Research of University Public Opinion Information Extraction Based on Micro-blog

Liang Hu, Jin Wen, Hao Wu, Jiangqing Liu and Hongmei Yu
Department of humanities and management, JiangXi Police College, NanChang, China

Abstract—**According to the management network of public opinion on the aging degree and real-time requirement, the algorithm is implemented to extract the classification of network popular university public opinion information, and determine the complex college network public opinion tendency, then find the center nodes that have influence on the sensitive points found in the mass of information. Thus the system could control the outbreak of hot events, establish the database of public opinion information resources in Chinese universities network, and provide college network public opinion assisted decision-making provide data and technical support for the university management.**

Keywords-public opinion; information extraction; micro-blog; network

I. INTRODUCTION

College network public opinion in Chinese refers to the university students in the network of some of the hot issues shown by the scale, with a certain scale and a clear tendency of the views or opinions of the situation. In the middle of twenty-first Century, the network technology has become an essential Internet communication skills, today's college students is a group of users[1]. According to the survey, the main problems of college students’ online attention, the international and domestic major events accounted for 25.4%, social hot and difficult issues accounted for 27.2%, the school's focus problems accounted for 29.1%, other issues accounted for 18.3%[2][3]. From these data, it is seen that the network of college students are more concerned about the problem of a wide range of ideas. With the rapid development of China's Internet, it provides a platform for the national consciousness to express more and more. Among the college students of this group have more convenient access conditions and higher level of knowledge and technical ability, and they put on enriching the road network as the most important places of national consciousness and patriotic expression. College students' ability to discern between network rumors and false information is weak, and the network public opinion tends to be emotional. According to the survey, there are 6.3% college students often spread on the Internet, the spread of false news, there are 34.4% of college students have, but not often on the Internet to spread, the spread of false news, there are 59.3% college students have spread and spread false news on the internet[4][5].

In recent years, the statistics show that China's colleges and universities for network public opinion information processing is not timely, in terms of response and processing is lagging behind, there is a certain percentage of college students have on public opinion information processing of the distrust mentality[6][7]. College students today is the use of a group of network more, the network is flooded with all kinds of information, ideas and behavior of these information for college students has a great influence, because of this university will become a public opinion event occurred is the most concentrated place.

The Internet public opinion is the influence of college students' collective action effectiveness indicator, and determines structure characteristics of network public opinion hot events perception will directly determine the development trend of college students' collective action based on it, in order to prevent the adverse effects caused by the social public opinion information on high school, effective monitoring of network public opinion, to provide decision-making basis for the management of personnel, the research aims to construct to guarantee the normal order of teaching and scientific research of college network public opinion decision support management platform, to promote college students' physical and mental development[8].

II. MICRO-BLOG TEXT INFORMATION EXTRACTION

Micro-blog has the following characteristics: First, the text length is short that is often just a word or a phrase and includes data sparse problem in text processing. Second, Micro-blog text more use of spoken language form, often contains abbreviations, omitted, refers to the generation of new words, expressions, such as symbols, and even the spelling errors. This has brought great difficulties to the understanding of the text; third. Micro-blog appears in the post in the form of text, often comments and reprint information, provides a rich context for the micro-blog fourth, semi-structured text comprehension. Micro-blog in addition to content, but also contains the post author, publishing time, comments, forwarding data, such as the number of metadata. Micro-blog public opinion analysis through the analysis of micro-blog and users of the text data, determine the hot topics, sensitive core users, so as to early warning of sensitive events of public opinion.

A. Micro-blog Topic Detection and Tracking

Micro-blog topic detection includes event review detection, online topic detection, new event monitoring and hierarchical topic detection, etc.. Incident detection of the micro-blog text over a period of time without being detected, from identifying related topics, using hierarchical agglomerative clustering algorithm and average packet clustering algorithm or a combination of both strategies, usually micro-blog text number, average packet clustering algorithm can effectively reduce the clustering time cost. Online topic detection is to build a topic detection model in the case of unknown topic, and deal with the arrival of the micro-blog text, from which to identify the latest topic. New event detection is generally based on the single channel clustering algorithm, the order of each micro-blog text and the existing class clusters for similarity comparison, determine whether it is classified into the existing class clusters. Micro-blog aims to topic detection method of text clustering of topic clusters in different micro-blog text, or the establishment of new topic clusters, including monitoring a topic of micro-blog, and micro-blog will involve a topic organized in a manner presented to the user, it emphasizes on the ability of the new information.

