2022 2nd International Signal Processing, Communications and Engineering Management Conference (ISPCEM) | 978-1-6654-9271-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISPCEM57418.2022.00009

Modeling and Analysis of Spatial-Temporal Relationship and Risk Evolution of Emergencies Based on Big Data from Social Media

Shituo Ma
School of Computer Science and Technology
Huazhong University of Science and Technology
Wuhan, Hubei, China
u201914900@hust.edu.cn

Pengkun Li
School of Electronic Information and Communications
Huazhong University of Science and Technology
Wuhan, Hubei, China
605534244@qq.com

Mingxin Yang
School of Computer Science and Technology
Huazhong University of Science and Technology
Wuhan, Hubei, China
1411477833@qq.com

Chulei Sun
School of Journalism and Information Communication
Huazhong University of Science and Technology
Wuhan, Hubei, China
785212931@qq.com

Ran Wang
School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan, Hubei, China
China Philosophy and Social Science Laboratory of Big data and national communication strategy, Ministry of Education, Wuhan, Hubei, China
rex_wang@hust.edu.cn

***Abstract*—There are many kinds of emergencies in the real world, and there is a close connection between different events. However, the current research on the evolution of emergency risk is limited by the volume and period of the data, which makes the conclusions drawn from these studies have certain limitations today. To describe the emergency's temporal and spatial distribution and its casual emergency chain from the perspective of statistics, we obtain a large amount of data in terms of emergency from Sina Weibo, one of the most famous social media in China. The research outcome is expected to contribute to early warning for preventing and controlling emergencies by fully using accumulated, updated, and rich emergency data from social media.**

Keywords—emergencies, emergency chain, data mining, co-occurrence matrix, complex network

I. INTRODUCTION

In real life, emergencies tend to be inter-relative rather than isolated. Then to describe the relevance of different emergencies, the concept of the emergency chain is introduced to model the analysis of emergencies. The modeling of an emergency chain is of great significance for the prevention and control of various emergencies.

Since the concept of the emergency chain was first proposed, some scholars have tried to infer from the geological mechanism or meteorological principles caused by emergencies, while studying the emergency chain based on data and statistical methods is relatively late. Early statistical inference is based on the simple co-occurrence matrix theory and insufficient amount of emergency samples, which mostly occurred in the late Qing Dynasty, the Republic of China, and even the ancient Tang and Song Dynasties. These factors make such analytical methods less suitable for today's emergencies and massive datasets.

With the advent of the information age, emergencies no longer depend on documents to record that are difficult to ensure their reliability such as history books but can be acquired and processed through network platforms such as Weibo. The means of big data mining can better ensure that the data can be fully mined and utilized. In recent years, as data mining research is higher, using data mining methods to extract the mode of complex information items becomes the most effective means of emergency chain modeling.

Currently, data mining based on emergency chain mode mainly depends on the clustering algorithm and group intelligence algorithm but is also limited to a smaller data set, the time complexity of the analysis algorithm, the volume of the data, and the complexity of the dataset's space. In addition, with technologies such as natural language processing, features of emergencies can be excavated from the text on the network.

Aiming at expanding the research method of emergency chain from the original natural disaster to network emergencies including social emergencies, at the same time and space modeling the emergency chain through massive data mining technology, the study topic is chosen. We will collect large amounts of data from the microblog database and filter out the data needed for further statistical analysis. The data source is real and reliable, and this massive data is of high value, which can be used to explore the application effect of some previous research methods in the massive data and to analyze whether there are some discoveries.

This topic will provide some reference for the prevention and control of major social emergencies, and reduce the losses caused by emergencies.

Compared with previous achievements, the main work of this study is as follows:

978-1-6654-9271-3/22/$31.00 ©2022 IEEE
DOI 10.1109/ISPCEM57418.2022.00009

1) Collect massive microblog data, screen out emergencies, analyze and visualize them in a statistical sense, and find their statistical regularities in spatial and temporal distribution utilizing density clustering.

