Postprint: Research on Weibo Emergency Event Detection Based on Bursty Topic Words and Agglomerative Hierarchical Clustering

Authors: Ding Shengchun, Gong Silan, Li Hongmei

Date: 2017-10-11T00:00:00+00:00

Abstract

【Objective】 To detect emergent events in massive Weibo data in real-time, accurately, and efficiently, providing important decision-making information support for public opinion emergency management.

【Method】 By introducing a reference time window mechanism, we design selection and calculation methods for four types of features: term frequency, document frequency, hashtags, and term frequency growth rate, to extract burst topic words based on dynamic thresholds. On this basis, Weibo texts are represented as feature vectors of burst topic words, and agglomerative hierarchical clustering algorithm is used to achieve emergent event detection.

【Results】By analyzing experimental results with actual cases, the emergent event detection achieves 80% accuracy, verifying the feasibility and effectiveness of the proposed method.

【Limitations】 Due to limitations in corpus data and research scope, automatic description of detected emergent events has not yet been realized, and there are also certain deficiencies in the analysis and consideration of elements such as netizen sentiment and semantic relationships between events.

【Conclusion】This study overcomes the limitations of previous related research in terms of text content quality, text format, and burst feature extraction results, and improves the efficiency of Weibo emergent event detection.

Full Text

A New Method for Detecting Emergencies from Microblog Posts Based on Bursty Topic Words and Agglomerative Hierarchical Clustering

Ding Shengchun, Gong Silan, Li Hongmei
(School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract

[**Objective**] This study proposes a method to detect emergencies from massive microblog posts in real time, accurately, and efficiently, providing critical decision-making support for public opinion emergency management.

[**Methods**] We introduced a reference time window mechanism and designed selection and calculation methods for four types of features: word frequency, document frequency, hashtags, and word frequency growth rate. Based on dynamic thresholds, we extracted bursty topic words. We then represented microblog texts as feature vectors of bursty topic words and applied an agglomerative hierarchical clustering algorithm to detect emergency events.

[**Results**] Experimental results analyzed with real-world cases demonstrated that our method achieved 80% accuracy in emergency event detection, verifying its feasibility and effectiveness.

[**Limitations**] Due to limitations in corpus data and research scope, the study has not yet achieved automatic description of detected emergencies, and the analysis of factors such as user sentiment and semantic relationships between events remains inadequate.

[**Conclusions**] This research overcomes previous limitations in text content quality, text format, and bursty feature extraction results, improving the efficiency of emergency event detection from microblog posts.

Keywords: Emergency event detection; Bursty topic words; Agglomerative hierarchical clustering algorithm; Network public opinion; Microblog

1. Introduction

As a new social media platform, microblog (Weibo) is characterized by convenient usage, rapid dissemination, strong interactivity, and comprehensive content, making it an important channel for the rapid aggregation and propagation of emergency information. Emergency events refer to sudden occurrences that cause or may cause serious social harm, requiring emergency response measures to address natural disasters, accidents, public health incidents, and social security events. Such events are instantaneous, with accidental 爆发 points, and their timing and location are highly unpredictable. When emergencies occur,

an increasing number of netizens habitually use microblogs to publish and obtain real-time information while expressing their personal views and attitudes. Furthermore, the frequent occurrence of emergencies has drawn widespread attention to network public opinion analysis on microblog platforms. Accurately and efficiently detecting emergencies from massive microblog posts at the first moment of their outbreak can not only help users obtain critical emergency information in real time and alleviate panic but also assist emergency management agencies in grasping the development trends of emergencies, rationally controlling and guiding public opinion, and providing decision-making information support for emergency management. This plays a significant role in leveraging the positive functions of network public opinion in ensuring citizens' right to know and maintaining social stability and healthy development.

Research on emergency event detection for microblogs has achieved certain results, mainly divided into document-centered detection research [1] and feature-centered detection research [2].

Document-centered emergency event detection techniques directly cluster documents, treating clusters as emergency events, and then extract event features to represent the detected emergencies [3-4]. Petrović et al. [5] proposed a Twitter text clustering algorithm based on LSH (Locality-Sensitive Hashing), which optimized the time efficiency of event detection on social media while maintaining constant time and space complexity. Phuvipadawat et al. [6] studied unsupervised clustering methods for breaking news events on Twitter, selecting both general features and microblog-specific features, and used the TF-IDF method to assign weights to each feature word, achieving good clustering results. Ge Gaofei [7] proposed an improved TC-LDA algorithm to address the noise problem in emergency event detection.