B. *Micro-blog Text Sentiment Analysis*

Micro-blog text sentiment analysis is to carry on the analysis processing, the induction and the inference process to the subjective text with emotion color, also called the opinion mining. In order to determine the emotional polarity of words, usually choose the interval [-1, 1] on a real number as emotional weight. If the emotional weight is greater than 0, said on the contrary, if the commendatory terms, the emotional weight is less than 0, as a derogatory term. With the absolute value of the degree of emotional weight said appraise words. Word polarity discrimination mainly by HowNet or WordNet Chinese dictionary provides semantic similarity or hierarchy judgment method based on dictionary word sentiment polarity, and the use of conjunctions between words and statistical characteristics to determine the corpus based approaches to Lexical Emotional polarity. Statement of sentiment analysis for the statement of distinction between subjective and objective, to appraise subjective sentences to judge, and the statement of fine-grained sentiment including evaluation object, opinion holder extraction etc..

C. *Micro-blog Public Opinion Attention Analysis*

Micro-blog public opinion attention can be reflected by micro-blog heat. Through the analysis of micro-blog public opinion attention, can be found in a timely manner, the focus of attention of public opinion, major events and the extent of the public's attention. Micro-blog concerned about the value of the number can be forwarded through micro-blog, the number of comments, the number of listeners, micro-blog release time to determine. For example, micro-blog sina is through the aggregation of the same keywords micro-blog, the statistics of the number of forwarding, comments, etc., in accordance with the statistical value of the introduction of hot topics. The formula is as follows:

$$P = \sum(\log(1+fS_i) + \sqrt{Zf_i} + pl_i) + \sum\frac{1+T_i-t_{i1}}{(1+t_{i2}-t_{i1})\sqrt{pl_i}}\log(1+fS_i) \qquad (1)$$

Among them, fS_i is the number of listeners in article a released by micro-blog users, Zf_i is the number of the micro-blog, pl_i is the number of times to be forwarded, T_i is the first comment release time, t_{i1} is for the first time micro-blog comments, t_{i2} is the last micro-blog a review of the time.

D. *Micro-blog Text Information Extraction Algorithm*

Micro-blog will feature a concept ontology algorithm and the network public opinion in the field to match the corresponding matching, if the match is successful on the use of the concept instead of features, and the concept of concept and features of items in the collection, the original weights and the feature weights by the concept of the feature set in the same item will be merged, with higher weight concept features retained, as shown in Figure 1.

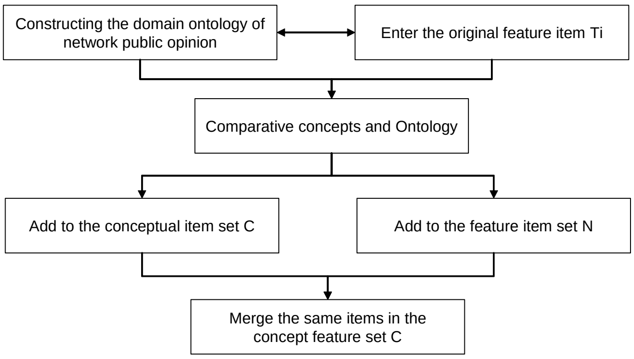
Constructing the domain ontology of network public opinion
Enter the original feature item Ti
Comparative concepts and Ontology
Add to the conceptual item set C
Add to the feature item set N
Merge the same items in the concept feature set C

FIGURE I. SEMANTIC FEATURE VECTOR GENERATION

According to the flow chart can be seen, the concept matching algorithm is the core of the process, the Web text of the original feature set for $T=\{t_1, t_2, t_3,... , t_n\}$, the semantic feature item set is represented as the concept feature set C, and the algorithm is described as follows:

Step 1 Loading network public opinion domain ontology;

Step 2 Input feature T_i;

Step 3 Traversal of all classes in the field of public opinion in the network of public opinion, if there is a corresponding class and the match, then the feature item T_i to join the concept of feature item set C;

Step 4 If there is no class to match, then the network public opinion in the field of public opinion all the attributes, if there is a corresponding attribute matching, then the attribute belongs to the concept into the concept of feature item set C;

Step 5 If there is no attribute matching, then all instances of ontology traversal in the field of network public opinion, if there is a corresponding instance match will be instances of their subordinate concepts into the concept of minimum feature set C;

Step 6 If there is no correlation between the concept feature and Ti matching, then it is put into the non matching feature item set N;

Step 7 Merge the same items in the concept feature set C, and keep the concept feature item of the weight;

Step 8 Return the concept feature set C.

By the above algorithm, the generation of Web text semantic features set $C=\{c_2, c_1, c_3, ..., c_m\}$, the weight of c_i could be determined by the weight of its original feature T_i, then the semantic feature vector of network public opinion is obtained.

III. Performance and Testing

In this experiment, text clustering is used to verify the effect of semantic feature extraction, in which the text similarity based on the semantic feature is used to measure the text similarity. In the usual clustering process cluster number *k* is generally specified in advance. In this paper, the evaluation indicators used in this paper are the standard of mutual information between clusters NMI that is independent of the number of clusters *K*.