2) From the perspective of a co-occurrence matrix and complex network, further analysis of emergency chain initiation is carried out utilizing graph theory modeling, and causality rather than traditional correlation.

3) From the perspective of statistics, the time difference between the outbreak time and occurrence time of different types of emergencies is analyzed, ultimately discovering the corresponding rules of propagation significance.

Our work's results revealed the regularity of the occurrence of emergencies in a statistical sense and explained the causality of the emergency chain to a certain extent, which has a certain reference significance for the prevention and control of disasters.

II. Related work

A. A review of emergency chain research

In 1987, Guo Zengjian, a Chinese seismologist, first proposed a theoretical concept of an emergency chain: an emergency chain is a phenomenon that a series of disasters occur successively [1]. The emergency chain describes the connection between various disasters in nature. Commonly, secondary emergencies such as landslides and debris flows associated with earthquakes are referred to as the first type of emergency chain, while the earth-air coupling other than the sea-air interaction, which is recognized by the scientific community, is referred to as the second type of emergency chain [2]. In addition to the trigger mechanism of the emergency chain, risk assessment also plays an important role in the risk prevention of the emergency chain. Wang Xiang et al. summarized the emergency chain risk assessment models and conducted a systematic analysis of regional emergency chain risk [3].

In addition, some scholars at home and abroad apply emergency chain research to specific event domains, such as flood disaster research [4], earthquake disaster research [5], power system disaster research [6], etc. This kind of research is of great significance in constructing coping strategies for similar events and reducing economic losses.

B. A review of applications of data mining in emergency chain research

In recent years, the popularity of data mining technology is increasing, and it matters to integrate data mining technology into emergency chain modeling.

Some clustering algorithms can be used to cluster the regions with dense emergence outbreaks to obtain their spatial and temporal characteristics. Among them, clusters based on density only need to consider inter-sample density, only neighborhood radius needed to be set in advance to describe the relationship between samples and clustering clusters, sidestepping the default setting of cluster number [7][8].

Some studies in the past also provide some feasible strategies for spatial and temporal data mining. Ray et al. proposed the concept of concept transfer by slicing the time series, considering the state of each slice, and then determining the change rule of the state over time [9]. In addition, for better processing of spatial-temporal data, the objective function of the K-means clustering algorithm is adjusted and improved to obtain the attribute characteristics of the harmonic K-means clustering algorithm used to describe the spatial-temporal data [10]. Karine et al. considered introducing subscripts of time series elements into the analysis process of space-time series as analysis features [11]. As for the efficient algorithm of spatial-temporal clustering, Vladimir et al. proposed the AUTOCLUST+ algorithm to promote the efficiency of the original spatial-temporal sequence clustering to a higher degree [12]. These studies provide a good method for us to analyze the temporal and spatial distribution of emergencies, but at present, a lack of evaluation of the dynamic changes of clustering results in the context of the whole temporal and spatial space still exists.

Today, data mining about the propagation of emergencies on the network is mainly based on some basic statistical methods. Chen et al. mined the impact and information dissemination of Hurricane Harvey in the United States based on statistical features of spatial-temporal series and text emotion recognition methods [13]. Yao et al. found that forwarding behavior shows aggregation characteristics, and users with different attributes have special forwarding habits on Sina Weibo [14]. Zhong et al. compared support vector machine, random forest, CatBoost, and LightGBM algorithms from the perspective of feature engineering, and mined the important features of emergencies in microblog communication from multiple perspectives such as text and user behavior [15][16][17].

All of these works have modeled and analyzed the emergencies on the network from the perspective of data mining, but the limitations are relatively strong, failing to consider all kinds of emergencies and the mechanism and characteristics of the mutual triggering of emergencies.

III. Emergency's temporal and spatial distribution

A. Data acquisition

We deployed a Web crawler based on the Scrapy framework to get reports of emergencies on Weibo. The collection covers a period of 11 years from 2009 to 2019, including 4 categories of events such as social security events, natural disasters, accidents and catastrophes, and public health events, which can be further divided into 18 subcategories, including criminal cases, geological disasters, and so on. The category relationship is shown in Fig 1.