Feature-centered emergency event detection techniques [8] focus on detecting bursty features that change over time in real-time data streams, i.e., extracting bursty topic words [9]. By clustering these bursty topic words or using them to represent texts before applying clustering algorithms, the goal of emergency event detection is achieved. This method can avoid data sparsity issues. However, microblog texts are short yet voluminous, containing substantial noise such as advertisements and online fraud, and are highly real-time, making emergency event detection on microblogs more susceptible to spam information [10]. To address noisy data, researchers emphasize utilizing temporal information and combining microblog' s inherent attribute functions like Hashtags [11] to mine bursty features presented during event periods.

Kleinberg [12] early discovered that document streams exhibit characteristics of suddenly appearing for a period before disappearing, proposing the classic Bursty mining method. He et al. [13] analyzed word trends in time series and applied them to unsupervised emergency event identification algorithms. Mathioudakis et al. [14] implemented the "TwitterMonitor" system, which clusters words with abnormally high frequency in Twitter within specific time periods to discover emerging emergencies in real time. Long et al. [15] introduced docu-

ment frequency, Hashtag, and information entropy factors in event detection to extract topic words representing emergencies, constructing word co-occurrence graphs and applying clustering algorithms to obtain events from microblogs. Zhao Wenqing et al. [16] used relative word frequency and word frequency growth rate to extract topic words of emergencies, clustering based on word co-occurrence graphs and treating clusters as microblog news events. Yao et al. [17] detected events in microblogs by monitoring changes in user-generated Hashtag markers. Wang Yong et al. [18] calculated word weights from three aspects—word frequency statistics, word frequency growth rate, and TF-PDF—to extract bursty word sets, proposing an "absolute clustering" algorithm to detect emergencies more accurately. Guo Yixiu et al. [19] integrated microblog text features, propagation features, and user influence to extract bursty words, using agglomerative hierarchical clustering on bursty words to detect emergencies on microblogs.

In summary, existing research on emergency event detection still has certain limitations, mostly constrained by factors such as text content quality, text format, and bursty feature extraction results. Based on this, our study introduces a filtering strategy for the three essential elements of microblog events to control text content quality. Simultaneously, considering microblog text format characteristics, we preprocess microblog data through traditional/simplified Chinese conversion, word segmentation, stop-word processing, and part-of-speech filtering to filter noise information that may affect bursty features and unify text formats. Then, based on comprehensive consideration of words' thematic expression capability and burstiness, we introduce a reference time window mechanism, design selection and calculation methods for four features—word frequency, document frequency, hashtag, and word frequency growth rate—and extract effective bursty topic words characterizing events based on dynamic thresholds. Finally, we represent microblog texts as feature vectors, construct microblog text similarity matrices, and use agglomerative hierarchical clustering algorithms to detect emergency events on microblogs.

3. Emergency Event Detection Research Framework

Public opinion on emergencies results from internet users holding respective viewpoints and communicating with each other around specific emergencies, forming certain information flows that exhibit periodic characteristics. After emergencies erupt on microblog platforms, some features used to describe events are widely mentioned. From a linguistic perspective, these features are bursty words appearing in microblog text content within specific time periods. However, using words alone cannot distinguish events; it is necessary to locate microblog texts within corresponding time periods using bursty words and achieve emergency event detection through text clustering.

The specific detection framework is shown in Figure 1 [Figure 1: see original paper]. This research mainly addresses the following issues:

(1) Microblog data contains much spam information, and bursty features are easily affected by noise. Therefore, special attention must be paid to noise and spam information filtering in microblog data before extracting bursty topic words, along with preprocessing operations such as text segmentation and part-of-speech tagging.

(2) Addressing the propagation characteristics of emergency events on microblogs, we divide collected microblogs into time windows to construct microblog data streams on time series. By capturing the temporal distribution and bursty patterns of words in different time windows, we utilize four features—word frequency, document frequency, hashtags, and word frequency growth rate—to extract event bursty topic words based on dynamic thresholds.

