ATLANTIS PRESS

Advances in Computer Science Research, (ACSR), volume 52

2016 International Conference on Computer Engineering and Information Systems (CEIS-16)

The Approach of Micro-blog Explosive Events Detection and Analysis in Real-time

Ying-Ying She, Qian Wang, Xiao-Yu Zhang,Wen-Yong Zheng, Yao Fu, Qing-Qiang Wu

Software School,

Xiamen University,

Xiamen, China

E-mail: yingyingshe@xmu.edu.cn

***Abstract*-Social network is an important platform for gathering social focus information. Explosive events from micro-blog can be used in the analysis and decision making in government monitoring, social phenomena research and other aspects. In this paper, we proposed an approach to detect and analysis explosive events from micro-blog in real-time. Our approach greatly improves the efficiency of explosive event tracking and analysis. It provides complete solution for social media event discovery. The experimental results show that the detection, tracking and analysis algorithms in the whole Micro-blog event detection process can efficiently improve the result of social media analysis in real-time.**

Keywords-component;social media analysis;micro-blog event detection; micro-blog explosion evaluation

I. INTRODUCTION

Social network [1] is a common platform for gathering information and provides a large number of data sources for big data era. Micro-blog is an emerging social network which is different from traditional blogs. If people can extract the valuable explosive events to do real-time tracking, it could produce inestimable social significance and commercial value. In this paper, we propose an approach of micro-blog explosive events detection and analysis. This approach is based on the research of unique attributes of micro-blog data and propagation characteristics of explosive events. In the experiment, we implement a Micro-blog explosive events analysis system. It takes Sina Micro-blog as a data source and uses the approach we proposed to do clustering and explosion analysis. Finally, we implement our method and visualize the analysis result in real-time.

II. RELATED WORK

According to the different data structures, information extraction usually comes in data mining method or natural language understanding and text mining [2]. In 2011, Jui-Weng, Cheng-Lun Yang and Bo-Nian Chen classified micro-blog text and extracted the important information viewpoint analysis and correlation detection module [3]. S. K. Endarnoto et al. used part of speech (POS) tag to mark and segment tweets text, and then extracted the information related to traffic congestion which is used to visualize on the map application of Android [4]. Jing Li et al. used mutual information calculation based on the assumptions of implicit topic relevance among micro-blog to extract topic vocabulary for hot recommended [5]. A micro-blog classification method based on machine learning has been proposed by M. Imran et al. [6].

Before the further analysis of explosion, we should classify events into right types via text clustering. There are five main approaches to solve clustering, division, hierarchical, density-based, grid-based and model-based [7]. The model-based clustering is used more frequently at present. Swit Phuvipadawat et al. proposed a method based on vector space model to calculate the relevance between two micro-blog vectors [8]. Xie Jing established similarity matrix of text and recalculate the similarity matrix by merging two texts with the maximum similarity [9]. TF-IDF (term frequency-inverse document frequency) is now commonly used method for text weighting. It is actually a product of TF (Term Frequency) and IDF (Inverse Document Frequency) [10].

T. Sakaki, M. Okazaki and Y.Matsuo presented a method for detecting the micro-blog topic about earthquake in Japan [11]. S. J. Zhang transformed text clustering problem into topic feature clustering problem, adopted a hot topic detection approach based on topic feature statistics and then established a comprehensive evaluation of public opinion security [12]. In 2012, Alan Ritter et al. created first open-domain events extraction and classification system of micro-blog [13]. In 2013, Z. X. Wang removed noise information in micro-blog data and obtained the keywords of events through recognizing the named entities, trigger words, time and other information in the sentences [14].