The data capacity is about 308G (30,000,000 pieces of data). It is difficult to analyze and process such massive data by traditional pure statistical methods. Based on this, the process strategy of data mining is adopted to analyze it.

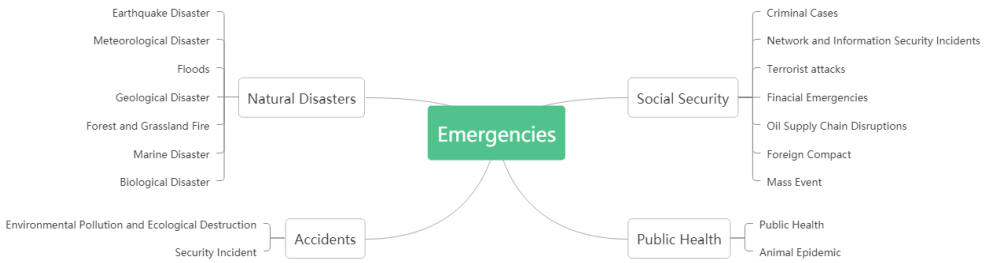


Fig 1. Schematic diagram of classification level of emergency relationship

B. Data preprocessing

First of all, the labels of data to pre-process. The text is classified based on the BERT model commonly used in the NLP field and then annotated based on the categories of emergencies. Secondly, we extracted the time and specific geographical location of the emergency outbreak from the text content based on regular expression and keyword retrieval as the space-time coordinates of the outbreak. The spatial and temporal coordinates of the emergency in the three-dimensional Euclidean space are the most important feature for us to analyze the emergency chain.

Considering that a large part of the data collected is not emergencies, a culling for non-emergencies is necessary. For simplicity and interpretability, entries are filtered by the time of the outbreak. Normally, if the news is not a breaking event, the specific time and place will not be reported. In addition, some non-emergencies also have time and place, so such events are eliminated according to the score of classification. After this step, the amount of data is greatly reduced, and the number of remaining data items is about 833,000. In these 833,000 Weibo data, there are a large number of repeated reports of information, which need to be de-duplicated. The coincidence of outbreak time points is used for further sifting and reserving the items with the earliest reporting time. Finally, about 156,000 pieces of data are screened out, and the skeleton of the data process is shown in Fig 2.

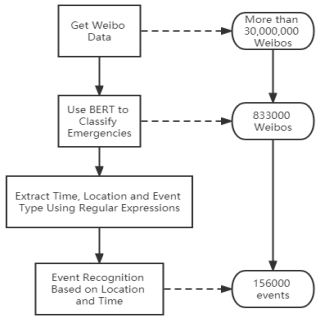


Fig 2. Pre-Process of Data

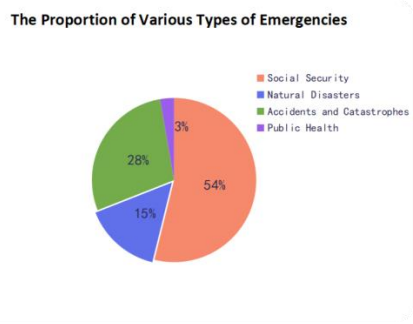


Fig 3. The proportion of each category

C. Statistical Characteristics and Spatio-Temporal Distribution of Emergencies

The proportions of the four categories in the data are 84,285 social security events, 44,214 accidents and catastrophes, 23,764 natural disasters, and 4,281 public health events. The number of items refined by each category and the proportion of each category are shown in Fig 3 and Fig 4. It can be found that social security incidents account for the largest proportion, and criminal cases account for the largest number of all secondary incidents, which shows that social emergencies are far more than natural disasters.