(3) After filtering texts describing emergencies, we use bursty topic words as bursty features to represent microblog texts and apply agglomerative hierarchical clustering strategies to cluster microblog texts into clusters, treating clustering results as emergency events.

(4) We verify the research method through experiments and analyze the emergency event detection effects with real-world cases.

4. Emergency Event Detection Method Based on Bursty Features

4.1 Bursty Topic Word Extraction (1) Bursty Topic Word Feature Analysis

Bursty topic words are content words that are extensively used within a certain time window but rarely used in previous time windows [9]. Based on microblogs' inherent characteristics of timeliness and fission propagation, before extracting bursty topic words for emergencies, it is necessary to first divide continuous microblog data streams into independent time periods. This paper divides microblog data into m×t time windows, where m is measured in "days." To more granularly detect the occurrence time of events on microblogs in real time, t can be further divided into finer time segments based on needs, measured in "days," "hours," "minutes," or "seconds." To enable extracted words from microblogs to more comprehensively describe emergencies, this paper establishes bursty topic word measurement standards from four aspects—word frequency, document frequency, hashtags, and word frequency growth rate—to determine whether a word can become a bursty topic word.

Word Frequency

Word frequency can measure the importance of a vocabulary item in a document. Under statistical significance, if a word appears frequently, it means the word is more likely related to the topic expressed in the text. Therefore, this paper adopts word frequency as one of the measurement methods for bursty topic word selection.

Document Frequency

For emergency events, if the number of microblogs containing a certain word is relatively high in the current time window, the word is more likely to be a feature word of an emergency event. To ensure the thematic expressiveness of selected feature words, this paper adjusts document frequency by introducing the concept of entropy to measure a word' s expressiveness for emergency event themes in that time window. Larger entropy indicates that the word can better express the theme.

Hashtag

As one of microblog' s most distinctive functional attributes, hashtags allow users to create topic tags for published information content [20]. Feature words more closely related to events are more likely to appear in microblog hashtags. This paper fully utilizes the hashtag feature of microblogs, measuring the degree of correlation between a word and an emergency event by calculating the word' s hashtag weight [21]. The calculation formula is as follows:

$$HT_{ij} = \begin{cases} \frac{h_i}{h_i'} & \text{if } l(w_i) = 1 \\ 0 & \text{if } l(w_i) = 0 \end{cases}$$

where HT_{ij} is the hashtag weight of word w_i in time window j, $l(w_i)$ is a discriminant function where $l(w_i) = 1$ indicates that at least one hashtag contains word w_i and $l(w_i) = 0$ indicates that no hashtag contains word w_i, h_i is the count of occurrences of word w_i in hashtags, and h_i' is the number of microblog posts in the current time window that contain word w_i and include the hashtag symbol #.

Word Frequency Growth Rate

The burstiness of words presents a state of sharp increase over time, with the most obvious feature being the use of word frequency increment to screen bursty topic words in the current time window. Word frequency increment is usually calculated using the proportional change of word frequency in adjacent time windows [18]. Meanwhile, to avoid interference from adjacent time windows during the event duration on word frequency growth rate results, this paper combines the relative time window and adjacent time window in the reference time window mechanism for comparison. The calculation formula is as follows:

$$FT_{ij} = \lambda_1 \cdot \frac{f_{ij} - f_{i(j-1)}}{f_{i(j-1)}} + \lambda_2 \cdot \frac{f_{ij} - f_{i'j'}}{f_{i'j'}}$$

where FT_{ij} represents the word frequency growth rate of word w_i in the current time window j, f_{ij} is the word frequency of word w_i in time window j, $f_{i(j-1)}$ is the word frequency of word w_i in the previous time window $j-1$. If "day" is used as the time unit, then $f_{i'j'}$ is the word frequency of word w_i in time window $j-2$. If "hour" is used as the time unit, then $f_{i'j'}$ corresponds to the word frequency of word w_i in the j' time window of the previous day. λ_1 and λ_2 are adjustment coefficients, with $\lambda_1 + \lambda_2 = 1$.

