Research Article

A Hotspot Information Extraction Hybrid Solution of Online Posts' Textual Data

HuiRu Cao,¹ Xiaomin Li²,³ Songyao Lian,³ and Choujun Zhan⁴

¹Department of Information Engineering, Guangzhou Institute of Technology, Guangzhou 510725, China

²College of Mechanical and Electrical Engineering, Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China

³Nanfeng College, Sun Yat-sen University, Guangzhou 510970, China

⁴School of Computer, South China Normal University, Guangzhou 510631, China

Correspondence should be addressed to Xiaomin Li; lixiaomin@zhku.edu.cn

Received 7 November 2020; Revised 28 March 2021; Accepted 7 April 2021; Published 15 April 2021

Academic Editor: J. R. Méndez

Copyright © 2021 HuiRu Cao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Online posts have gradually become a major carrier of network public opinion in social media, and the social network hotspots are the important basis for the study of network public opinion. Therefore, it is significant to extract hotspots for monitoring Internet public opinion from online posts textual big data. However, the current hotspot extraction methods are focused on the users' features that are based on textual big data with spam and low-quality content. Meanwhile, these methods seldomly consider the time span of posts and the popularity of users. Accordingly, this article presents a hotspots information extraction hybrid solution of online posts' textual data. Firstly, a filtering strategy to obtain more high-quality textual data is designed. Secondly, the topic hot degree is presented by considering the average number of replies and the popularity of the participant. Thirdly, an improved co-word analysis technology is used to search the same topic posts and Bisecting k-means clustering algorithm using repliers' popularity and key posts are designed for studying and monitoring the hotspots of online posts in a valid big data environment. Finally, the proposed algorithms are verified in experiments by extracting the hotspots of online posts from the dataset. The results show that the data filtering strategy can help to obtain more valuable information and decrease the computing time. The results also demonstrate that the proposed solution can help to obtain hotspots comparing the traditional methods, and the hot degree can reflect the trend of the online post by comparing the traditional methods.

1. Introduction

With the rapid development of mobile communications and networks, the Internet increasingly integrates into our life. It is reported that there are now more than 4 billion Internet users around the world. Most Internet users spend an average of six hours surfing the Internet, and 3 billion people now use social media, such as Twitter, blogs, Bulletin Board System (BBS), and podcasts [1, 2]. It is known that online posts have gradually become an important tool in social media for the exchange of information. An increasing amount of public opinion is now spread by social media, especially through BBS [3–5]. Since hotspots directly reflect

public opinion, studying and monitoring the hotspots of social media becomes more important for public affairs.

Social media has become one of the most important and popular carriers and distributors of the current online public opinion [6, 7]. Compared to the traditional public opinion channels, online posts have some unique features, such as a wider audience range, greater influence, faster propagation speed, and large amount of data [8–10]. For obtaining and monitoring public opinion hotspots, an increased number of studies focus on this field from different perspectives. In general, the current studies mainly use natural language processing, data mining technologies, machine learning, and other methods to monitor hotspots and explore propagation [11–14].

Currently, text data is still an important medium for information dissemination on social networks [15, 16]. To study complex dynamics in social networks, the extraction of hotspots from massive textual data becomes one of the important steps. On the one hand, the current hotspots' extraction methods are simple to collect the user-related feature and mostly based on textual big data with spam, irrelevant, and low-quality content. In social media, there are many spam information [17–19], such as paid posters and fake replies, as shown in Figure 1. Advertising posts and replies are a good example in BBS. Such corresponding users' featured information based on invalid or incomplete data can be very different from real one, especially for hotspots and public opinions. On the other hand, the main methods seldomly take into account the time span of posts and popularity of repliers. Firstly, time span of posts is a significant factor in hotspots extraction, as hotspots of social networks are the collective action of users in a short time (for example, a collective response to an event in BBS). Secondly, it is reported that popularity users (repliers and main posters) play a significant role in Internet public opinion [20]. However, few studies address these problems. Due to the complexity and features of social media such as BBS, monitoring of public opinion hotspots still faces the following challenges:

- (1) How to obtain more valuable data by filtering a large amount of spam textual data
- (2) How to find the key posts according to the association among multiple posts for the same topic
- (3) How to search real hotspots by considering the valuable repliers and key posts

Accordingly, for solving the problems, a hotspots' information extraction hybrid solution of online posts' textual data is proposed based on the feature of users in social networks. The solution contains three main steps. Firstly, a textual data filtering strategy is used to obtain a more valid dataset. An improved co-word analysis technology is used to search the same topic posts. Secondly, bisecting k-means clustering algorithm based on poster popularity and key posts are proposed to obtain the hotspots of online posts. Then, the hot degree is proposed to search the real hotspots. The proposed methods are implemented in a real experiment, where the results demonstrate the effectiveness of the solution.

