Send Orders for Reprints to reprints@benthamscience.ae

The Open Cybernetics & Systemics Journal, 2014, *8,* 821-828

Open Access

A Multi-Strategy Quantitative Analysis Model for Network Emergency

Yuan Sun* and Wenbin Guo

School of Information Engineering, Minzu University of China, Haidian District, Beijing, 100081, China

Abstract: How to quickly and efficiently get real-time emergency data from a massive network is becoming a major problem in monitoring internet public opinion. In this paper, we propose a clustering algorithm based on simplified cluster hypothesis to improve system efficiency. Meanwhile, by analysis of the emergency's characteristics under the dynamic web environment, we propose a multi-strategy quantitative analysis model for network emergency by researching on the following aspects: (1) Using field feature vector similarity to dividing emergency category. (2) Using Sum Limit method to calculate the emergency's maximum effect range. (3) Using distribution to calculate the emergency's timeliness. Finally, through analysis dynamic data from Tianya.cn, the experimental results prove the model is effective.

Keywords: Emergency, internet public opinion monitoring, quantitative analysis model, text clustering.

1. INTRODUCTION

Chinese society is in a rapid transition at present. During this special period, the political, economic, cultural, and other fields are facing severe social challenges because of the adjustment of social structure. Meanwhile, with the rapid development of internet, especially with the advent of Web 2.0 era, persons are not only the Web content browser, but also its maker. If the dissemination of information on the internet do not get effective control, it is easy to further deteriorate the current social problems and affect social stability.

The Third Plenary Session of the 18th Communist Party of China (CPC) Central Committee has been held from Nov. 9 to 12 in Beijing. The CPC approved a decision on "major issues concerning comprehensively deepening reforms" [1]. One of most important issues is to establish and improve network emergency response mechanisms. "2012 Chinese Internet public opinion analysis report" shows the topic of "Diaoyu Island and anti-Japanese demonstrations" has more than mega posts and forwards on Tianya.cn, SINA Weibo and other SNS platform [2]. Meanwhile, the number of posts that these forum and SNS platform produced every day is far exceed millions of data. How to make quantitative analysis on different types of emergencies under dynamic network environment of huge amounts of data is one of the most problems in current public sentiment work.

Nowadays, some researchers have already carried out many works on the system efficiency improvement and emergency index system, early warning mechanism and detecting methods. These works also have achieved many fruitful results.

In text clustering methods, K-means algorithm [3] and the center point algorithm [4] divide the given n objects or the database of data record into k classes. References [5-7] are primarily based on the "neighbourhood" thought. It will continuously cluster as long as the density (number of objects or data points) in the "neighbourhood" exceeds a certain threshold. Reference [8] is a grid-based approach: the object space is quantized to a finite number of elements to form a grid structure. All cluster operations are conducted on the grid structure. Reference [9] combines linguistic characteristics, defines and extracts "theme elements" and using it for indexing base, with a good clustering effect. In addition, there are also many text clustering method at present [10-12]. These methods have been used in practice successfully.

In terms of emergencies index system, Zeng used AHP method to construct the network opinion emergencies early warning index system including 3 types of factors and phenomena such as warning source, warning sign and warning information [13]. In Reference [14], Ju toke an in-depth discussion in emergency concepts, elements, and feature types. Zhang established index system to measure and evaluate enthusiasm of nonconventional emergency network public opinion, determining deep impact factor in deep public opinion fluctuations and its inner mechanism [15]. In Reference [16], the author used system theory, viewing emergency event as a system and analyzing the generic behavior mode about various events. Then the method for constructing the Bayesian networks model of predicting emergency events is proposed to deal with the uncertainty of the events. Reference [17] proposed a system to detect hot web event automatically. The system is focused on the stream of news report on the Internet, sorting the candidate events by calculating their hot degree. In Reference [18], the author constructed sentiment vectors and built hierarchical structures, monitoring the states of sentiments words to discover burst events and burst periods using improved method. In Reference [19], the author proposed a blog emergent detection method based on temporal distribution.