Based on the above analysis, words with high word frequency, document frequency, hashtag weight, and word frequency growth rate in microblogs are more likely to become bursty topic words describing events. The bursty topic degree of words will be calculated by combining the normalized results of these four features. The calculation formula is as follows:

$$BTword_{ij} = F'_{ij} + DF'_{ij} + HT'_{ij} + FT'_{ij}$$

where $BTword_{ij}$ represents the bursty topic degree of word w_i in time window j, and F'_{ij}, DF'_{ij}, HT'_{ij}, and FT'_{ij} are the normalized word frequency, document frequency, hashtag weight, and word frequency growth rate, respectively. The final bursty topic word set $BTword$ is represented as: $BTword = \{word_1, word_2, word_3, \dots, word_k\}$, where $word_k$ represents the k-th bursty topic word in the current time window j.

(2) Bursty Topic Word Extraction Algorithm

Whether a word can become a bursty topic word must first meet the set threshold δ standard, and then the bursty topic degree of all words meeting the standard is calculated. Threshold δ includes: the average value δ_1 of word frequency of all words in the current time window; the average value δ_2 of document frequency of all words in the current time window; the empirical dynamic threshold δ_3 adjusting the bursty characteristics of words; and the average value δ_4 of bursty topic degree of words meeting the first three thresholds in the current time window. The specific process of the bursty topic word extraction algorithm is as follows:

Input the microblog data stream, assign it to different time windows according to the post publication time, then perform statistics after preprocessing microblogs in each window to obtain the total word list W in each time window.

Read word w_i from word sequence W and execute step .

Calculate the word frequency of word w_i and determine whether it is greater than threshold δ_1. If greater, retain the word and execute step ; otherwise, filter out word w_i, set $i = i + 1$, and jump to step .

Calculate the document frequency of word w_i and determine whether it is greater than threshold δ_2. If greater, retain the word and execute step ; otherwise, filter out word w_i, set $i = i + 1$, and jump to step .

Calculate the word frequency growth rate of word w_i and determine whether it is greater than threshold δ_3. If greater, retain the word and execute step ; otherwise, filter out word w_i, set $i = i + 1$, and jump to step .

For words w_i meeting the above threshold standards, first calculate the hashtag rate, then comprehensively calculate their bursty topic degree, and determine whether it is greater than threshold δ_4. If greater, retain the word and execute step ; otherwise, filter out word w_i, set $i = i + 1$, and jump to step .

Add word w_i to the event bursty topic word list $BTword$, and finally output all bursty topic words in that time window.

By processing all words in the time window according to the above process, retaining all words meeting the threshold as bursty topic words, these bursty topic words have both high thematic expressiveness and can reflect the bursty characteristics of events, thus effectively characterizing emergency events.

4.2 Emergency Event Detection (1) Microblog Text Feature Representation Based on Bursty Topic Words

For any microblog text in a certain time window, we construct text feature vectors based on the bursty topic word set $BTword = \{word_1, word_2, word_3, \dots, word_k\}$ in the current time window. The formal vector representation of microblog texts is defined as follows:

$$\text{text}_i = \{term_{i1}, term_{i2}, term_{i3}, \dots, term_{ik}\}$$

where text_i represents the i-th microblog text, and $term_{ik}$ indicates whether the i-th microblog text contains the k-th bursty topic word, where $term_{ik} = 1$ means it contains the bursty topic word and $term_{ik} = 0$ means it does not. For example, if the bursty topic word set in time window j is {nurse, Nanjing, Yuan Yaping, official, paralysis}, and the bursty topic words contained in microblog text text_i are {nurse, Nanjing, paralysis}, then text_i can be represented as: $\text{text}_i = \{1, 1, 0, 0, 1\}$.

Drawing on the microblog text filtering principle in literature [18], we consider that a microblog text describing an event should contain at least any 3 elements of "5W1H." However, element types are no longer distinguished in the microblog text feature vector. Specifically, when applied to bursty topic word feature vectors of microblog texts, they should contain at least 3 bursty topic words. By removing all microblog texts containing fewer than 3 bursty topic words, we can effectively reduce the sparsity of the microblog text–bursty topic word matrix, improve the efficiency of emergency event detection, and simultaneously ensure the completeness of detection results.

(2) Emergency Event Detection Algorithm Based on Agglomerative Hierarchical Clustering

After feature representation of microblog texts, we find that users' language expressing emergency events on microblogs is relatively similar, showing a "onlooking" phenomenon. Microblogs related to events generally appear concentrated, and words in microblog texts usually revolve around certain event feature words with high repetition rates. Therefore, we believe that the more bursty topic words different microblog texts contain in common, the more likely they are describing the same emergency event.

