Microblog Emergency Detection Model Based on Big Data Cluster Analysis

Yu Wang*, Xiangming Kong, Nan Wang

Beijing Polytechnic, Beijing, 100176, China

*Corresponding author e-mail:17805006@qq.com

Abstract. This paper studies the microblog emergency detection model based on the big data cluster analysis method. A large number of users and relatively free speech information make the microblog a powerful tool that significantly influences society. Using the crawler designed for the page information structure of the microblog platform, the detection of emergencies in this paper takes the detection of event subject words as the clue. Firstly, the eigenvalues and data organization methods suitable for the microblog corpus are selected. Subsequently, the feature trajectories of each word in the time window are constructed. Combined with the time-domain and frequency-domain characteristics of the feature trajectories, the burst of words is determined. A nonlinear model of microblog attention is established, the data in key microblog articles is mined, and the correlation coefficient is determined. Additionally, we conducted the simulated experimental detection on the model. The result showed a significant improvement in clustering time cost and an improved detection efficiency of burst time.

Keywords: Big data; Cluster analysis method; Microblog emergencies; Detection model.

1 Introduction

The microblog has attracted a strong user group because of its convenient release, timeliness, and strong interaction. At the same time, it also reflects high research value due to its wide distribution of participants, a huge amount of information, and a high degree of freedom of speech. The research based on the microblog at home and abroad is mainly divided into three aspects: the research on the statistical nature and social networking of the microblog, the performance of the microblog in special fields, and the application research and development based on the microblog platform.

Microblog emergency monitoring refers to the integration of Internet information collection technology and intelligent information processing technology to realize users' information needs such as online microblog emergency monitoring and news topic tracking. These are achieved through automatic capture, automatic classification and clustering, topic monitoring, and topic focus of massive Internet information, forming analysis results such as briefings, reports, and charts. This provides an ana

M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_14

lytical basis for customers to fully grasp the ideological trends of the masses, offering correct public opinion guidance and providing an analysis basis [1].

In short, microblog emergency monitoring is to make comprehensive monitoring around the development or change of existing social events in a specific social space, and many social groups will participate in it. Now microblog emergency monitoring is widely used. With the presence of AI, when doing microblog emergency monitoring, many people may consider what the characteristics of these microblog emergency monitoring are.

The adoption of big data, artificial intelligence, and other technologies to collect, store, study, and analyze massive amounts of Internet information, provides enterprises, brands, governments, enterprises, and institutions with the discovery and analysis services of various microblog emergencies. These emergencies include social conditions and public opinion, livelihood hotspots, or emergencies. Therefore, users can accurately, timely, and comprehensively grasp the Internet microblog emergency information related to themselves. This has improved the ability of the microblog to guide emergencies and resolve contradictions in time.

Recommend this microblog emergency monitoring manufacturer level data (ii. pro), no omission of key microblog emergencies, timely delivery of multi-terminal push, and minute level capture and push of microblog emergencies. Microblog emergency trend development, tracing to the source, emotional analysis, geographical distribution, and other 10 dimensions to comprehensively mine the data value. One million high-quality sources, covering microblog, WeChat, websites, pictures, small videos, microblog emergency retrieval by emergency, etc. Support microblog source type, region, and emotion [2]. Target microblog emergencies are accurately located in massive information, and the accuracy of garbage filtering and emotion analysis is up to 90%.

Emergency refers to natural disasters, accidents and disasters, public health events, and social security incidents that occur suddenly, cause or may cause serious social harm, and need to take emergency measures to deal with them [12]. Table 1 presents the national emergency warning information for the fourth and the third quarter of 2023.

Table 1. Monthly report of national emergency early warning information

Time	Number of social security incidents	Total number of information
In December, 2023	1	34106
In November, 2023	2	21543
In October, 2023	0	12898
In September, 2023	27	26525
In August, 2023	8	60266
In July, 2023	14	68821

From the above table, there is relatively little early-warning information on social security incidents, which is the weak link in detection. With the development of social networks, it is of great significance to detect emergencies in time.