These methods discuss topics and events in different aspects and provide good reference for us. However, the emer-

1874-110X/14

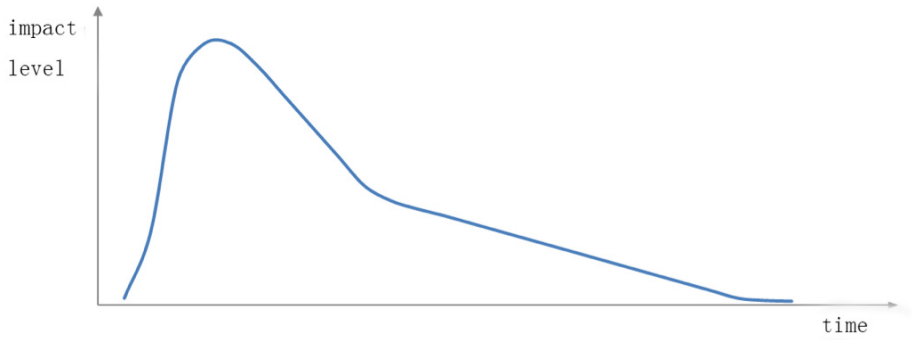

Fig. (1). Network public opinion influence level.

gency types, influence area, timeliness and its quantitative analysis are seldom discussed.

In this paper, we firstly build One-Next text clustering algorithm based on simplify the text clustering hypothesis to improve system efficiency. Secondly, we propose a quantitative analysis model for network emergency to calculate emergency quantitative value. We analysis different aspects (post number, forward number, reply number, *et al.*) which influence emergency outreach.

The rest of this paper is organized as follows. The next section describes the features of network emergency and shows the process of emergency detection. Section 3 introduces a new text clustering method based on simplified hypothesis. The multi-strategy network emergency quantitative analysis model is presented and discussed in section 4. Some experimental results by simulation are presented in section 5. Finally, the conclusion is drawn in section 6.

2. NETWORK EMERGENCY ANALYSIS

Typically, the communication of network event source always follows a fixed mode whatever platform (Micro-blog, forum, SNS, *et al.*) user posted a message on. Emergency life cycle in the network environment is generally divided into the incubation period, the significantly period, height period and receding period [20], as shown in Fig. (**1**).

When emergency occurs in a network, texts associated with this matter will report on the Internet. People will quickly start to pay attention to it within a short time. Factors influencing the trend of public opinion including emergency type, the number of people hearing the related event and especially time that the event occurs. Each factor reflects varying degrees of people's interest in this emergency.

(1) Emergency Type: people have different attitude towards different emergency type. For natural disaster, it is generally viewed with a sympathetic attitude. For terrorist attacks, people may have angry feelings.

(2) Emergency Spread Area: all of network users' operations will leave a mark on the website which can enhance the emergency's diffusion. In network environment, the emergency spread area may include viewing number, replying number and forwarding number.

(3) Emergency Timeliness: emergency impact is huge when it happens, and it will slowly recede as time goes. So time is an important factor that influence emergency.

For example, when "7.21" heavy rain disaster occurred in Beijing. A wide of reports related to it quickly appeared on the Internet. Suddenly, almost all Chinese people pay attention to this emergency. We can see it on every Internet platform. After a period of time, the number of posts began to decline which means that emergency have slowly faded out of people's lives.

In this paper, we establish a network emergency assessment process based on the above characteristics, shown in Fig. (**2**).

(1) Filtering Web data while crawling in real time.

(2) Extracting basic information from the event, including event title, content, release time reviewing number, forwarding number and replying number and save them to the basic properties of the event source database.

(3) Using text cluster method to put events with same theme into the same category.

(4) Using Multi-Strategy network emergency quantitative model to quantify each emergency value.

(5) Sorting every emergency by its value.

3. TEXT CLUSTERING METHOD

3.1. Clustering Hypothetical Problem

Text clustering is based primarily on the famous clustering hypothesis: the same kind of document's similarity is larger, but the different kind of document's similarity is smaller. We call the cluster hypothesis distance relativity hypothesis because it is not only in full consideration of the distance within the class at the same time, but also necessary to consider the distance between classes.