Regarding the selection of similarity calculation methods, using the Jaccard coefficient method to determine the similarity between microblog text feature vectors better conforms to the real situation of emergency event aggregation

and can reflect the true similarity between microblog texts. Specifically, two microblog texts discussing the same event should have relatively high overlap. The specific Jaccard similarity calculation formula is as follows:

$$S(\text{text}_i, \text{text}_j) = \frac{|\text{text}_i \cap \text{text}_j|}{|\text{text}_i \cup \text{text}_j|}$$

where $S(\text{text}_i, \text{text}_j)$ is the similarity between two microblog texts, text_i represents the microblog text feature vector, $\text{text}_i \cap \text{text}_j$ represents the intersection of text_i and text_j, and $\text{text}_i \cup \text{text}_j$ represents the union of text_i and text_j.

The process of the agglomerative hierarchical clustering-based event detection algorithm is as follows:

Input the microblog set of time window j, represent microblog texts as bursty topic word feature vectors, denoted as text_i, filter out vectors text_i with fewer than 3 bursty topic words, and form the microblog text–bursty topic word matrix D.

Initialize each microblog text feature vector as a class, calculate pairwise similarity values $S_{i,j}$ of microblog text feature vectors using the Jaccard coefficient, and construct the microblog text similarity matrix S.

Find the maximum value $\max\{S_{i,j}\}$ in similarity matrix S.

According to the merging rules of hierarchical clustering, merge event class i and event class j into a new vector, simultaneously recalculate the similarity between this new vector and existing event class vectors, and readjust similarity matrix S.

Determine whether the number of columns or rows in matrix S meets the preset threshold. If satisfied, execute step ; otherwise, jump to step .

Through this clustering process, all microblog texts are finally clustered into n clusters. Map microblog text feature vectors text_i back to original microblog texts and output the final clustering results, where each cluster represents an emergency event.

5. Experiments and Results Analysis

5.1 Experimental Data Source and Evaluation Metrics (1) Data Source and Preprocessing

In the research field of emergency event detection for microblog data, there is currently no internationally recognized standard test corpus. The experimental data in this paper comes from Sina Weibo, with data crawling implemented through a microblog crawler based on its open platform API. Limited by the frequency and quantity restrictions of the API interface, only partial data from Sina Weibo was obtained (over 1.8 million microblog posts from February 25 to March 11, 2014). From an experimental perspective, these microblog data can serve as a sample representative of the complete data to support the experimental analysis and research in this paper.

Microblog data is filled with substantial spam and noise information, which can seriously impact emergency event detection results. Before conducting emergency event detection on microblogs, preprocessing of microblog data is required, including noise filtering, traditional/simplified Chinese conversion, word segmentation, stop-word processing, and part-of-speech filtering. Noise in microblogs mainly includes @XXX noise, URL link noise, and emoticon symbols. This paper sets regular expressions for specific noise types to filter them. Further, according to the "Common Standard Chinese Characters Table" [22], all traditional characters and their corresponding simplified characters are extracted to respectively construct the "Traditional Chinese Characters Table"and "Simplified-Traditional Chinese Characters Correspondence Table"to achieve traditional/simplified Chinese conversion of microblogs. Subsequently, the NLPIR Chinese word segmentation system [23] is used for microblog word segmentation, and part-of-speech filtering is implemented based on its annotations, retaining nouns and verbs. Finally, a stop-word list is used to filter stop words through vocabulary matching, completing the preprocessing of microblog data.

(2) Experimental Results Evaluation Metrics

Bursty Topic Word Extraction Evaluation

Traditional evaluation metrics include three parameters: Precision, Recall, and F-measure. Since it is impossible to obtain all bursty topic words in the current time window, Recall is difficult to calculate directly. Therefore, this paper uses words that are correctly extracted as bursty topic words without filtering by the bursty topic degree average threshold δ_4 in the current situation as the overall bursty topic words of that time window to calculate recall. Words extracted under the condition of meeting threshold δ_4 are considered the final words needed for the experiment. Precision, Recall, and F-measure are used for evaluation. Whether bursty topic words are correctly extracted is manually judged by determining whether the extracted bursty topic words in that time window can describe or summarize emergencies occurring in real life. The specific evaluation formulas are as follows:

$$\text{Precision}(BTword) = \frac{k}{K}$$

$$\text{Recall}(BTword) = \frac{k}{S}$$

$$\text{F-measure}(BTword) = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where Precision($BTword$) represents the precision of bursty topic word extraction, Recall($BTword$) represents the recall of bursty topic word extraction, k represents the number of correctly extracted bursty topic words in the current

time window, K is the total number of bursty topic words extracted in the current time window, and S is the total number of correctly extracted bursty topic words among all words not filtered by the bursty topic degree average threshold δ_4.