III. PROPOSED ALGORITHM

The research content of this paper includes a micro-blog explosive event detection and tracking solution that can extract the keywords and cluster data from micro-blog, real-time analysis of explosion and the implement of explosive events detection system which provides a visual interaction platform for these algorithms. The system adopts a natural language understanding algorithm with self-learning capability, analyses the raw data of micro-blog in real-time and then calculates the explosion of events. In order to enable users to understand the analysis better, this system realize the visualization of analysis results so that users can observe the development trend of explosive events through a series of real-time image and track the evolution process of events. Figure 1 shows the modular structure of the explosive event real-time analysis system. The detail of the data acquisition and processing module, micro-blog clustering module and explosion analysis module are explained in this paper.

The raw data of micro-blog can be obtained via API provided by Micro-blog open platform and stored as text files in txt format. After the redundancy information filtering,

feature words filtering, word segmentation and other pre-processing operation of raw data, these unstructured data are stored into a relational database to form micro-blog text dataset.

The raw data for clustering analysis in this paper are obtained from the API provided by Micro-blog, such as the Sina. Data pre-processing is the basis of data analysis, for text data, heuristic rules and natural language processing techniques is commonly used to extract representative features from the text. In this paper, we use three methods, redundant information filtering, text word segmentation and POS tagging, and feature selection.

Redundant information filtering means deleting the micro-blog content which are meaningless to the explosive event detection before clustering. This process consists of four parts, word count filtering, time filter, Stop Words filtering, and user-defined keyword filtering.

Word segmentation is the process to reassemble the continuous word sequence into terms sequence according to certain norms. Chinese Word Segmentation refers to segment a Chinese character sequence into separate words.

Feature selection refers to selecting some of the most effective feature from the feature set based on some of rules. There are two feature selection methods in this paper, document frequency feature extraction and information gain. DF (Document Frequency) is the number of the documents in the training corpus in which the term appears. If the DF value of a term is below a certain threshold, it is a low frequency words with less or no amount of information, so we can filter out it from the initial feature space to reduce the feature spatial dimension, simple and workable. IG (Information Gain) is an evaluation method based on entropy. Information entropy is a quantitative measurement of information to measure the uncertainty degree of random variable.

The detection of explosive events requires a lot of aggregated data. Relational database provides a platform to pre-process the messy data and add them into a data tables queue successively. The relational database contains an original aggregate data table and classification tables after clustering (one table corresponds to one category).

The main function of micro-blog clustering module is the twice clustering of original micro-blog text data based on the events topics, primarily through forwarding relationship and text similarity. Each table integrates all micro-blog that contains forwarding relationship or high similarity feature words of an event. The micro-blogs with the same event are stored in the same table, other micro-blog are uniformly stored in irrelevant data table.

In order to extract the popular micro-blogs, we defined a quantity coefficient *SingleAmo* for a single micro-blog. It is a comprehensive assessment of the micro-blog relevant property value, and its specific Equation is as follows:

$$SingleAmo = W_1 \times f + W_2 \times c + W_3 \times b \qquad (1)$$

Equation 1 is the forwarding count of a micro-blog, c is the comments amount, b presents the influence of blogger. W_1, W_2 and W_3 are the weighted coefficient of these three variables, respectively. After having a clear selection criterion of popular micro-blogs, it has to traverse all micro-blogs in the database and calculate the $SingleAmo_i$ of each micro-blog. Meanwhile, the module has to construct a *MaxHeap* to store the first k micro-blogs with largest $SingleAmo_i$. When the algorithm is complete, there should be no more than k micro-blog in *MaxHeap*. The *SingleAmo* of these micro-blogs are the largest in the existing micro-blog dataset, and these micro-blogs are the candidates of explosive event source micro-blogs. After establishing k database tables, the k micro-blog in *MaxHeap* are placed in each table respectively as the candidate sets of explosive event, the rest micro-blogs are placed into the pending micro-blog set.

Before the clustering we can cluster the micro-blog according to an important attribute--forwarding relationship based on text similarity. Under normal circumstances, forwarded micro-blog and original micro-blog are related to the same event, but the bloggers of forwarded micro-blog append their own comments about the event. Some bloggers with high influence may cause the second forwarding by tracking these forwarding relationships. We can get most of the micro-blog associated with the event from the dataset. This process is iterated until we add all the micro-blog with forwarding relationships into the corresponding candidate set.