The rest of the paper is structured as follows. Section 2 discusses related studies on the Internet public opinion and current challenges. Section 3 introduces hotspot monitoring and public opinion communication characteristics. Section 4 introduces the cluster hotspot monitoring based on PR values and bisecting k-mean algorithms. Section 4 presents the results of an experiment using the proposed methods and our dataset. Section 5 concludes and discusses the paper.

2. Related Work

In this section, we present existing studies on the public opinion analysis of BBS and monitoring hotspots. These studies are used as a basis for our work. We review related research from two aspects: network public opinion and hotspots.

2.1. Public Opinion. In [21], the natural language processing and machine learning techniques are used to interpret sentimental tendencies related to users' opinions and predict real events. In [22], a public opinion dynamics model for an online-offline social network context is provided and conditions to form a consensus in the proposed model are analyzed. In [23], the authors propose a method to recognize network public opinion leaders by using Markov logic networks, and a recognition system is designed and implemented. In [24], a cross-network public opinion spreading model is created in a combined social network environment. Two network nodes are assumed in this paper. In [25], the author constructs a dictionary monitoring sentiment computing model using text words and labels as the input parameters. In [26], a new method is provided for sentiment computing for news and events by constructing a word emotion association network. The authors provide a word emotion computation method to obtain initial words. These studies mostly focus on public opinion based on the assumption that the dataset is always valid.

2.2. Hotspots. Zhao et al. [27] present a Social Sentiment Sensor (SSS) system on Sina Weibo to detect daily hotspots and analyze sentiment distributions related to these topics. Clusters of topics that describe the same issue are formed and ranked based on popularity to exploit the resulting hotspots. In [28], the authors use a clustering method to obtain candidate topics on BBS and the evolution theory to calculate the heat of candidate topics and obtain hotspots based on it. Hao and Hu [29] propose a method based on a baseline model to solve the topic drift problem of network BBS. Liu and Li [30] adopt text mining approaches based on a vector space model and k-means clustering to group Internet public opinion hotspots. Li [31] uses an emotion analysis technology to analyze the emotional polarity of network BBS Chinese texts and a k-means algorithm and the SVM to cluster the contents of posts considering each class as a hot topic. Chen et al. [32] design a similarity analysis algorithm of Internet public opinion based on information entropy, which can cluster and identify hotspots and crisis events.

The above studies provide a useful basis for this study, but there are some gaps that need to be filled, especially regarding the public opinion analysis of online posts. The unresolved issues are related to the validity and usefulness of data and the popularity of posters when used in the clustering of hotspots. Based on the above research, a data filtering strategy is introduced to improve the quality of data. Improved co-word analysis and Bisecting k-means clustering algorithms are designed using time spans and popularity to obtain more accurate results.

3. Mathematical Model of Online Post

In this section, we first build a mathematical model of online posts based on their characteristics and then use it to study hotspots.

FIGURE 1: The structure of BBS post.

Let S_{all} be the set of all posts at time t . Let S_{valid} and S_{invalid} be the set of valid and invalid posts during period t . So, S_{all} can be represented as follows:

$$S_{\text{all}} = S_{\text{valid}} + S_{\text{invalid}}. \quad (1)$$

Assume that there are m valid posts, so S_{valid} can be expressed as $S_{\text{valid}} = \{s_1^{\text{valid}}, s_2^{\text{valid}}, \dots, s_m^{\text{valid}}\} \dots$. For any posts, the post usually contains the first replies set $F(s_i) = \{f_1^{s_i}, f_2^{s_i}, \dots\}$. Each $f_j^{s_i}$ has its replying content $\text{text}(f_j^{s_i})$, replying time $\text{time}(f_j^{s_i})$, and the second reply number $\text{num}(f_j^{s_i})$. The details are shown in Figure 2.

For time period t , the all topic set is

$$\begin{aligned} T_{\text{all}} &= T + T_{\text{similar}}, \\ T &= \{T_1, T_2, \dots, T_n\}, \end{aligned} \quad (2)$$

where n is the size of the valid topics set T and T_{similar} is similar to T on the topic set in period t . In other words, T_{similar} and T have similar keywords.

It is known that all posts (S_{all}) during a period have multiple topics; therefore, getting the post relevant to the one topic is the basement for extracting the hotspots. Each topic $T_i \in T$ corresponds to the keywords set ($X_{\text{all}}(T_i)$), which contains similar keywords' set ($X_{\text{all}}(T_i) = X(T_i) + X_{\text{similar}}(T_i)$). The valid keywords set has n_i valid keywords, and we can assume that its keyword set is $X(T_i) = \{x_1(T_i), x_2(T_i), \dots, x_{n_i}(T_i)\}$ ($(x_i(T_i) \in X(T_i), 1 \leq i \leq n)$).