Emergency Event Detection Evaluation

For emergency event detection results, since emergencies occurring in real life cannot be known in advance, i.e., the total number of all emergency events in microblogs within a certain time window is difficult to obtain beforehand, the recall of this result cannot be directly calculated. Therefore, this experiment only selects precision to evaluate whether the detected emergency events are correct. The experimental detection results of emergency events are manually compared by judging whether the detected emergency events reflect real emergencies. If they do, they are considered correctly identified; otherwise, they are considered incorrect. The emergency event detection evaluation formula is as follows:

$$\text{Precision(event)} = \frac{e}{E}$$

where Precision(event) is the precision of emergency event detection, e represents the number of correctly detected emergency events in the current time window, and E is the total number of emergency events detected in the current time window.

5.2 Bursty Topic Word Extraction Experiments and Results Analysis

For bursty topic word extraction of events in microblog texts, this paper selects four types of features: word frequency, document frequency, hashtag rate, and word frequency growth rate, as shown in Table 1 . Five groups of feature combination calculation methods are designed for comparative experiments to demonstrate the effectiveness of the selected feature calculation methods. Method 1 is used to examine the impact of the four-feature combination calculation method on bursty topic word extraction, while Methods 2 to 5 are used to analyze the impact of word frequency growth rate, hashtag rate, document frequency, and word frequency calculation methods on bursty topic word extraction.

Table 1. Comparative Design of Feature Calculation Method Combinations for Bursty Topic Words

Feature Calculation Method Combination
Word frequency, Document frequency, Hashtag rate, Word frequency growth rate
Word frequency, Document frequency, Hashtag rate
Word frequency, Document frequency, Word frequency growth rate
Word frequency, Hashtag rate, Word frequency growth rate
Document frequency, Hashtag rate, Word frequency growth rate

Using "day" as the time window, experiments were conducted with data from February 25 to February 27, 2014. The adjustment coefficients in the word frequency growth rate calculation formula were set as $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, and the word frequency growth rate threshold $\delta_3 = 0.5$. The statistical results of bursty topic word extraction for February 27 data under different feature calculation methods are shown in Figure 2 [Figure 2: see original paper].

According to Figure 2, comparing Method 2 with Method 1, both precision and recall decrease significantly, indicating that the word frequency growth rate calculation method is particularly important for extracting bursty topic words. Using word frequency growth rate to judge the bursty characteristics of words can improve the precision and recall of bursty topic word extraction. The results of Methods 3 to 5 demonstrate the effectiveness of hashtag rate, document frequency, and word frequency calculation methods in this method. Comparing precision shows that word frequency contributes most to improving precision, followed by document frequency and hashtag rate. From the recall perspective, document frequency can enhance a word' s importance and make it more likely to be selected as a bursty topic word. Overall, Method 1, which combines these four features, clearly achieves the best experimental results.

5.3 Emergency Event Detection Experiments and Results Comparison Analysis This experiment verifies the feasibility of the emergency event detection algorithm by dividing different time windows and using the four-feature combination calculation method to extract bursty topic words within time windows. This part of the experiment selects agglomerative hierarchical clustering algorithm and K-means algorithm for clustering and comparison. Using "day" as the time window and taking data from February 27, 2014 as an example, the number of clusters was set to 5, 10, 15, and 20 for experiments. Precision was used to evaluate the emergency event detection effect based on clustering methods. The final statistical results are shown in Figure 3 [Figure 3: see original paper].

According to the experimental results, in terms of emergency event detection algorithm, agglomerative hierarchical clustering outperforms K-means. When the number of clusters K is set to 10, the accuracy rate of emergency event detection reaches 80%, obtaining a relatively optimal result. As K increases, the accuracy rate decreases. Analysis of experimental data and results reveals that an emergency event may involve multiple aspects of content. When K is excessively large, an emergency event may be divided into multiple fine-grained side information pieces related to the event but unable to become independent events, and these side information clusters may not be emergency events.