2 Related Work

2.1 Cluster analysis

The common methods of data analysis using data mining mainly include classification, regression analysis, clustering, association rules, feature, change and deviation analysis, web page mining, and so on. They mine data from different angles.

Classification is to find out the common characteristics of a group of data objects in the database and divide them into different classes according to the classification mode. Its purpose is to map the data items in the database to a given category through the classification model. It can be applied to customer classification, customer attribute and characteristic analysis, customer satisfaction analysis, customer purchase trend prediction, etc. For example, an auto retailer divides customers into different categories according to their preferences for cars, so that marketers can mail the advertising manual of new cars directly to customers with such preferences, which greatly increases business opportunities.

$$\left\|\Delta x_{k+1}(t)\right\| \leq k_f \int_0^t \left\|\Delta x_{k+1}(t)\right\| \tag{1}$$

The regression analysis method reflects the time characteristics of attribute values in the transaction database, generates a function that maps data items to a real value prediction variable, and finds the dependence between variables or attributes [3]. Its main research problems include the trend characteristics of data series, the prediction of data series, and the correlation between data. It can be applied to all aspects of marketing, such as customer seeking, maintaining and preventing customer loss activities, product life cycle analysis, sales trend prediction, and targeted promotion activities

Cluster analysis can automatically classify the data of a batch of samples (or variables) according to their many characteristics, according to the degree of affinity in nature (the overall difference in the value of each variable) without prior knowledge (no prior specified classification criteria), generating multiple classification results [10]. Individuals within classes share similarities in characteristics, with large differences between different classes. Its purpose is to map the data items in the database to a given category through the classification model. The shortest and longest distances of interclass distances are shown in Figure 1.

$$\mathrm{E(t)} \in \mathrm{Rn{\times}n,B(t)} \in \mathrm{Rn{\times}m,C(t)} \in \mathrm{Rr{\times}n} \tag{2}$$

$$\sup_{0 \leq t \leq T} \left\|d_k(t)\right\| \leq \varphi \tag{3}$$

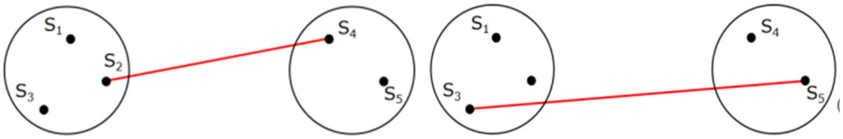

Fig. 1. Shortest and longest distances in interclass distances for cluster analysis

Association analysis uses an indicator to indicate the closeness of the interdependence between the phenomena. The coefficient used to measure the simple linear correlation is the Pearson simple correlation coefficient. The classification of the correlations is shown in Figure 2.

Feature analysis is to extract the feature expressions of these data from a set of data in the database, which express the overall characteristics of the data set.

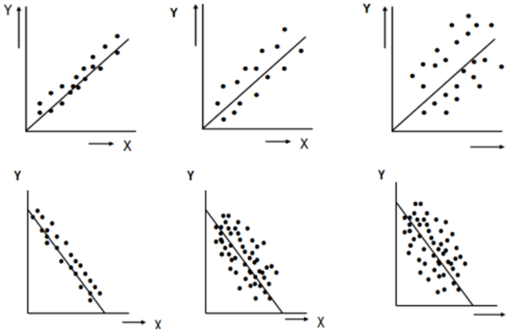

Fig. 2. Association analysis scatter plot (positive and negative correlation)

2.2 Information definition of microblog

Microblog contains various information such as users and microblog items, as well as hot information pushed by microblog platforms such as leaderboards. Each aspect of information contains a lot of detailed information. This section defines all kinds of information and details [4].