At present, most clustering method are based on distance construction of relativity hypothesis. They adopt iterative relocation technique and try to move between objects in a group to optimize model parameter which can improve division. Take K-means algorithm as an example, the relative distance assumption makes the original text is not a kind of imputation to a class, as shown in Fig. (**3**). Fig. (**3a**) is classes before clustering, Fig. (**3b**) classes is after clustering. According to the relative distance assumption, due to the distance C from A is less than the distance C from B. C is divided into category A, and in fact, C should not be classified into B or A in any category, but should stand alone as a class exists.

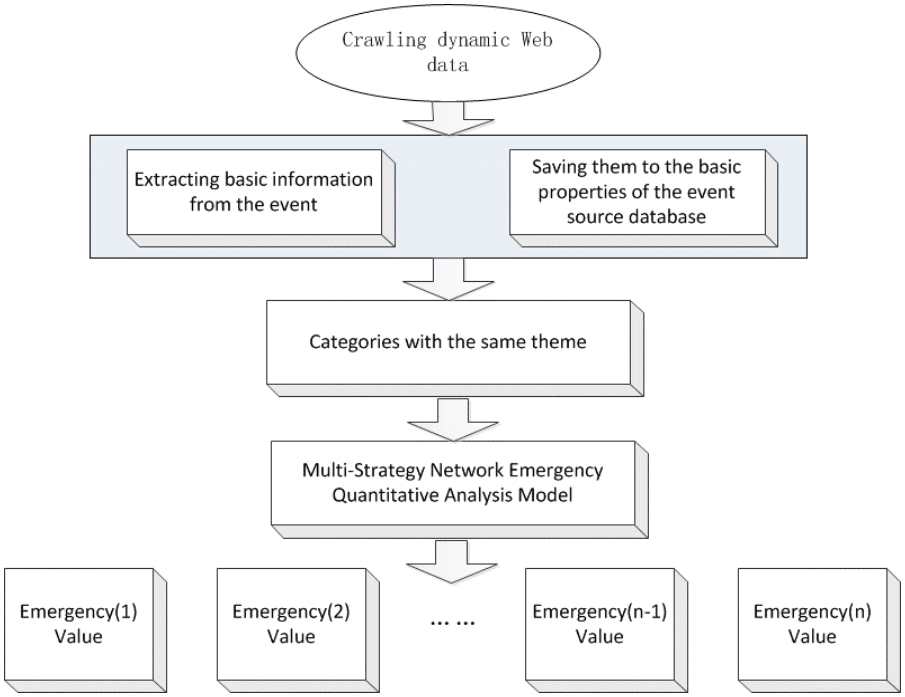

Fig. (2). Multi-strategy network emergency quantitative analysis model.

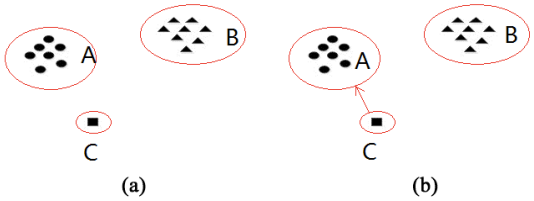

Fig. (3). Three classes clustering.

These algorithms use iterative relocation technique, making the clustering in the process of calculation, continuous counting of objects that have been categorized. This will lead to high time complexity of the calculations, especially when the text feature vector has high dimension, this computational complexity will be greatly enhanced. Fig. (**4**) shows an iterative process of K-means cluster algorithm. We use the black star to represent the k classes center randomly selected. The arrow indicates the process of centroid movement. Each iteration will calculate the distance all point to the new cluster center, complexity of the calculation process. Because the algorithm first select k center point is random, they will often cause a reduction in accuracy of clustering results when the k center points selected are unevenly distributed.

The above phenomenon occurs because the cluster that we assume is positioned "similar documents' similarity is larger, and different kind of documents' similarity is smaller". We believe that it does not need to determine the similarity of the text with other text set when the text is divided into a category. Based on this phenomenon, we simplified cluster hypothesis to "texts with same characteristic have a strong probability to be divided as one cluster".