The space vector dimensional model refers to the algebraic operations of mapping the calculation of text to the space vector dimension. Each explosive event candidate micro-blog set has a space vector. This vector space is composed of words with largest TF-TDF weight, the value of each dimension is corresponding TF-TDF weights. TF-IDF (Term Frequency-Inverse Document Frequency) is a common weighting technique used for text retrieval and semantic mining to assess the importance of the fragment after word segmentation to a particular event in the micro-blog clustering dataset. In this paper, TF (Term Frequency) refers to the occurrence frequency of a word in the current candidate set. This number is the normalization of phrase counts in order to prevent it biasing to the long documents. The same word may have a higher count in long micro-blog text than in short one, regardless of the importance of the word. For a given term t_i in the documents, its importance value can be calculated through Equation 2, in which, $n_{i,j}$ is the occurrence count of a term in the micro-blog set_j, and the denominator is the sum of the occurrence counts of all terms in the set_j .

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{m} n_{i,j}} \qquad (2)$$

IDF (Inverse Document Frequency) is a measure of the common importance of a term. We can get the IDF of a term by Equation 3.

$$idf_i = \log \frac{|D|}{|\{j:t_i \in set_j\}|} \quad (3)$$

$|D|$ presents the total number of documents in micro-blog text library. $\{j:t_i \epsilon set_j\}$ shows the number of the document that contains term t_i (the document counts of $n_{i,j}$). If the term is not in the micro-blog text library, the dividend is zero, so we use $1+\{j:t_i \epsilon set_j\}$ generally. Finally, we can calculate the TF-IDF weights of this term in the micro-blog text library by using Equation 4.

$$TF - IDF_{i,j} = tf_{i,j} \times idf_j \quad (4)$$

The TF-IDF weights correspond to every terms become the feature vector of this current micro-blog candidate set.

The purpose of the second clustering is mainly for the original micro-blogs in which have no forwarding relationship for an explosive event. In this paper, cosine similarity is used to analyze the similarity between a micro-blog in pending micro-blog set and the current micro-blog candidate set. The vector space adopted in this calculation is consists of feature words of current micro-blog candidate set, value of each dimension is the corresponding TF-IDF weight. We use a dot product to present the cosine similarity Θ between a micro-blog A and micro-blog candidate set B, the calculation method is shown in Equation 5.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (5)$$

In order to improve system efficiency and reduce memory consumption, noise reduction step is necessary. This process is mainly to remove the candidate set with a small amount of micro-blog and non-explosive events in pending micro-blog set.

In the explosion analysis module, we analyze the explosion of events based on micro-blog quantity coefficient, acceleration, diffusion rate and some other factors. Additionally, this module constructs a classification model based on fuzzy comprehensive evaluation to make an overall rating to the explosion of events. This model integrates the influence of all factors and also considers the assessment of single factor. After these processes, the events selected by this module are explosive events rather than simply popular events. The analysis results are stored in cartographic database for design and implementation of visualization.

IV. EXPERIMENT

During the real-time implement process, we consider the view layout and overall design of the system combined with the easy operation, interactivity and aesthetics principles. We select event extraction accuracy, intensity of clustering and deviation of explosion calculation as evaluation criteria to test the analysis results of system.

The event extraction accuracy AEvent can be calculated by Equation 10. Compare explosive events set extracted by system with the events set in explosive micro-blog subset, and obtain the number of the same events in both sets Ncross. NEventPre is the amount of all events in explosive micro-blog subset.

$$A_{Event} = \frac{N_{cross}}{N_{EventPre}} \times 100\% \quad (10)$$

Nextract reprensents the number of all micro-blogs related to a certain event that extracted by our clustering algorithm. It should not less than the amount of micro-blogs related to the same event in explosive micro-blog subset, which expressed by NOriginal. Only use NExtract to evaluate clustering is unreasonable, so the calculation method of intensity of clustering is shown in Equation 11. NNoise is the amount of noise micro-blogs which mistake clustering due to the constraint limits.