We use the notation sT_i to indicate that post s ($s \in S_{\text{all}}$) is relevant to T_i . That is, s contains the keywords of T_i . We can use the following equation to define this relationship between a post and a topic:

$$f(s, T_i) = \begin{cases} 1, & sT_i, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

According to (3), the relevant posts are the post contents that contain the keywords of the one topic. Therefore, we can obtain the relevant posts set $S_{\text{all}}(T_i)$ of topic T_i as

$$S_{\text{all}}(T_i) = \{s_{\text{all}}(T_i) | f(s, T_i) = 1\}. \quad (4)$$

Definition 1. Post lifespan $t(s)$ is

$$t(s) = \text{day}(\text{time}_{f_{\text{end}}} - \text{time}_{f_0}), \quad (5)$$

where time_{f_0} and $\text{time}_{f_{\text{end}}}$ represent the time of the first and last reply of post s during a certain time, respectively, and $\text{day}(\cdot)$ is the number of days disregarding hours and minutes.

Definition 2. The reply number (the total number of replies of one post) is equal to

$$\text{TR}(s) = \sum_{j=1}^{|F|} \text{num}(f_j^s) + |F|, \quad (6)$$

where $\text{TR}(s)$ is the total number of replies of post s and $|F|$ is the total number of the first replies.

Definition 3. Post participants (p) are the post creator and repliers. Let $P(s_i)$ be the participant's participation number:

$$P(s_i) = P_{\text{first}}(s_i) + P_{\text{second}}(s_i), \quad (7)$$

where $P_{\text{first}}(s_i)$ and $P_{\text{second}}(s_i)$ are the first repliers and second repliers, respectively.

Definition 4. (degree of participation (DoP)). The degree of participation of post s_i is the ratio of the reply number and participant reply number

$$\delta_i = \frac{P(s_i)}{\text{TR}(s_i)}. \quad (8)$$

Definition 5. (minimum number of replies (MNR)). The MNR is the least reply limitation of the post s_i . Let ε_i be the MNR of the post s_i . MNR aims to simplify the valid post dataset without considering the posts with no replies and few replies.

It is clear that different posters and repliers have different influences. To calculate such influence, we use the popularity of a participant as the value. The popularity of a participant mostly depends on two factors: the frequency and the

FIGURE 2: The framework of a post or BBS.

popularity of replies. The frequency can be expressed as the average number of participants in posts, and the popularity can be expressed as the number of replies per day or per post. Accordingly, the post participant p 's popularity (forces(p)) can be expressed as

$$\text{forces}(p) = a \cdot fr + b \cdot \frac{1}{N} \cdot TSP, \quad (9)$$

where fr is the frequency of the posts of participant p , N is the total number of the discussion posts of p , TSP is the total number of repliers to all discussion posts of p , and a and b are the coefficients corresponding to the frequency and the total number of repliers, respectively.

We use different values to denote different post values for a topic. The value of a post is mainly determined by two criteria: the average number of replies and the popularity of the participant. Accordingly, the following formula is used to calculate the value:

$$\text{weight}(s) = \alpha \cdot TR(s) + \beta \cdot \sum \text{forces}(p(s)), \quad (10)$$

where α and β are the coefficients corresponding to the average number of replies and popularity, respectively.

Based on the above discussion, we can calculate the topic hot degree (HD) $\text{hot}(T)$ as

$$\text{hot}(T) = \vartheta \cdot |S(T)| + \theta \cdot \sum_{i=1}^m \text{weight}(s_i), \quad (11)$$

where $S(T)$ is the valid post set about topic T and ϑ, θ are the coefficients of the total number of posts and post values, respectively.

Definition 6. Hotspot is the topic that gets the maximum value of the hot degree. For the topic set $T = \{T_1, T_2, \dots\}$, the problem of hotspot search becomes

$$\begin{aligned} & \arg \max(\text{hot}(T_i)), \\ & \text{Subject To : } s_i \in S(T_i), \\ & T_i \in T, \\ & X(T_i) \in X_{\text{all}}(T_i), \end{aligned} \quad (12)$$

where $S(T_i)$ is the valid post set of topic T_i and the constraint conditions in equation (12) are restricted vales scope of the post, topic, and keywords, respectively.