3.2. One-Next Clustering Algorithm

In this paper, we propose a new clustering algorithm named One-Next text clustering algorithm, which is based on a simplified cluster hypothesis. The basic idea is taking one-time text classification according to the similarity of text to avoid the same text "division-clustering, division again-clustering again" repeated iteration process.

For a collection of texts $D=\{D_1,D_2,\cdots,D_n\}$ that are not clustered, it put the text's initial position D_i as the original center of mass of the class and sequential scan $D_1,D_2,\cdots,D_n$ gradually, judging it is similar distance with each text. If the similarity is greater than d, we put D_i into the same class

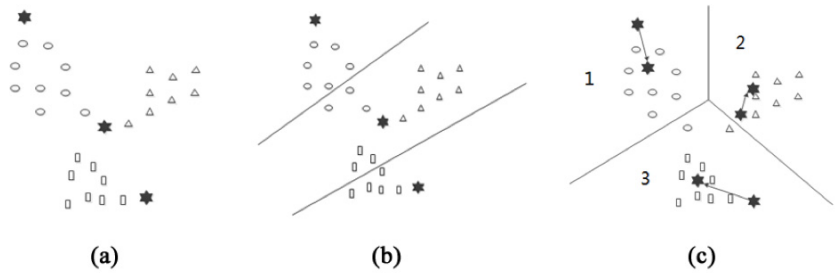

Fig. (4). K-means algorithm process.

Table 1. Network emergency classification.

Category	Subclasses
Natural Disasters	Weather, Forest, Marine, Geology, Earthquake, Flood and drought, Bio.
Accidents	Mining business security incidents, Transport accidents, Nuclear radiation, Environmental pollution and ecological destruction, Public facilities.
Public Health	Epidemic situation of infectious diseases, Mass illness of unknown cause, Food safety and occupational hazards, Animal disease outbreaks, Life safety event.
Social Security	Terrorist attacks, Ethnic and religious affairs, Economic security, Foreign Affairs-related safety, Political security, Group events.

C, then calculate the new center of the mass of the class C. Assuming that class C contains n texts, each of them has m dimensions, so the text feature vector is $v_i = v_{i1}, v_{i2}, \cdots, v_{im}$. The k-th dimensions value of the new class center is shown in (1).

$$w_k = \frac{v_{1k} + v_{2k} + \cdots + v_{nk}}{n} \ (1 \le k \le m) \tag{1}$$

The algorithm will stop until the text collection, which contains the un-classed text is empty.

4. MUTI-STRATEGY EMERGENCY QUANTITATIVE MODEL

Based on the above network emergency analysis and the text clustering results, we mainly discuss the model under the following aspects. Each aspect will be replaced by an emergency value.

4.1. Categories Division Method Based on Field Feature Vector Similarity

There are a lot of topics in the network, some of them are not emergencies. We should select the specific topics from a large number of network events flows. We construct field feature vectors based on the categories divided by "Emergency Response Law of the People's Republic of China", as shown in Table **1**.

Each subclass has a specific feature vector $\vec{c_i} = (c_{i1}, c_{i2}, \cdots, c_{in})$. When a network event X_i comes, we use a feature vector $\vec{v_i} = (v_{i1}, v_{i2}, \cdots, v_{in})$ to represent this event. According to c_i and v_i, we establish emergency type characteristic function $Term(\vec{v_i}, \vec{c_i})$, Which is to determine whether a network event is an emergency:

$$Term(\vec{v}, \vec{c}) = \begin{cases} 1, & sim(\vec{v}, \vec{c}) \ge \lambda \\ 0, & sim(\vec{v}, \vec{c}) < \lambda \end{cases} \tag{2}$$

where λ is similarity regulatory factor of $\vec{c}$ and $\vec{v}$, and $sim(\vec{v}, \vec{c})$ is feature vector similarity of $\vec{c}$ and $\vec{v}$:

$$sim(\vec{v}, \vec{c}) = \cos\theta = \frac{\sum_{k=1}^{n} v_k \times c_k}{\sqrt{(\sum_{k=1}^{n} v_k^2)(\sum_{k=1}^{n} c_k^2)}} \tag{3}$$

4.2. Emergency Maximum Effect Range

Emergency effect range is the number of people who notice this emergency over a period of time.