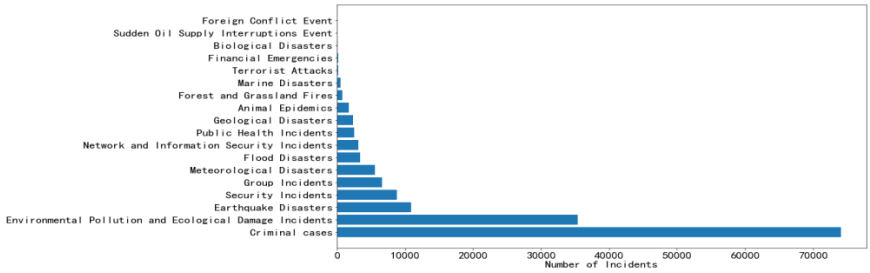
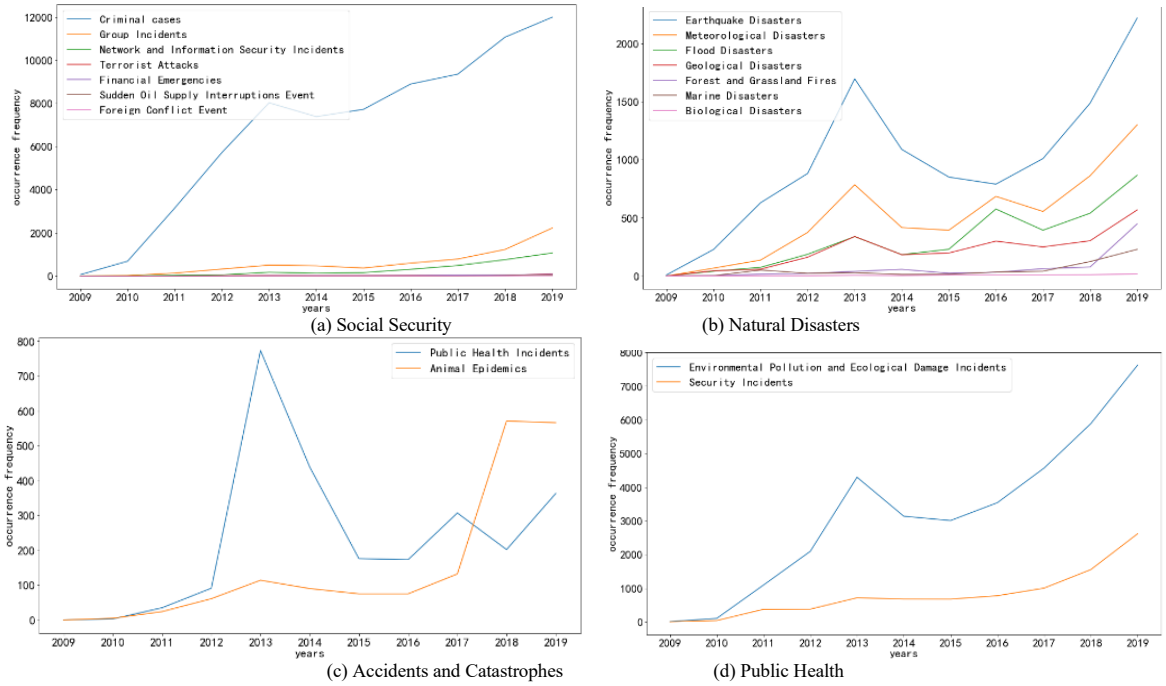


Fig 4. The number of items refined by each category

In order to research the spatial statistical distribution of the four types of emergencies, we plotted the outbreak locations of emergencies on the map of China.

It can be seen that social security events account for the highest proportion and occur the most frequently among all kinds of emergencies. Besides, social security events occur frequently in the “paradise” in our eyes – the areas of Beijing,

Tianjin, Jiangsu, Zhejiang, Shanghai, Guangdong, Hongkong, and Macao. Such a phenomenon generates by the joint force of the dense population attracted by the advanced economy, which is the trigger of social security problems, the regulation burdens, and the higher frequency of contrived incidents compared to natural disasters. The special geographical location and topographical structure make the Yunnan-Guizhou-Sichuan region a high-earthquake-prone area in a geological sense, and the vast forest area and abundant rainfall also lay hidden dangers for natural disasters such as flash floods and forest fires. The outbreak of public health events was the least (In 2019, COVID-19 has not yet broken), however, Zhejiang was the area with a high incidence of public health events. For example, the outbreak of swine fever in Jiaxing, Zhejiang on March 11, 2013, the H7N9 influenza virus in 2013-2014, and the high incidence of rabies in Zhejiang in recent years are all public health events.