Definition 7. Hot post is the post with the maximum value of a post. And, the hot post $s_{\max}$ of topic T_i can be expressed as follows:

$$\begin{aligned} s_{\max}(T_i) &= \arg \max(\text{weight}(s_i)), \\ \text{Subject To : } s_i &\in S(T_i). \end{aligned} \quad (13)$$

It is clear that the main problem related to formulas (12) and (13) is to obtain valid posts, replies, and keywords.

4. Spam Data Filtering Mechanism and Improved Cluster Hotspot Monitoring Algorithm

To determine the hotspots of BBS, we use a filtering mechanism to obtain more valuable data and an improved cluster hotspot monitoring algorithm to find hotspots. We focus on text data filtering, extracting keywords, constructing the common word matrix, and searching the hotspots and hotposts.

The main process involves the following steps: the identification of postspamming and fake replies to increase the post and reply values, the application of a text rank-based keyword extraction algorithm to calculate the PageRank Value of the candidate keywords and obtain their PR values, the determination of the keywords based on posts' PR (PageRank) values, and the construction of the co-word matrix for these keywords, as shown in Figure 3. As a result, we determine the hotspots by sorting the above results.

4.1. Post Filtering. Paid posts and invalid replies are well-known phenomena in BBS networks. However, they affect the results of hotspot and hot postsearch. Accordingly, we must filter out invalid data. We adopt the following rules.

Rule 1(degree of participation): if the degree of participation (DoP) is above a predefined constant, we consider the post as a paid post or spam post. And, we delete the post from the post set. The DoP is the ratio of

FIGURE 3: The main steps of the proposal strategy.

the participant's reply number and the total reply number of the post. Rule 1 can be formulated as follows:

$$S = \begin{cases} S_{all} \setminus \{s_i\}, & \text{if } \delta_i > \delta, \\ S_{all}, & \text{otherwise,} \end{cases} \quad (14)$$

where δ is the DoP constant.

Rule 2 (MNR): if the MMR of the post is less than a predefined constant, we consider the post as an invalid post or spam post. And, we delete the post from the post set. The MNR is the ratio of the total reply number and the participant's number of the post. Rule 2 can be formulated as follows:

$$S = \begin{cases} S_{all} \setminus \{s_i\}, & \text{if } \varepsilon_i < \varepsilon, \\ S_{all}, & \text{otherwise,} \end{cases} \quad (15)$$

where ε is the MNR constant.

We use Rule 1 and 2 to filter out paid or spam posts. It is known that, in a real BBS, there are many fake replies, which are not related to the topic, such as advertising. Such replies must be deleted from the post set as well.

Rule 3 (lexical filtering): the predefined vocabulary set is denoted as $A = \{a_1, a_2, \dots\}$ for topic T_i . If a reply text ($\text{Text}_{\text{reply}}(f_{ij})$) does not contain an element of A , it is considered invalid and gets deleted from the replies set. The rule for getting a valid reply can be described as follows:

$$F = \begin{cases} S = \begin{cases} F_{all} \setminus \{f_{ij}\}, & \text{if } \text{Text}_{\text{reply}}(f_{ij}) \not\supset A, \\ F_{all}, & \text{otherwise,} \end{cases} \end{cases} \quad (16)$$

where $\text{Text}_{\text{reply}}(f_{ij}) \not\supset A$ means the reply text contains a predefined vocabulary element of A , F_{all} is all reply for the topic T_i , and f_{ij} is the j th reply text of topic T_i .

Algorithm 1 shows the details of the post-text filtering for the topic T_i . In algorithm 1, the text data filtering can be mainly divided into the following steps. Firstly, the relevant post set of T_i can be obtained with equations (3) and (4). Then, according to Rule 1 and 2, the degree of participation and the minimum number

replies are employed for deleting the invalid posts which are beyond the constraints of (14) and (15). Moreover, using Rule 3, the invalid replies are selected and removed from the reply set F_{all} . Finally, the valid post and replying sets (S , F) are returned.

4.2. TextRank-Based Keyword Extraction Algorithm.

Based on Section 4.1, we can obtain valid posts and replies by adopting the filtering mechanism. In this section, we further extract keywords based on their text ranks. The main steps are as follows.

Step 1: we divide the text of a reply into a word list. Then, we order the words in the list. Namely, $\text{list}(V) = [v_1, v_2, \dots]$.

Step 2: after filtering the element of $\text{list}(V)$ according to the following rules, we obtain the list of candidate keywords.

Step 3: we use the following synonym processing rule to build candidate keywords.

Rule 4 (synonym processing): let C be the synonym keywords set:

$$C = \{C_1[c_1(\text{main}), C_1(\text{syn})], C_2[c_2(\text{main}), C_2(\text{syn})], \dots\} \quad (17)$$

where $c_i(\text{main})$ and $C_i(\text{syn})$ are the main word and its synonym set, respectively. If a word is a synonym, we use the main keyword to replace it and then merge the same words and build candidate keywords ($\text{list}(X)$).