Assuming $f(t)$ is the rate of message diffusion. *Sum* is emergency effect rage during Δt:

$$Sum = \int_{t1}^{t2} f(t)dt \tag{4}$$

where $\Delta t = t_2 - t_1$, t_1, t_2 is represents start time point and end time point that we set.

In this paper, we use Sum Limit method to calculate *Sum* value. Inserting some split point in $[t_1, t_2]$ to separate it into several small intervals:

$$[t_0, t_1], [t_1, t_2], \cdots, [t_{n-1}, t_n]$$

The length of each interval is:

$$\Delta t_1 = t_1 - t_0, \Delta t_2 = t_2 - t_1, \cdots, \Delta t_n = t_n - t_{n-1}$$

Firstly, we select one point $\xi_i \ (t_{i-1} \le \xi_i \le t_i)$ randomly from every small interval $[t_{i-1}, t_i]$. Secondly, multiplying $f(\xi_i)$ and the length of Δt_i. Finally, evaluating Sum Limit:

$$Sum = \sum_{i=1}^{n} f(\xi_i)\Delta t_i \tag{5}$$

4.3. Emergency Timeliness Based on χ^2 Distribution

In probability theory, the χ^2 distribution with k degrees of freedom is the distribution of a sum the squares of k independent standard normal random variables.

In this paper, we use χ^2 distribution to calculate emergency timeliness:

$$f_k(x) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2} \tag{6}$$

where $\Gamma(k/2)$ denotes the Gamma function, which has closed-form values for integer k. The probability density changes when integer k has different value. $f_k(x)$ denotes emergency impact factor.

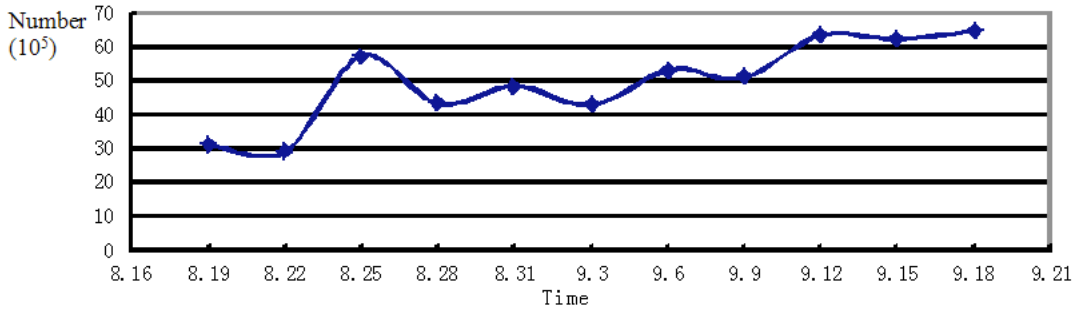

Fig. (5). Data sets of tianya.cn.

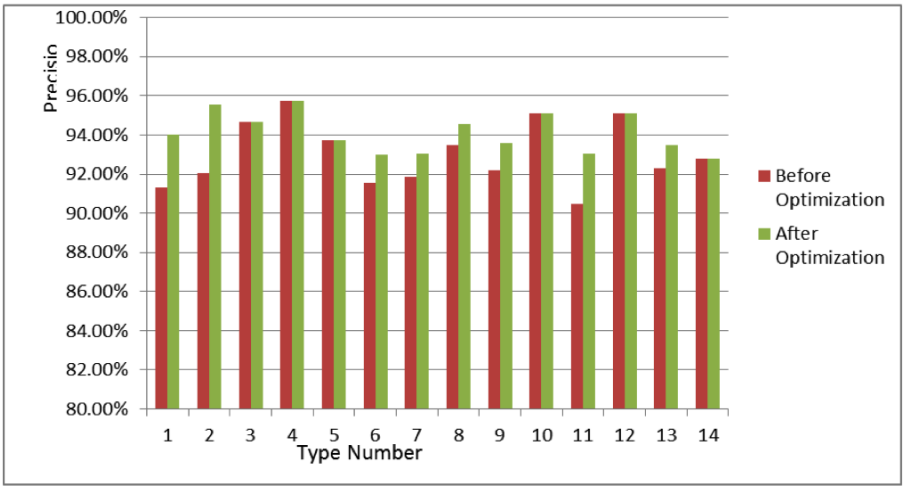

Fig. (6). Comparable precision.