(a) Social Security (b) Natural Disasters

(c) Accidents and Catastrophes (d) Public Health

Fig 5. The curve of the change in the number of emergencies over time

As can be seen from Fig 5, even with data deduplication, the total number of emergencies each year still increases exponentially. There were no more than 1000 emergencies in 2009, but they quickly rose over the next decade. For different classes of events, these curves are generally in an upward trend; In absolute terms, as of 2019, social security events far outnumbered natural disasters, followed by accidents and public health. This phenomenon seems to be that emergencies are increasing year by year. In fact, since the data we use is Weibo data, this phenomenon reflects that Weibo and other social media are gradually playing an increasingly important role in media early warning of emergencies. This reflects that emergencies in the information age can be reported more quickly so that users can grasp first-hand information timely and accurately. In recent years, the network public opinion of emergencies has gradually shown the characteristics of rapid diffusion and a large amount of information, and more emergencies have been further reported, but at the same time, it has also increased the difficulty of the dissemination and control of network public opinion [18].

D. Cluster analysis of outbreak points

In order to further mine the regional distribution of emergencies and find the densely distributed regions in the space of emergencies and the evolution of these regions over time, we divided the data according to each day and clustered the data at each time node geographically. Since the transmission of emergencies has a certain geographical continuity and should not be too long or too short, we selected 100 kilometers as the clustering radius and used DBSCAN for density clustering [19,20]. Compared with the K-Means algorithm, DBSCAN does not need to set the K value, and areas with low density will not be included in the clustering results, so the spatial clustering results of the outbreak sites of emergencies can intuitively reflect the spatial density distribution of the outbreak of emergencies. In order to explore the evolution of the areas with the most intensive outbreaks over time, we also tried to cluster the outbreak sites of emergencies by using the DBSCAN clustering algorithm to explore the evolution law of the points with the most intensive outbreaks.

We set the neighborhood radius as 100km (empirical value, which can be adjusted), that is, the connecting radius between two adjacent clusters does not exceed 100km. Considering that the outbreak has a certain time, we slice the data according to the outbreak time.In order to better describe the high-density region, we took the average coordinate of the samples in each cluster as the sample center point to describe the evolution law within a certain range.Several rules can be found from the clustering results:

1) With time passing by, the number of areas where the burst density is concentrated is also increasing. This result is consistent with the problem reflected in Fig 5. As the media plays an increasingly important role in media early warning,

more emergencies can be reported promptly, resulting in an increasing number of incidents and more intensive areas. Although the outbreak areas have been changing, they are concentrated in Central China, North China, East China, South China, and Sichuan. Among them, Guangdong is almost unmoved, and Jiangsu, Zhejiang, Shanghai, Beijing, Tianjin, and Hebei are also regulars. These areas with high density in time and space are also areas with high GDP and population density, so they are prone to social security incidents and accidents.

2) More emergencies occur in summer than in winter. This is because the comfortable weather in summer not only leads to more frequent social activities but also provides a hotbed for natural disasters such as rainstorms, floods, and forest fires. Secondary emergencies caused by these natural disasters follow a specific emergency chain.

IV. Conclusion

This paper analyzes emergencies and their secondary derivative events based on complex network theory and improved co-occurrence matrix model. On the basis of massive social media data crawled, based on natural language processing methods and statistical analysis methods, this paper statistically describes the temporal and spatial distribution of emergencies from 2009 to 2019 and analyzes its statistical characteristics. The temporal and spatial distribution and evolution patterns of events analyzed by massive data mining and modeling are of great significance to the early warning and emergency response of emergencies.