$$I_{Cluster} = \frac{N_{Extract} + N_{Noise}}{N_{Original}} \quad (11)$$

As shown in Equation 12, deviation of explosion calculation is associated with two variables. Epre is the explosion of each events in explosive micro-blog subset, calculated by system. After weighted average, the explosion from volunteers is represented by Epost. Intuitively, it is clear that the smaller DExplosion, the higher the accuracy of explosion calculation.

$$D_{Explosion} = \sqrt{(E_{pre} - E_{post})^2} \quad (12)$$

According to the results in Table 1, it can be seen that the results after clusterings can basically meet the requirements of system. Eighty percent of intensity of clustering is greater than 1, and the deviation of explosion calculation is about 0.22 within an acceptable range. This evaluation results show that the detection and analysis algorithms implemented by this system has a good effect and achieve the expected goal.

TABLE I. EVALUATION RESULTS OF EXPERIMENT.

Events	A_{Event}	$I_{Cluster}$	$D_{Explosion}$
Puer Ms6.6 earthquake in Yunnan	83.33%	1.16	0.21
Conflict of “defend the homeland” in Kunming		1.28	0.18
Performing arts companies will not hire the drug-related actors		1.07	0.21
Some people occupied central in Hong Kong		1.13	0.19
The micro-blog of Lu Han create a Guinness record		1.23	0.22
Sanitation worker stopped a BMW owners littering, got slapper in face		0.94	0.28
Chinese National Day anniversary		1.13	0.24
Ebola virus in Guangzhou		0.93	0.30
Bank card was drew back, a women dismantle ATM by hands		1.10	0.22
4th plenary session of 18th CPC Central Committee		1.02	0.22

V. CONCLUSION

In this paper, we propose an approach of analysis explosive events in social media micro-blog. It greatly improves the efficiency of clustering and reduces the computational burden. It has been implemented in the real-time Micro-blog explosive events analysis system. Considering the various factors that impact the explosion of events, the extraction accuracy rate of explosive events is more than 85%. The future work of this project will endeavour to expand this work to the detection and warning of explosive events on the social network platform with various types, in order to form a complete public opinion monitoring system of social network.

ACKNOWLEDGMENT

The project is supported by Fujian Science and Technology Plan Projects (no. 2015R0079).

REFERENCES

N. B. ELLISON. SOCIAL NETWORK SITES: DEFINITION, HISTORY, AND SCHOLARSHIP. JOURNAL OF COMPUTER-MEDIATED COMMUNICATION 2007; 13(1): 210-230

[1] B. L. Li, Y. Z. Chen and S. W. Yu. Research on Information Extraction: A Survey. Computer Engineering and Applications 2003; 39(10):1-5.

[2] J. Y. Weng, C. L. Yang, B. N. Chen, Y. K. Wang and S. D. Lin. IMASS: An Intelligent Microblog Analysis and Summarization System. In Proceedings of the ACL-HLT 2011 System Demonstrations, Portland, Oregon 2011; 133–138.

[3] S. K. Endarnoto, S. Pradipta, A. S. Nugroho and J. Purnama. Traffic condition information extraction & visualization from social media twitter for android mobile application. Electrical Engineering and Informatics (ICEEI), 2011 International Conference on. IEEE 2011; 1-4.

[4] J. Li, H. Zhang, H. X. Wu and J. Xiang. BTopicMiner: domain-specific topic mining system for Chinese microblog. Journal of Computer Applications 2012; 32(08): 2346-2349.

[5] M. Imran, S. M. Elbassuoni, C. Castillo, F. Diaz and P. Meier. Extracting information nuggets from disaster-related messages in social media. Proc. of ISCRAM, Baden-Baden, Germany 2013.

[6] J. G. Sun, J. Liu and L. Y. Zhao. Clustering Algorithms Research. Journal of Software 2008; 19 (1): 48-61.