The experiment achieved good results, which the authors believe mainly stem from the following aspects:

(1) Addressing the issue that bursty features are easily affected by noise data in microblogs, we studied noise processing methods on microblogs.

Through traditional/simplified Chinese conversion, filtering of @XXX symbols, URL links, and emoticon symbols, part-of-speech filtering, and stopword processing, we effectively improved microblog text quality, providing better data support for emergency event detection.

(2) Fully utilizing the bursty patterns of bursty features and combining microblog' s inherent Hashtag attribute, we proposed a dynamic threshold-based bursty topic word extraction algorithm. Selecting word frequency, document frequency, hashtag, and word frequency growth rate calculation methods, the designed extraction algorithm can effectively screen high-quality bursty topic words with both thematic expressiveness and bursty characteristics from a large number of words. Moreover, dynamically adjusting the word frequency growth rate threshold can obtain different numbers of bursty topic words, thereby affecting the number of detected emergency events.

(3) Based on bursty topic words, we represented microblog texts as feature vectors, combined with set filtering strategies to retain effective microblog vectors, used Jaccard similarity coefficient to calculate similarity between microblog feature vectors, better reflecting the similarity between emergency event texts. On this basis, we used agglomerative hierarchical clustering to achieve emergency event detection, ensuring the feasibility and effectiveness of emergency event detection.

6. Conclusion

This study takes microblog as the research platform, conducts research on emergency event detection, designs and implements a bursty feature-centered emergency event detection method combined with microblog characteristics, and performs bursty topic word extraction and emergency event detection experiments, achieving high precision. Due to limitations in corpus data and research scope, the study has not yet achieved automatic description of detected emergencies, and the analysis and consideration of elements such as netizens' user sentiment features and semantic relationships between events, which are important for emergency event detection, remain inadequate. Therefore, future work will further attempt to combine netizens' user sentiment features and semantic relationships between events to assist emergency event detection research on microblog platforms, expecting to obtain better detection results.

References

[1] Wang X, Zhai C X, Hu X, et al. Mining Correlated Bursty Topic Patterns from Coordinated Text Streams[C]. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007: 784-793.
[2] Du Y, Wu W, He Y, et al. Microblog Bursty Feature Detection Based on Dynamics Model [C]. In: Proceedings of 2012 International Conference on

Systems and Informatics (ICSAI). IEEE, 2012: 2304-2308.
[3] Aggarwal C C, Zhai C X. A Survey of Text Clustering Algorithms [A]. //Mining Text Data [M]. Springer US, 2012:
[4] Yang Y, Pierce T, Carbonell J. A Study of Retrospective and On-line Event Detection[C]. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1998:
[5] Petrović S, Osborne M, Lavrenko V. Streaming First Story Detection with Application to Twitter[C]. In: Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 181-189.
[6] Phuvipadawat S, Murata T. Breaking News Detection and Tracking in Twitter [C]. In: Proceedings of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE, 2010:
[7] Ge Gaofei. Design and Implementation of New Topic Detection and Tracking of Microblog Based on Emergency [D]. Beijing: Beijing University of Posts and Telecomunications, 2014.
[8] Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter[C]. Proceedings of the 5th International AAAI Conference on Weblogs and Social Media. 2011: 438-441.
[9] Du Y, He Y, Tian Y, et al. Microblog Bursty Topic Detection Based on User Relationship [C]. In: Proceedings of 2011 6th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2011:
[10] Benevenuto F, Magno G, Rodrigues T, et al. Detecting Spammers on Twitter In: Proceedings of Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS). 2010(6): 12-20.
[11] Weng J, Lee B S. Event Detection in Twitter [C]. In: Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain. 2011:
[12] Kleinberg J. Bursty and Hierarchical Structure in Streams [J]. Data Mining and Knowledge Discovery, 2003, 7(4): 373-397.
[13] He Q, Chang K, Lim E P. Analyzing Feature Trajectories for Event Detection [C]. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2007: 207-214.
[14] Mathioudakis M, Koudas N. Twittermonitor: Trend Detection over the Twitter Stream[C]. In: Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. ACM, 2010: 1155-1158.
[15] Long R, Wang H, Chen Y, et al. Towards Effective Event Detection, Tracking and Summarization on Microblog Data [A]. //Web-Age Information Management [M]. Springer Berlin Heidelberg, 2011: 652-663.
[16] Zhao Wenqing, Hou Xiaoke. News Topic Recognition of Chinese Microblog Based on Word Co-occurrence Graph [J]. CAAI Transactions on Intelligent Systems, 2012, 7(5): 444-449.