Acknowledgement

This research is founded by National Social Science Fund of China award number(s): 19CXW032. Ran Wang is the corresponding author (rex_wang@hust.edu.cn).

References

[1] Guo Zengjian, Qin Baoyan. A Brief Discussion on Disaster Physics [J]. Disaster Science, 1987(2):30-38.

[2] Gao Jianguo. Research on China's emergency chain [C]// 2007 Sino-US Disaster Prevention Symposium. 2007.

[3] Wang Xiang. Research on regional emergency chain risk assessment [D]. Dalian University of Technology, 2011.

[4] Liu Yongzhi, Tang Wenwen, Zhang Wenting, Zhang Xingnan, Niu Shuai. Overview of flood disaster risk analysis based on emergency chain [J]. Water Resources Protection, 2021, 37(01): 20-27.

[5] Yu Shizhou, Zhang Lingxin, Zhao Zhendong, et al. Probability Analysis of Earthquake emergency chain and Disaster Mitigation Method for Broken Chain [J]. Chinese Journal of Civil Engineering, 2010(S1):479-483.

[6] Sun Baojun. Analysis of Natural emergency chain of Inner Mongolia Power System [J]. Disaster Science, 2020, v.35;No.139(04):10-14+49.

[7] Wang Guihong. Research on density-based clustering algorithm [J]. Journal of Quanzhou Normal University, 2009(02):44-49.

[8] Xia Luning, Jing Jiwu SA-DBSCAN: An Adaptive Density Based Clustering Algorithm [J] Journal of Graduate School of Chinese Academy of Sciences, 2009, 26 (4): 9

[9] Hickey R J , Black M M . Refined Time Stamps for Concept Drift Detection During Mining for Classification Rules [J]. Springer-Verlag, 2000.

[10] Zhang B , Hsu M , Dayal U . K-Harmonic Means - A Spatial Clustering Algorithm with Boosting [M]. Springer Berlin Heidelberg, 2001.

[11] Roddick JF , Hornsby K . [Lecture Notes in Computer Science] Temporal, Spatial, and Spatio-Temporal Data Mining Volume 2007 || Join Indices as a Tool for Spatial Data Mining [J]. 2001, 10.1007/3-540 -45244-3(Chapter 9):105-116.

[12] Estivill-Castro V , Lee I . AUTOCLUST+: Automatic Clustering of Point-Data Sets in the Presence of Obstacles [C]// Temporal, Spatial, & Spatio-temporal Data Mining, First International Workshop Tsdm Lyon, France, September 12, Revised Papers. DBLP, 2001.

[13] Chen S , Mao J , Li G , et al. Uncovering Sentiment and Retweet Patterns of Disaster-related Tweets from a Spatiotemporal Perspective—A Case Study of Hurricane Harvey [J]. Telematics and Informatics, 2019, 47:101326.

[14] Yao, W., Jiao, P., Wang, W. and Sun, Y., “Understanding human reposting patterns on Sina Weibo from a global perspective”, Physica A: Statical Mechanics and its Applications 518, 374-383 (2019)

[15] Prokhorenkova L , Gusev G , Vorobev A , et al. CatBoost: unbiased boosting with categorical features [J]. 2017.

[16] Meng Q . LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2018.

[17] Shengtao Zhong, Rui Sheng, Ran Wang, Yiyao Li. Prediction of the propagation effect of emergencies microblog [P]. International Symposium on Multispectral Image Processing and Pattern Recognition, 2020.

[18] Li Gang, Chen Jinghao. A review of online public opinion research on public emergencies [J]. Library and Information Knowledge, 2014(02): 111-119. DOI: 10.13366/j.dik.2014.02.111.

[19] Rong Qiusheng, Yan Junbiao, Guo Guoqiang. Research and implementation of clustering algorithm based on DBSCAN [J]. Computer Applications, 2004(04):45-46+61.

[20] Hu Qinglin, Ye Nianyu, Zhu Mingfu. A Survey of Clustering Algorithms in Data Mining [J]. Computer and Digital Engineering, 2007(02):17-20+188.