In this paper, we determine integer k value by calculating the slope s:

$$s = \frac{f(t_{i+1}) - f(t_i)}{\Delta t} \tag{7}$$

Based on the above factors, we establish the timeliness function:

$$Time(X) = \alpha f_k(x) \tag{8}$$

where $f_k(x)$ is the probability density value, α denotes χ^2 distribution adjustment coefficient, it is influenced by $f_k(x)$ and $Sum(X)$.

4.4. Muti-Strategy Network Emergency Quantitative Analysis Model

Based on the above analysis, we propose a multi-strategy network emergency quantitative analysis model:

$$Event(X) = Term(\vec{v}, \vec{c}) \times \frac{Sum(X)}{10^6} \times Time(X) \tag{9}$$

(1) $Term(\vec{v}, \vec{c})$ is emergency type feature function.

(2) $Sum(X)$ is emergency maximum effect range. Topically, $Sum(X)$ is a very large number, so we need to reduce its dimensions.

(3) $Time(X)$ is emergency timeliness function.

5. EXPERIMENTAL RESULTS

5.1. Experimental Data Sets

In this paper, we statistic data of Tianya.cn that crawling by web crawler from August 19 to September 18 in 2012, which contains 13.1 million short texts. Fig. (**5**) shows forum data counted by every 3 days as a time point.

5.2. One-Next Clustering Method

In order to improve the clustering results, we use the feature vector grading extraction method to filter feature items, optimizing the feature vector.

Fig. (**6**) shows the comparable precision before and after using the feature vector grading optimization. After using the method, the precision is significantly increased. The optimization results are different because each text cluster has different number of passage and word.

During the experiment, we use the DBSCAN, K-Means as a comparable algorithm to prove the One-Next clustering algorithm has a high time efficiency. The three algorithms' consuming time is shown in Fig. (**7**).

Generally, the time complexity of traditional clustering algorithms is very high. For example, the time complexity of K-Means algorithm is $O(n*k*t)$, where t is iterations, k is the number of clusters, n is the number of clustering texts. While the DBSCAN's time complexity is $O(n^2)$.

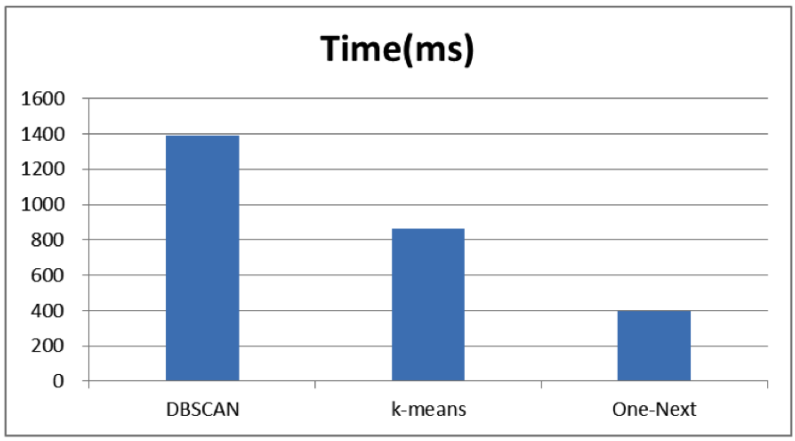

Fig. (7). Comparable time of complexity.

Table 3. Emergencies form.

X	$Term(\vec{v},\vec{c})$		$Sum(X)$	$Event(X)$ **/10^3**
	Value	**Type**	**Value**	**Value**
Diaoyu Island and anti-Japanese demonstration	1	group event	6327821	60.13
Earthquake in Yi Liang	1	earthquake	2720652	8.01
Yang Mingtan bridge collapses	1	public facilities	1024489	3.87
360 and Baidu business battle	1	economic security	880421	24.98
Watch Gate of smiling secretary	1	political security	762557	1.48

Table 4. Emergencies timeliness value.