Step 4 (build candidate keyword map): the candidate keyword map $G = (i, \text{list}(i))$, where i is the candidate keyword, $\text{list}(i)$ is the set of words co-existing with i in the window, and, for the word j in $\text{list}(i)$, the co-occurrence number between i and j is denoted as weight w_{ij} .

Step 5 (iterate operation): we set the number of iterations (L), according to the page rank algorithm [33, 34], we can calculate the page rank value (PR value) of each word and then construct a sequence according to the reverse order of PR values. The formula for the PR value is


```

Input:  $S_{all}, \varepsilon, A, \delta$ 
Output:  $S(T_i), F(T_i)$ 
Initialization  $S = S(T_i) \leftarrow S_{all}(T_i), F = F(T_i) \leftarrow F_{all}(T_i), S_{all}(T_i) \leftarrow S_{all}$ 
Computing the relevant post set  $S_{all}(T_i)$  of  $T_i$  // According to equations (3) and (4)
  for  $i = 1: |S|$  // Degree of participation filtering
    Calculate  $P(s_i)$ 
     $\delta_i \leftarrow P(s_i) / TR(s_i)$ 
    if  $\delta_i > \delta$ 
       $S \leftarrow S / s_i$ 
    end if
  end for
  for  $i = 1: |S|$  // MNR filtering
    Calculate  $TR(s_i)$ 
     $\varepsilon_i \leftarrow TR(s_i)$ 
    if  $\varepsilon_i < \varepsilon$ 
       $S \leftarrow S / s_i$ 
    end if
  end for
  for  $i = 1: |S|$ 
    for  $j = 1: |F|$  // Lexical filtering
      if  $Text_{reply}(f_{ij}) \triangleright A$ 
         $F \leftarrow F / f_i$ 
      end if
    end for
  end for
Return  $S(T_i), F(T_i)$ 

```

ALGORITHM 1: The post-text filtering.

$$PR(i) = (1 - d) + d^* \sum_{j \in \text{list}(i)} \frac{\omega_{ij}}{\sum_{l \in \text{list}(j)} \omega_{jl}} PR(j), \quad (18)$$

where $PR(i)$ refers to the PR value of keyword i , j denotes the keywords co-existing with i , l denotes the keywords co-existing with j , and d is the damping coefficient.

4.3. Common Word Matrix for Obtaining the Same Hotspots and Hot Posts. Common word matrix: n keywords are selected according to their PR values (Section 4.2). The keyword set is $W = \{w_1, w_2, \dots, w_n\}$. The position of the co-word matrix corresponds to the semantic distance between two keywords. The formula for the semantic distance $\text{dist}(w_i, w_j)$ between two keywords (w_i, w_j) is

$$\text{dist}(w_i, w_j) = \frac{1}{\text{count}(w_i, w_j) + 1}, \quad (19)$$

where $\text{count}(w_i, w_j)$ represents the number of co-occurrence events between keywords (w_i, w_j). The smaller the semantic distance between two keywords is, the more likely the two keywords belong to the same hotspot. Therefore, the common word matrix (CA) can be represented as

$$CA = \begin{pmatrix} \text{dist}_{11} & \dots & \text{dist}_{1n} \\ \vdots & \ddots & \vdots \\ \text{dist}_{n1} & \dots & \text{dist}_{nn} \end{pmatrix}. \quad (20)$$

Searching the hotspots and hot posts: the common word matrix CA is transformed into a point set. Then, all the

points are treated as the first cluster, and the first cluster is divided into two parts. Select each cluster that can minimize the SSE (sum of squared errors) value and divide it into two new clusters. This loop continues until the number of clusters equals the predefined number K .

The hotspots and keywords are obtained based on the above steps. Then, we can obtain the hot post using equation (10). By using the strategy explaining in equation (11), we can get the topic hot degree for every topic. Also, the hotspot and hot post by sorting can be identified.

5. Experimental Results

Experiments were conducted to evaluate the performance of the proposed algorithm using a real dataset. The results of the experiment are used to analyze the proposed approach. This section covers the simulation parameters, setup, and results.

5.1. Dataset and Experiment Setup. Dataset: the dataset is gathered from three online post websites (W1 (Baidu Tieba post): <https://tieba.baidu.com>; W2: <https://bbs.tianya.cn>; W3: <http://www.xici.net>). The three websites are the most famous online BBS platforms in China, which have more than 150 million active users in 2020. The post textual data was collected from these BBS and covers the whole year of 2018. Figure 4 shows a screenshot of the online post of W1, which is a classical online community post based on textual data.