[7] S. Phuvipadawat and T. Murata. Breaking News Detection and Tracking in Twitter. International Conference on Web Intelligence and Intelligent Agent Technology 2010; 205:120-123.

[8] J. Xie. Topic Detection and Tweet’s Trends Warning for Chinese Microblog. Master's thesis, Shanghai Jiao Tong University 2012.

[9] Q. Zhou and S. Y. Feng. Build a relation network representation for How-net. Journal of Chinese Information Processing 2000; 06.

[10] T. Sakaki, M. Okazaki and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World wide web 2010; 851-860.

[11] S. J. Zhang. Analysis model and implementation of public sentiment based on microblog social network. Doctoral dissertation, South China University of Technology 2011; 11.

[12] A. Ritter, Mausam, O.Etzioni and S. Clark. Open domain event extraction from twitter. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining 2012; 1104-1112.

[13] Z. X. Wang. Event Extraction and Sentiment Analysis on Microblog. Master's thesis, Shanghai Jiao Tong University 2013

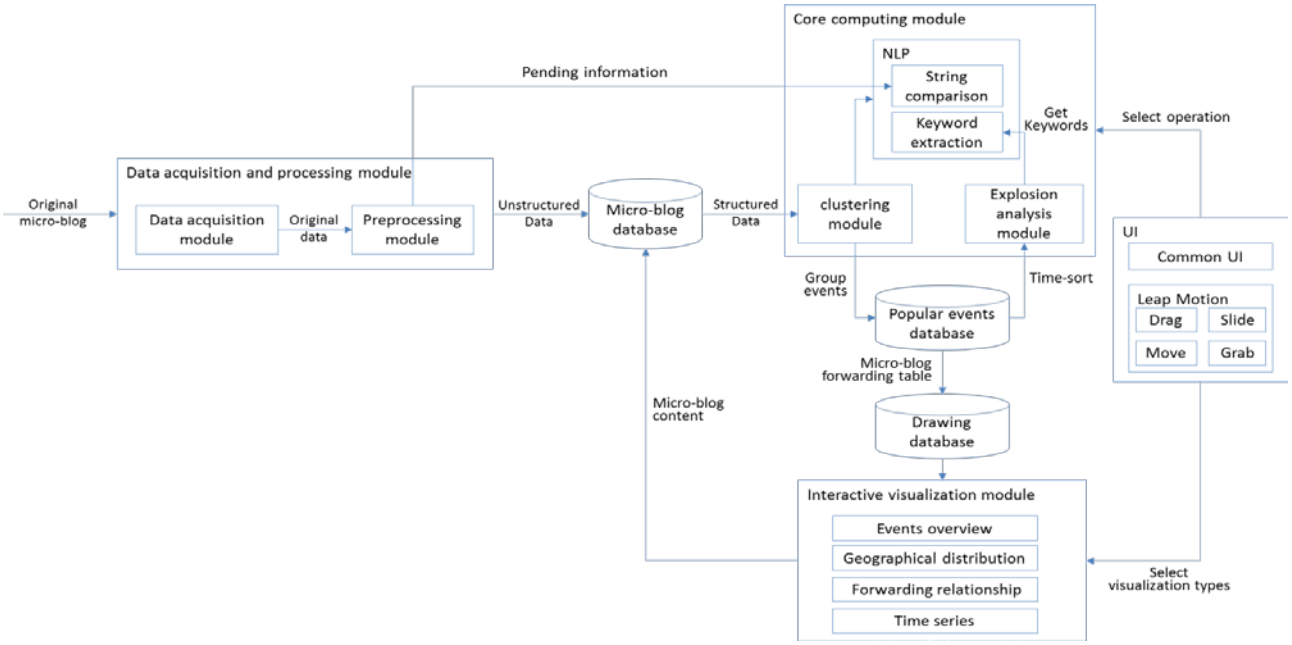

Figure 1. The structure diagram of micro-blog explosive event detection system.