X	**Time Point**	α	$f_k(x)$	$Time(X)$
Diaoyu Island and anti-Japanese demonstrations	2012/9/18	125	0.076	9.51
Earthquake in Yi Liang, Yun Nan	2012/9/18	19	0.153	2.94
Yang Mingtan bridge collapses	2012/9/18	118	0.032	3.78
360 and Baidu business battle	2012/9/18	195	0.145	28.37
Watch Gate of smiling secretary	2012/9/18	121	0.016	1.94

The time complexity of One-Next clustering method that we proposed is:

$$. \ Time\,Complexity = \begin{cases} O(n) & only\ one\ class \\ \frac{1}{2}(k+1)n & average\ case \\ O(n^2) & n\ classes \end{cases} .$$

where k is the number of clusters, n is the number of clustering texts.

5.3. Experimental Results of Emergency Type Feature Function

To divide emergencies into the correct 22 categories, we select 10000 forum data samples to test similarity λ. Table **2** shows similarity parameters while setting λ different values.

Table **2** shows when λ's value is 0.3, which means the similarity of event feature vector $\vec{v} = (v_1, v_2, \cdots, v_n)$ and field feature vector $\vec{c} = (c_1, c_2, \cdots, c_n)$ is 0.3, the accuracy of emergency recognition is highest.

5.4. Emergency Quantitative Analysis

We set September 18 as the end time point, analysing 13.1 million shot texts in Tianya.cn. Calculating every emergency X's evaluation $Event(X)$, shown in Table **3**.

5.5. Experimental Results of Emergency Timeliness

Each emergency has its timeliness. Table **4** shows every parameter's value of the above emergency.

The above emergencies have their own changing maps during the time between August 19 and September 18 in 2012, as shown in Fig. (**8**).

Table 5. λ **Evaluation form.**

λ	P	R	F
0.7	93.94%	70.33%	80.43%
0.6	92.56%	79.21%	85.37%
0.5	89.22%	84.35%	86.73%
0.4	87.92%	87.22%	87.57%
0.3	84.71%	91.22%	87.84%
0.2	79.08%	92.36%	84.62%
0.1	68.33%	93.01%	78.78%

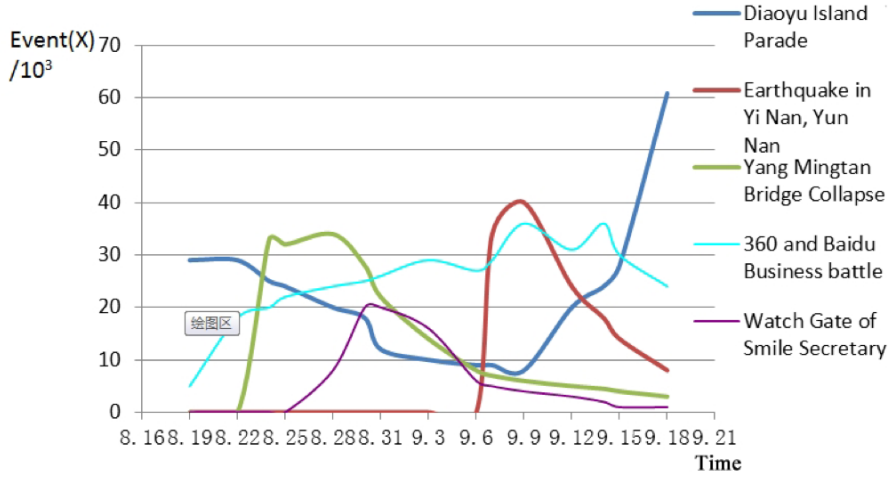

Fig. (8). Emergencies changing maps.

