDF algorithm based on channel distribution information for Web news feature extraction," in *Proc. 2nd Int. Workshop Educ. Technol. Comput. Sci.*, 2010, pp. 15–19.

[18] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: Computer Science and Information Technology: CS & IT for GATE*. Boston, MA, USA: Addison Wesley, 1999.

[19] K. Church and W. Gale, "Inverse document frequency (IDF): A measure of deviations from Poisson," in *Natural Language Processing Using Very Large Corpora* (Text, Speech and Language Technology), vol. 11. Dordrecht, The Netherlands: Springer, 1999.

[20] H. Chen, D. Han, Y. Dai, and L. Zhao, "Design of automatic extraction algorithm of knowledge points for MOOCs," *Comput. Intell. Neurosci.*, vol. 2015, 2015.

[21] C. Shi, C. Xu, and X. Yang, "Study of TFIDF algorithm," *J. Comput. Appl.*, vol. 29, no. S1, pp. 167–170, 2009.

[22] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," in *Proc. Int. Conf. Very Large Data Bases*, Trondheim, Norway, Aug./Sep. 2005, pp. 181–192. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-33745624002&partnerID=tZOtx3y1

[23] T.-P. Liang and H.-J. Lai, "Discovering user interests from Web browsing behavior: An application to Internet news services," in *Proc. 35th Annu. Hawaii Int. Conf. Syst. Sci.*, 2002, pp. 2718–2727.

ZHILIANG ZHU received the M.S. degree in computer applications and the Ph.D. degree in computer science from Northeastern University, China. He has authored or co-authored over 130 international journal papers and 100 conference papers. In addition, he has published five books, including *Introduction to Communication and Program Designing of Visual Basic .NET*. His research interests include information integration, complexity software systems, network coding and communication security, chaos-based digital communications, applications of complex network theories, and cryptography. He is a Fellow of the China Institute of Communications. He is a Senior Member of the Chinese Institute of Electronics and the Teaching Guiding Committee for Software Engineering under the Ministry of Education. He was a recipient of nine academic awards at national, ministerial, and provincial levels. He has served in different capacities at many international journals and conferences. He currently serves as the Co-Chair of the 1st–9th International Workshop on Chaos–Fractals Theories and Applications.

JIE LIANG is currently pursuing the master's degree with the Software College, Northeastern University, China. Her research interests include information retrieval, data mining, and information processing.

DEYANG LI is currently pursuing the master's degree with the Software College, Northeastern University, China. His research interests include recommendation systems, data mining, and machine learning.

HAI YU received the B.E. degree in electronic engineering from Jilin University, China, in 1993, and the Ph.D. degree in computer software and theory from Northeastern University, China, in 2006, where he is currently an Associate Professor of software engineering. His research interests include complex networks, chaotic encryption, software testing, software refactoring, and software architecture. He has served in different roles at several international conferences, such as the Associate Chair for the 7th IWCFTA, in 2014, the Program Committee Chair for the 4th IWCFTA, in 2010, the Chair of the Best Paper Award Committee at the 9th International Conference for Young Computer Scientists, in 2008, and a Program Committee Member for the 3rd–9th IWCFTA and the 5th Asia–Pacific Workshop on Chaos Control and Synchronization. He was a Lead Guest Editor of *Mathematical Problems in Engineering*, in 2013. He also serves as an Associate Editor for the *International Journal of Bifurcation and Chaos*, as a Guest Editor for *Entropy*, and as a Guest Editor for the *Journal of Applied Analysis and Computation*.

GUOQI LIU received the B.E. degree in computer science and technology from Jilin University, China, in 2002, and the Ph.D. degree in computer application technology from Northeastern University, China, in 2011, where he is currently an Associate Professor of software engineering. His research interests include complex networks, service computing, data warehousing and mining, and software engineering.