[17] Yao J, Cui B, Huang Y, et al. Bursty Event Detection from Collaborative Tags [J]. World Wide Web, 2012, 15(2):
[18] Wang Yong, Xiao Shibin, Guo Yixiu, et al. Research on Chinese Micro-blog Bursty Topics Detection [J]. New Technology of Library and Information Service, 2013 (2): 57-62.
[19] Guo Yixiu, Lv Xueqiang, Li Zhuo. Bursty Topics Detection Approach on Chinese Microblog Based on Burst Words Clustering [J]. Journal of Computer Applications, 2014, 34(2): 486-490, 505.
[20] Small T A. What the Hashtag? A Content Analysis of Canadian Politics on Twitter [J]. Information, Communication & Society, 2011, 14(6): 872-895.
[21] Zhang Zhiying. Research on Hot Event Detection in Micro-blog Based on Topic Model and Community Discovery [D]. Chongqing: Southwest University, 2014.
[22] National Languages Committee. The Common Standard Chinese Characters Table [K]. 2013.08. http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm.
[23] NLPIR Chinese Word Segmentation System [CP/OL]. http://ictclas.nlpir.org/downloads.

Author Contributions

Ding Shengchun: Topic selection, research ideas and methods, paper revision.
Gong Silan: Responsible for data collection, paper drafting and revision.
Li Hongmei: Data collection and processing analysis, paper drafting.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data [1] can be found in the journal' s online version at http://www.infotech.ac.cn; supporting data [2-4] are self-archived by the authors, E-mail: njustgsl@163.com.
[1] Gong Silan, Li Hongmei. data_results.rar. Data processing results.
[2] Gong Silan, Li Hongmei. Weibo_crawler.py. Microblog corpus crawling program.
[3] Gong Silan. weibo_0225-0311.sql. Original microblog corpus.
[4] Li Hongmei. weibo_process.jar. Corpus processing program.

EBSCO Information Services Assists Global Researchers in Studying the Belt and Road Trade Initiative Across Multiple Regions

EBSCO has recently launched an authoritative international Belt and Road reference resource database, collecting journals and publications from over 60 countries. This Belt and Road reference resource database helps researchers

better understand the cultural and economic conditions of countries along the Belt and Road and discover new trade opportunities.

The Belt and Road Initiative is a trade and economic growth strategy proposed by the People' s Republic of China. The initiative aims to further connect mainland China with Western European trade partners along the Belt (the "New Silk Road") through the development of the Maritime Silk Road. The upcoming maritime improvement strategies include new freight infrastructure and regional port construction to support overseas shipping initiatives.

EBSCO' s Belt and Road reference resource database provides over 5,300 full-text journals, including many hard-to-find local publications from Belt and Road countries. The database also includes nearly 65 full-text newspapers and over 270 reports and conference proceedings. Building this database is another measure of EBSCO' s commitment to global academic research. By providing high-quality content, the Belt and Road reference resource database can offer researchers both global and local perspectives in a multi-country environment.

The database covers multidisciplinary content with a wide range of sources, including: *Journal of Architecture and Civil Engineering* (China), *Journal of Surveying, Construction and Property* (Malaysia), *Educational Sciences* (Turkey), *Journal of Theoretical and Applied Information Technology* (Pakistan), *Biomedical Chemistry* (Russia), *GSTF Journal of Mathematics, Statistics and Operations Research* (Singapore), *China Economist* (China), and many others.

For more information about the Belt and Road reference resource database, please visit: https://www.ebscohost.com/academic/one-belt-one-road-reference-source.
(Compiled from: https://www.ebsco.com/news-center/press-releases/ebsco-information-services-helps-global-researchers-prepare-for-the-one-belt)
(Journal News)

Note: Figure translations are in progress. See original paper for figures.