CONCLUSION

In this paper, One-Next text clustering method based on simplified hypothesis and a quantitative analysis model for network emergency is proposed. We mainly discuss three important factors: (1) emergency type. (2) emergency maximum effect range. (3) emergency timeliness. We also use mathematical methods to calculate each factor's value. Finally, through analysis data sets from Tianya.cn, we prove the model can accurately achieve good results.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by National Nature Science Foundation (No. 61331013), Beijing Higher Education Young Elite Teacher Project (No. YETP1291), National Language Committee Project (No. YB125-139, ZDI125-36), and Minzu University of China "First Class University and First Discipline Project" (No. 10301-0150200518).

REFERENCES

[1] Xinhua News Agency, "Major Issues Concerning Comprehensively Deepening Reforms," 2013

[2] People's Daily Online, "2012 Chinese Internet public opinion analysis report," 2012.

[3] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297, 1967.

[4] Kaufan L and Rousseeuw PJ, "Finding Groups in Data:An Introduction to Cluster Analysis," *New York:John Wiley & Sons*, 1990.

[5] Ester M, Kriegel HP, Sander J. and Xu X, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *in Simoudis E, Han JW and Fayyad UM, Eds, Proceedings of the 2th International Conference on Knowledge Discovery and Data Mining, Portland: AAAI Press*, pp. 226-231, 1966.

[6] Ankerst M, Breunig M, Kriegel HP and Sander J, "OPTICS: Ordering Points To Identify the Clustering Structure," *Proc. ACM SIGMOD Int. Conf. on Management of Data, Philadelphia, PA*, pp.49-60, 1999.

[7] A.Hinneburg and D.A.Keim. "DENCLUE: An Efficient Approach to Clustering in Large Multimedia Databases with Noise," *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD'98), AAAI Press*, pp. 58-65, 1998.

[8] W.Wang, J.Yang and R.Muntz. "STING: A Statistical Information Grid Approach to Spatial Data," *Proceedings of the 23rd VLDB Conference, Morgan Kaufmann*, pp. 186-195, 1997.

[9] Zhao Shiqi, Liu Ting and Li Sheng, "A Topic Document Clustering Method," *Journal of Chinese Information Processing*, vol. 21, no. 2, pp. 58-62, Mar. 2007.

[10] Ma Shuai and Wang Tengjiao, "A Fast Clustering Algorithm Based on Reference and Density," *Journal of Software*, vol.14, no.6, 2003.

[11] Gu Bo and Li Jihong, "Chinese Text Clustering Based on COSA Algorithm," *Journal of Chinese Information Processing*, vol.21 no.6, 2007.

[12] Zhai Donghai and Yu Jiang. "K-means text clustering algorithm based on initial cluster centers selection according to maximum distance," *Application Research of Computers*, vol.31, 2013.

[13] Zeng Runxi, "Constructions on Internet Emergencies Early Warning Index System," *Information studies: Theory & Application*, pp. 77-80, 2010.

[14] Zhu Li, "Emergency Concept Element and Type," *Social Sciences in Nanjing*, no. 11, pp. 81-88, Nov. 2007.

[15] Zhang Yiwen, Qi Jiayin, "Research on the Index System of Public Opinion on Internet for Abnormal Emergency," *Journal of Intelligence*, vol. 29, no. 11, pp. 71-75, Nov 2010.

[16] Qiu Jiangnan, Wang Yanzhang, "A Model for Predicting Emergency Event Based on Bayesian Networks," *Journal of Systems & Management*, vol. 20, no. 1, pp. 98-103, Feb. 2011.

[17] Liu Xingxing, He Tingting, "Design of Hot Web Event Detection System," *Journal of Chinese Information Processing*, vol.22, no.6, pp. 80-85, Nov. 2008.

[18] Zhang Lumin, Jia Nan, "Bursty Event Detection in Microblogging based on Sentiment Computing," *Netinfo Security*, Aug. 2008.

[19] Lin Dazhen, Li Shaozi, "Blog Emergent Event Detection Based on Temporal Distribution," *Computer Engineering & Science*, vol. 32, no. 10, pp. 145-149, 2010.

[20] Lan Yuexin, Deng Xinyuan, "Research on the Evolution Model of Network Public Opinion of Sudden Events," *Journal of Intelligence*, vol. 20, no.8, pp. 47-50. Aug. 2011.

Received: September 16, 2014 Revised: December 23, 2014 Accepted: December 31, 2014