To test the proposed strategies, we select three typical subjects ("Computer game" (S1), "Exam" (S2), and

FIGURE 4: The screenshot of a Baidu Tieba post.

“Nanfang College” (S3)) from the above BBS websites. The data was obtained using a crawler. Meanwhile, data visualization software was designed for analyzing these textual data of BBS, and the data filtering algorithm was used in the software, as shown in Figure 5. The dataset obtained has 16,373 posts and 100,197 replies from January 1, 2018, to December 31, 2018. The parameters of the experiment are shown in Table 1.

FIGURE 5: The data visualization software screenshot of BBS.

5.2. Results' Analysis. Valid posts and replies: the proposed data filtering mechanism is used to obtain valid data from the dataset. Figure 6 shows the results of the valid posts and replies of the above three subjects (S1, S2, S3). It is a well-known fact that, by using filtering strategies, we can effectively delete spam posts and replies. Figure 6(a) provides the comparison of the results of our filtering rules and the raw data in some different subjects for different subject posts. Our mechanism can decrease the number of posts by 109 and 513 by using Rules 1 and 2, respectively, compared to the raw data that was not filtered in S1. Accordingly, by using our filtering rules, more than 13% of the invalid post is obtained. Figure 6(b) shows the results of the filtering of online post repliers based on the proposed methods in different subjects by using rule 3. Similarly, the methods can effectively filter out invalid replies. Particularly, Rule 3 can filter more than 30% invalid replies.

The results demonstrate that the proposed filtering mechanisms can decrease the number of invalid posts and replies. Also, the filter can reduce datasets and improve the efficiency of searching for hotspots. Furthermore, the results show that the proposed data filtering algorithm has different post and replies effects on a different subject. In other words, the larger the scope of the subject, the bigger the post and replies. Topics with a wide scope of topics are more likely to have spam posts and replies.

Precision: for verifying the data filtering algorithm performance of precision, the part raw data (10%) of the subject of S3 is selected. Then, these BBS data are filtered by the manual and the proposed data filtering algorithm, respectively. Figure 7 shows the precision percentage results of posts and replies of subject S3 by using the proposed method in different BBS websites. From the results, it is easy to get that the precision of filtering posts is more than 92%, and the precision of filtering replies is large than 85%. The results

demonstrate that the proposed data filtering algorithm has a good effect on the precision of spam posts and replies.

Computing time: computing time is an important metric to evaluate the performance of the data filtering algorithm. Therefore, the computing time results to collect the number of users in different subjects is given in raw and filtered data, as shown in Figure 8. It is easy to get that the S1 spends the most computing time in three subjects. And, the proposed filtering can decrease more than 15% computing time. In other words, the filtered data used less time to search the number of users than the raw dataset in all subjects. The results show that the presented strategy can save more computing time by using a data filtering algorithm.

Hot degrees: we use hot degrees to search for hotspots and posts. When the hot degree of a post reaches 3, we consider it a hotspot. After calculating hot degrees and searching for hotspots, five hotspots were selected based on their hot degrees. Hot degrees of different topics can be obtained using the hot degree calculation method. Meanwhile, the same five hotspots of the maximum number of post and repliers are calculated in the same dataset. The results of different metrics are shown in Figure 9. Figure 9(a) provides the hot degrees of different topics from the 90th to 105th days. The values of the hot degrees of the five topics are 3.6, 5.6, 5.6, 11.75, and 5.08. It is noticeable that topic 4 is the hottest topic during this period. Figures 9(b) and 9(c) show the total numbers of post and repliers in the same time period. It can be seen that topic T1 is the hotspot using the different degree, and it has the highest values of the three metrics. Namely, the hot degree can reflect the hotspots of online posts.

The values of hot degrees on subject S1 are obtained in different datasets from different websites. Figure 10 shows the hot degree results from the 5th to 235th days related to topic 1. From Figure 7, it can be seen that topic 1 has three

TABLE 1: The experiment parameters.

Parameters	Values	Descriptions
A	0.7	The coefficients of the frequency of repliers
b	0.3	The coefficients of the total number of repliers
α	0.9	The coefficients of the average number of replies
β	0.1	The coefficients of the popularity of replier
ϑ	0.1	The coefficients of the total number of posts
θ	0.9	The coefficients of the post value
δ	0.5	The DoP constant value
$window$	5	The number of accommodated keywords
L	100	The number of iterations
d	0.85	The damping coefficient
K	3	The number of cluster centers
ε	5	The minimum number of replies constant value

FIGURE 6: The comparison of valid posts and replies. (a) Filtering of posts. (b) Filtering of replies.