User information is divided into user identification information and user basic information. User identification information refers to the information that can identify the user, such as user name and user number or user link. User link is not only the unique mark of the user but also a direct interface to find the user information. The user number is the same as the user domain name for non-real name authenticated users. When the user is a real name authentication user, the user's domain name is the full spelling of the real name, and the user can also jump to the user page through the user number. There are many contents of users' basic information, but the data are uneven. The more unified one is the user city, followed by whether the user's real name authenticates the user and whether it contains a personal description. User activity includes "attention", "fans", and "microblog", and each part has a number and list of information.

Hot information. There are three kinds of hot keywords, hot users, and hot microblogs. The hot keyword is a list with a small length (usually 8), which provides a link to the microblog item corresponding to the keyword. Hot user information includes recommended real-name users and user ranking lists sorted according to different indicators such as attention and number of microblogs. These lists mainly provide user identification information. Hot microblog item information is a ranking list sorted according to the number of forwarding or comments within a certain time. The microblog item information in the ranking list is complete [5].

3 Research on Microblog Emergency Detection Model Based on Big Data Cluster Analysis Method

Hot information includes hot keyword information, hot user information, and hot microblog item information. The hot keyword information is shown as a keyword ranking list [11]. The hot user information includes recommended users and a user ranking list sorted according to different indicators such as attention and microblog number. The hot microblog item information is based on a ranking list sorted by the number of forwarding or comments within a given time. There are a large number of hot users and hot microblog item leaderboards, but each type of leaderboard information is centrally displayed on the web page. The extraction of hot information first filters out the node where the information block is located, and then accurately extracts the target information. The hot users and hot microblog item leaderboards filter the corresponding nodes [6].

The task of overall information extraction is to comprehensively extract the user and microblog item information of the microblog platform. Similar to the extraction of hot information, the overall information extraction also adopts the accurate information extraction method based on page format. The difference is that the overall amount of information is very large, and the page where the information is located can no longer be obtained by enumeration. Users in the microblog are connected to each other through the relationship of “attention” and "being concerned", forming a network structure. A user can access many users, and each microblog item must belong to a user, through which the user can access the microblog item. Therefore, the user list is an important clue in the overall information extraction process. In addition, the comprehensiveness of the overall information extraction depends on the simulated login process. Before the overall information extraction, log in to the microblog website as a user [7].

The data storage part adopts the MySQL database. MySQL is a small relational database management system. Compared with Oracle, DB2, SQL Server, and other large databases, it has weak functions but high portability, good operation efficiency, and rich network support. The data table design in this section mostly meets the third paradigm, that is, there is no transfer function dependence of non-key fields on any candidate key fields. There are two kinds of information subjects in microblog, namely, user information and microblog item information. However, there is another kind of implicit important information, that is, user relationship information, which includes the relationship of "concern" and "being concerned" contained in user information. The "forward" and "reply" relationships contained in microblog items are very important in the process of analyzing the information dissemination in the microblog. In addition, the microblog platform also provides popular information covering user information, microblog item information, and keyword information, which can be used as the source of information extraction, and can also preliminarily test the experimental results.

The propagation attributes of microblog articles are related to emergencies, a nonlinear model of microblog attention is established, the data in key microblog articles is mined, and the correlation coefficient is determined. The model is shown in Figure 3.

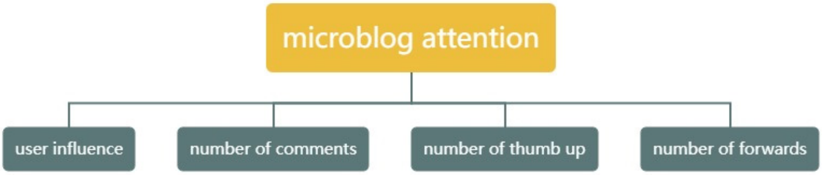

Fig. 3. Non-linear model of microblog attention

The data used in Table 2 are 300 key microblogs extracted manually. By calculating the correlation between user influence, number of comments, number of thumbs up, and number of forwards, the characteristics affecting the attention of microblogs are analyzed.