FIGURE 7: The comparison of precision of data filtering.

peaks at the 60th, 120th, and 190th days, and the hot degree of three websites has the same trend in different online post websites during the monitored period. The proposed hot degree can directly reflect fluctuation trends. Specifically, a

hot degree can demonstrate the trend in terms of repliers and users, as in our strategies we merge post users and replies. In summary, the proposed method can effectively solve the social media hotspot problem.

FIGURE 8: The results of computing time.

FIGURE 9: The results on the hot degrees (a), the number of post (b), and repliers (c) for five topics.

FIGURE 10: The hot degree results in different datasets.

6. Conclusion

The online posts have become public platforms for expressing personal opinion, so their monitoring and online hot topic search gained more significance. Considering the weight of different users, the extraction of hotspots from massive textual data with spam data become one of the important bases for study the public opinion of the social network. By collecting and analyzing text information on online posts, current hotspots can be obtained. This article adopts a data filtering mechanism, common words, and clustering technology for online hotspots search, using a time span, poster popularity, and PR values. Then, hot degree is used to evaluate the hotspots of online posts based on the number of replies and the popularity of the participant. The proposed methods are implemented and applied to a BBS dataset. The results show that the proposed method can effectively filter out invalid data, compress datasets, save more computing time, and improve performance. At the same time, the results demonstrate that the proposed method and hot degree can also reflect changes in the trend of the hotspots of online posts.

Data Availability

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

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Hui-Ru Cao conceived and wrote the manuscript; Songyao Lian analyzed the data and performed the experiments; Chou-Jun Zhan and Xiaomin Li analyzed the experimental results.

Acknowledgments

This work was supported by the Ministry of Education in China Liberal Arts and Social Sciences Foundation under Grant no. 20YJCZH004, Natural Science Foundation of Guangdong Province of China under Grant 2019A1515011346, Featured Innovation Projects of Guangdong Province universities of China under Grant no. 2019GKTSCX075, and National Science Foundation of China Project under Grant no. 61703355. This work was also partially supported by colleagues at the Department of Electronic Communication and Software Engineering of Sun Yat-sen University.

References

- [1] W. Hanson, "A global internet: the next four billion users," *New Space*, vol. 3, no. 3, pp. 204–207, 2015.
- [2] A. Zubiaga, A. Aker, K. Bontcheva et al., "Detection and resolution of rumours in social media: a survey," *Acm Computing Surveys*, vol. 51, no. 2, 2018.
- [3] J. Katz and P. Aspden, "Motivations for and barriers to Internet usage: results of a national public opinion survey," *Internet Research*, vol. 7, no. 3, pp. 170–188, 1997.
- [4] J. Zeng, S. Zhang, C. Wu, and J. Xie, "Predictive model for internet public opinion," in *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, pp. 1–11, IEEE, Haikou, China, 2007.
- [5] D. Pokotylo, *Online Public Opinion and Archaeological Heritage Conservation: A Case Study from Western Canada*, pp. 35–48, Relevance and Application of Heritage in Contemporary Society, London, UK, 2018.
- [6] X. Chen, M. Xia, J. Cheng et al., "Trend prediction of internet public opinion based on collaborative filtering," in *Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*, pp. 583–588, IEEE, Xian, China, 2016.
- [7] Q. Zhang and X. Zhang, "Group decision method for internet public opinion emergency with linguistic preference," in *Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 439–443, IEEE, Changsha, China, 2016.
- [8] Y. Song, X. Y. Dai, and J. Wang, "Not all emotions are created equal: expressive behavior of the networked public on China's social media site," *Computers in Human Behavior*, vol. 60, pp. 525–533, 2016.
- [9] C. Wang, H. Liu, and X. Guan, "Response effect assessment of internet public opinion based on fuzzy comprehensive evaluation," in *Proceedings of the International Conference on Logistics, Informatics and Service Sciences*, pp. 1–6, IEEE, Changsha, China, 2016.
- [10] Y. Yang, "Research and realization of internet public opinion analysis based on improved TF-IDF algorithm," in *Proceedings Of The International Symposium On Distributed Computing And Applications To Business, Engineering And Science*, pp. 80–83, IEEE Computer Society, Hong Kong, China, 2017.
- [11] E. Oster, E. Gilad, and A. Feigel, "Internet comments as a barometer of public opinion," *EPL (Europhysics Letters)*, vol. 111, no. 2, p. 28005, 2015.
- [12] X. Gao and L. Fu, "Methods of uncertain partial differential equation with application to internet public opinion problem," *Journal of Intelligent & Fuzzy Systems*, vol. 33, no. 1, pp. 1–11, 2017.
- [13] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada, and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.
- [14] H. Zhu, P. Liu, and X. Shan, "Analysis of internet-based public opinion in China, 2012," *Journal of Molecular Neuroscience*, vol. 49, no. 3, pp. 614–617, 2015.
- [15] S. A. Salloum, C. Mhamdi, M. Al-Emran et al., "Analysis and classification of Arabic newspapers' Facebook pages using text mining techniques," *International Journal of Information Technology and Language Studies*, vol. 1, no. 2, pp. 8–17, 2017.
- [16] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *Computer*, vol. 52, no. 1, pp. 36–45, 2019.
- [17] D. Ruano-Ordás, J. Fdez-Glez, F. Fdez-Riverola, and J. R. Méndez, "Using new scheduling heuristics based on resource consumption information for increasing throughput on rule-based spam filtering systems," *Software: Practice and Experience*, vol. 46, no. 8, pp. 1035–1051, 2016.
- [18] D. Ruano-Ordás, J. Fdez-Glez, F. Fdez-Riverola, V. Basto Fernandes, and J. R. Méndez, "RuleSIM: a toolkit for simulating the operation and improving throughput of rule-based

- spam filters,” *Software: Practice and Experience*, vol. 46, no. 8, pp. 1091–1108, 2016.
- [19] D. Ruano-Ordás, J. Fdez-Glez, F. Fdez-Riverola, and J. R. Méndez, “Effective scheduling strategies for boosting performance on rule-based spam filtering frameworks,” *Journal of Systems and Software*, vol. 86, no. 12, pp. 3151–3161, 2013.
 - [20] G. Neubaum and N. C. Krämer, “Monitoring the opinion of the crowd: psychological mechanisms underlying public opinion perceptions on social media,” *Media Psychology*, vol. 20, no. 3, pp. 502–531, 2017.
 - [21] A. Hernandez-Suarez, G. Sanchez-Perez, K. Toscano-Medina et al., “Social sentiment sensor in twitter for predicting cyber-attacks using ℓ_1 regularization,” *Sensors*, vol. 18, no. 5, 2018.
 - [22] Y. Dong, Z. Ding, F. Chiclana, and E. Herrera-Viedma, “Dynamics of public opinions in an online and offline social network,” *IEEE Transactions on Big Data*, p. 1, 2017.
 - [23] W. Zhang, X. Li, H. He et al., “Identifying network public opinion leaders based on Markov Logic Networks,” *The Scientific World Journal*, vol. 2014, no. 5, p. 268592, 2014.
 - [24] L. Zhang, C. Su, Y. Jin, M. Goh, and Z. Wu, “Cross-network dissemination model of public opinion in coupled networks,” *Information Sciences*, vol. 451–452, pp. 240–252, 2018.
 - [25] Z. Feng, “Hot news mining and public opinion guidance analysis based on sentiment computing in network social media,” *Personal and Ubiquitous Computing*, vol. 23, pp. 373–381.
 - [26] D. Jiang, X. Luo, J. Xuan, and Z. Xu, “Sentiment computing for the news event based on the social media big data,” *IEEE Access*, vol. 5, no. 99, pp. 2373–2382, 2017.
 - [27] Y. Zhao, B. Qin, T. Liu, and D. Tang, “Social sentiment sensor: a visualization system for topic detection and topic sentiment analysis on microblog,” *Multimedia Tools and Applications*, vol. 75, no. 15, pp. 8843–8860, 2016.
 - [28] D. Zheng and F. Li, “Hot topic detection on bbs using aging,” in *Proceedings of International Conference on Web Information Systems and Mining (WISM’09)*, Shanghai, China, November 2009.
 - [29] X. iulan Hao and Y. Hu, “Topic detection and tracking oriented to bbs,” in *Proceedings of 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE)*, Changchun, China, 2010.
 - [30] H. Liu and X. Li, “Internet public opinion hotspot detection research based on k-means algorithm,” in *Proceedings of Advances in Swarm Intelligence, First International Conference, ICSI 2010*, Beijing, China, June 2010.
 - [31] N. Li and D.D. Wu, “Using text mining and sentiment analysis for online forums hotspot detection,” *Decision Support Systems*, vol. 48, no. 2, pp. 354–368, 2010.
 - [32] X. G. Chen, S. Duan, and L. D. Wang, “Research on clustering analysis of Internet public opinion,” *Cluster Computing*, vol. 22, no. 3, 2019.
 - [33] Y. Gao, X. Yu, and H. Zhang, “Overlapping community detection by constrained personalized PageRank,” *Expert Systems with Applications*, vol. 173, Article ID 114682, 2021.
 - [34] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 107–117, 1998.