Table 2. Correlation analysis of the attention of microblog in emergencies

characteristic	mean value	standard deviation	relativity	User influence	number of comments	number of thumb-up	Forward the number
User influence	6.549	0.893	Pearson relativity	1	0.093	0.0236	0.204
number of comments	23483.531	106321.693	Pearson relativity	0.093	1	0.522	0.652
number of thumb-up	34500.142	15283.695	Pearson relativity	0.0236	0.522	1	0.807
number of forwards	7943.235	62728.399	Pearson relativity	0.204	0.652	0.807	1

As shown in Table 1, in terms of attention to the microblog in emergencies, the number of comments on the microblog, the number of likes, and the number of comments showed a strong correlation. Here we propose the calculation method of attention on the microblog as shown in Formula (4).

$$key(w_i) = \alpha \times \log_3(x_{1i}+1) + \beta \times \log_3(x_{2i}+1) + \gamma \times \log_2(x_{3i}+1) + \delta \times x_{4i} \tag{4}$$

α、β、γ、δ are coefficients, x_{1i}、x_{2i}、x_{3i}、x_{4i} are user influence, number of comments, number of thumbs up, and number of forwards, respectively. $key(w_i)$ is the attention of the microblog, The greater the value is, the more important it is, and the more convincing its content is. Therefore, the accuracy of the key microblog calculation model depends on the threshold θ of $key(w_i)$ and the values of the regulatory parameters α、β、γ、δ.

4 Simulation Analysis

The data table reads the microblog item text within the time window, and the burst word detection should be in the valuable microblog item text. Therefore, microblog items with small information content are filtered out. For example, the microblog item body is "forward microblog", which is the default publishing information provided by the system when users directly forward microblog items, or the microblog item body is composed of a continuous forwarding relationship [8]. The approximate judgment method of continuous forwarding relationship is to divide the microblog item body into several segments. If each segment contains a forwarding relationship, It is considered that there is a continuous forwarding relationship in the text of microblog items.

Organize the text of microblog items into text blocks with days as time units, match each text block backward to the maximum, and establish an inverted index after word segmentation: Word -{microblog item serial number{word order number}}the number of microblog items containing words, then filter non-Chinese information and stop word information, filter non-Chinese information, and retain the characters in the Chinese coding area according to the Chinese coding range in GBK. Otherwise, it will be filtered out. Stop words will be filtered according to the frequency of words. If the word frequency is higher than a certain value, it will be filtered (as shown in Figure 4 below).

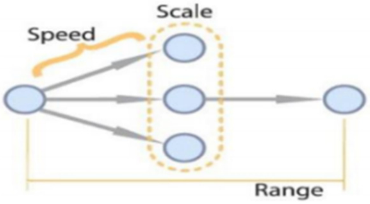

Fig. 4. Filter if the word frequency is a certain value.

Word filtering helps to eliminate the noise of sudden word detection and improves the efficiency of new word discovery. The purpose of word filtering after establishing an inverted index is to ensure the accuracy of word serial numbers so as not to affect the process of new word recognition. The process of new word recognition is based on the word string "shredded" by the dictionary According to the frequency of adjacent words, the frequency value of adjacent words is compared with the frequency of two words. The higher ratio and the lower ratio must meet a certain threshold [9]. In addition, if the word under merger investigation is a new word, its corresponding merger threshold will rise, and the appropriate threshold will be taken.

The programming language is Python. K-means Cluster analysis of potential incident data sets for incident detection. The experiment collected Weibo data from January 16, 2024, to January 17, 2024, one day apart with nearly 80,000 records. Then parsed useful data items for later analysis and stored them in Mongodb. This contains event information and is shown in Table 3. Here are Python code snippets for collecting Weibo data:

First, define a request header to simulate web browser access

```
# request header
headers = {
    "User-Agent": "Mozilla/5.0 (Linux; Android 6.0; Nexus 5 Build/MRA58N) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/99.0.4844.51 Mobile Safari/537.36",
    "accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,image/apng,*/*;q=0.8,application/signed-exchange;v=b3;q=0.9",
    "accept-encoding": "gzip, deflate, br",
}
```

Then send request

```
# request url
url = 'https://m.weibo.cn/api/container/getIndex'
# request parameters
params = {
    "containerid": "100103type=1",
    "page_type": "searchall",
    "page": 1
}
# send request
r = requests.get(url, headers=headers, params=params)
```

Next parse returned data

```
# parse json data
cards = r.json()["data"]["cards"]
```

Use jasonpath to parse data items needed in later analysis:

```
# reposts count
reposts_count_list = jsonpath(cards, '$..mblog.reposts_count')
# comments count
comments_count_list = jsonpath(cards, '$..mblog.comments_count')
…
```

Combine all of data items to form a whole record

```
# dataframe data
df = pd.DataFrame(
    {
        'Weibo id': id_list,
        'Author': author_list,
        'Create time': time_list,
        'Text': text2_list,
        'Reposts': reposts_count_list,
        'Comments': comments_count_list,
        'Attitudes': attitudes_count_list,
        …
    }
)
```

The storage format of microblog is as follows:

```
{
"Weibo id" : "5b63a8b4bf7a7d40fc7d9cbf",
"Author" : "1195379710",
"Create time" : "2024-01-17 19:31",
"Text" : "#Yang Zi new drama net name#",
"Reposts" : 4302,
"Comments" : 4366,
"Attitudes" : 57413,
… }
```

Table 3. Event Information

Number	Event Information
1	The girl who suffered a bench fracture has voluntarily quit school
2	Shenyang zhan is changed back to Shenyang zhanzhan
3	Use Chinese poetry and painting to divide the local cultural travel
4	The boy learns to cook is a great success!
5	Students said they did not want to graduate with fake results
6	Yang Zi new drama net name

The experimental set time interval was one day, and Euclidean distances were clustered as a metric of inter-text similarity, to form the final set of events. In this paper, 1000 key microblog and randomly selected ordinary microblog data are taken as the training set to train the key microblog computing model, and calculate the regulation parameters α, β, γ, δ of the model and the threshold of microblog attention θ. Calculated by simulation, when the adjustment parameters α, β, γ, δ are set to 0.1,0.1,0.7,0.1, respectively, a large differentiation degree can be obtained, and the average value of key microblogs (9.23) is greater than the maximum value of ordinary microblogs (7.26), so the threshold of attention $\theta = 9.23$. For the data of the 18th, a total of 312 key microblogs. The corresponding user influence and microblog attention are shown in Table 4. The sudden word cloud is shown in Figure 5.

Table 4. Calculation parameters of key microblog posts

number	user name	user influence	thumb up number	Comment on the number	Forward the number	microblog attention
1	CCTV News	8.036	1863	1457	2588	10.190
2	Cover News	7.536	3398	8170	2376	10.228
3	Upstream News	7.439	1803	1463	1862	10.097
4	Southern Weekend	7.891	4102	7356	1699	10.126
5	Xinyue Entertainment	6.878	2043	4673	1356	9.998

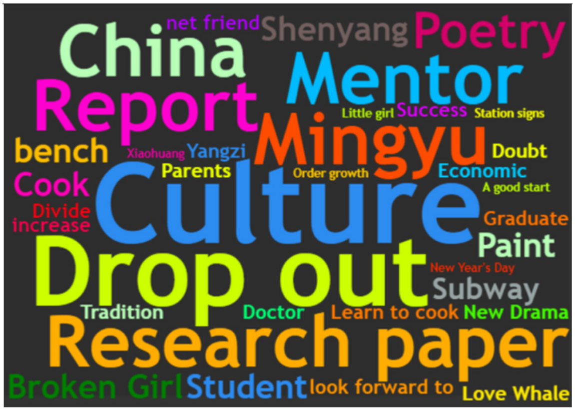

Fig. 5. Sudden word cloud diagram

We validated the effectiveness of the proposed method with the three commonly used event detection methods: Detection method based on blog TFIDF feature clustering, detection methods based on text-based motif discovery LDA, and detection methods based on burst words BBW. Comparisons were made in terms of accuracy, recall rate, and time cost. The accuracy are 0.67, 0.50, 0.56, and 0.62 respectively; the recall rates are 0.86, 0.57, 0.72, 0.75 respectively; cluster time (S) are 68, 308, 178, and 202 respectively. The method proposed in this paper combines the advantages of text-based and burst-based word event detection and optimizes the detection time while ensuring the event detection accuracy and recall rate. By considering the key microblog posts, we have improved the efficiency of event detection, with a significant improvement in clustering time cost.

Through the microblog attention division, the user attribute information and user behavior information are used to build the key text computing model of Weibo. Through the influence division, the burst words in the key posts of the blog are extracted, and the burst words are extracted with the burst characteristics of the events to detect emergencies in the microblog. The experiment shows that considering the influence of the model can effectively reduce the interference of noise data to the detection results and improve the detection performance.

5 Conclusion

This paper proposes a method to detect microblog topic emergencies. Compared with traditional methods, combined with user influence, this paper proposes a sudden word extraction algorithm, which has more practical significance and can describe an event more clearly. The attractive agglomerative hierarchical clustering algorithm adopts the word similarity calculation method based on word cooccurrence in the clustering process, which solves the problem that words belonging to the same event but with low semantic similarity cannot be clustered, and the clustering nestles are identical.

References

1. Qiu Y F A, Cheng L B. Research on Sudden Topic Detection Method for Microblog[J]. Computer Engineering, 2012, 38(9): 288-290.
2. Comito C, Forestiero A, Pizzuti C. Bursty Event Detection in Twitter Streams[J]. ACM Transactions on Knowledge Discovery from Data, 2019, 13(4): 41.
3. Lu X, Yu Z, Guo B, et al. Trending Words Based Event Detection in Sina Weibo[C]. Proceedings of the 2014 International Conference on Big Data Science and Computing, 2014: 4.
4. Alsaedi N, Burnap P, Rana O. Can we predict a riot Disruptive event detection using Twitter[J]. ACM Transactions on Internet Technology (TOIT), 2017, 17(2): 1-26.
5. Zhong Zhaoman, Guan Yan, Li Cunhua, etc. Top-k incident detection in Weibo network region [J]. Journal of Computer Science, 2018,41 (07): 1504-1516.
6. Zhang Lezhong. Research on sudden event detection method based on social network [D]. University of TC, 2019
7. Yang S-F, Rayz J T. An event detection approach based on Twitter hashtags[J]. arXiv preprint arXiv:1804.11243, 2018.
8. Zhou L, Du J, Cui W, et al. Discovering Bursty Events Based on Enhanced Bursty Term Detection[C], 2020: 656-663.
9. Soares V H A, Campello R J, Nourashrafeddin S, et al. Combining semantic and term frequency similarities for text clustering[J]. Knowledge and Information Systems, 2019, 61(3): 1485-1516.
10. Huan Z, Pengzhou Z, Zeyang G. K-means Text Dynamic Clustering Algorithm Based on KL Divergence[C]. 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018: 659-663.
11. Xia Y, Tang N, Hussain A, et al. Discriminative Bi-Term Topic Model for Headline-Based Social News Clustering[C]. The Florida AI Research Society, 2015: 311-316.
12. A.L. Hughes, L. Palen. Twitter Adoption and Use in Mass Convergence and Emergency Events[J]. International Journal of Emergency Management. 2009, 6 (3):248-260.