The test data set is assumed to contain *m* text, and there are *n* classes $C=\{c_1, c_2, \ldots, c_n\}$, if the clustering results will *m* the text is divided into *k* cluster, in which the first class c_i and *j* cluster c_i are included in the m_i and m_j text, then the formula is as follows:

$$NMI = \frac{\sum\sum m_{ij}\log\frac{m\times m_{ij}}{m_i\times m_j}}{\sqrt{\sum m_i\log\frac{m_i}{m}\sum m_j\log\frac{m_j}{m}}} \qquad (2)$$

In this paper, we have labeled a good data set for each category in 1000 randomly selected as a test set. In order to verify the actual effect of the network public opinion information clustering algorithm, this paper makes a comparison between the network public opinion clustering algorithm based on semantic feature extraction and other clustering algorithms on the same data set. This paper selects the reference algorithm such as K-Means algorithm, MBM(Multivariate Bernoulli Model) clustering algorithm based on probability model, MM(Multinomial Model) clustering algorithm based on polynomial model. As shown in Table 1, all of the NMI values are expressed as mean ± standard deviation in the form of *K*, the table shows the number of clusters average.

TABLE I. COMPARISON OF NMI VALUE OF CLUSTERING RESULTS

k	11	15	17	21
K-Means	0.55±0.02	0.58±0.01	0.58±0.01	0.57±0.01
MM	0.52±0.02	0.54±0.04	0.54±0.04	0.56±0.02
MBM	0.36±0.01	0.46±0.01	0.50±0.01	0.51±0.01
Micro-blog	0.57±0.02	0.65±0.03	0.68±0.01	0.69±0.01

The experimental results show that the clustering method proposed in this paper by using the network public opinion domain ontology keywords feature set into Web vector containing text semantic concept features of ontology in the field of network public opinion collection based on the full vector expression of network public opinion Web text semantic content can be set, and can improve the accuracy of clustering, as shown in Figure 2.

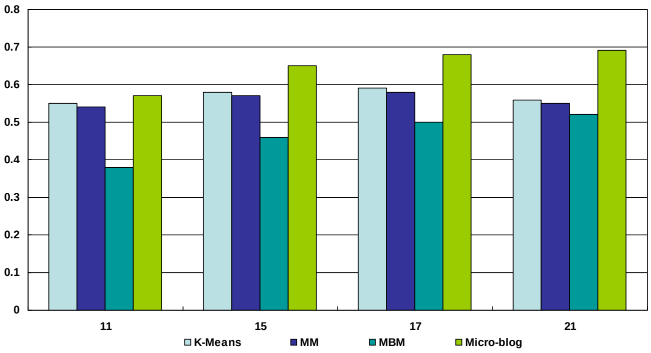

FIGURE II. COMPARISON OF CLUSTERING ALGORITHMS

In terms of clustering efficiency, the average runtime of each algorithm is shown in Figure 3, in which the K-Means algorithm has the longest running time, and the efficiency of this algorithm is in the same order of magnitude as MBM and MM. The conversion algorithm process keywords space mapped to semantic concept space in the Web text semantic feature extraction, using ontology in the field of network public opinion, attributes, instances to represent the semantic features of Web text, so that the dimension of feature space is greatly reduced, in the process of making text similarity computing Web complexity is simplified, operation time and shorten the algorithm.

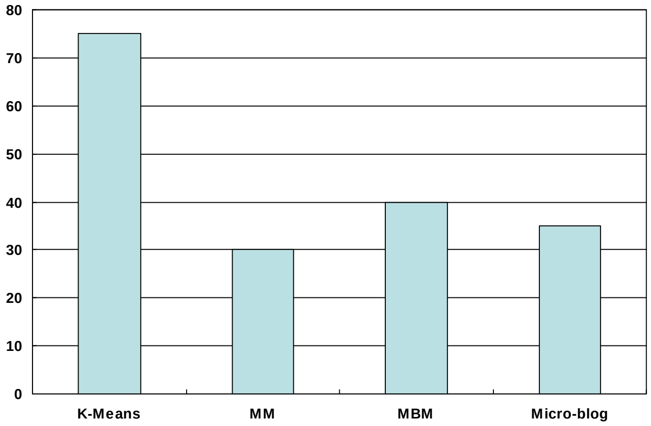

FIGURE III. ALGORITHM RUNNING TIME COMPARISON

IV. Discussion and Future Work

This study attempts to processing, classification and annotation of micro-blog data extraction, hot public opinion data contain additional elements of college students attribute the use of templates, the construction of college network public opinion information extraction model, is a valuable supplement to the field of public opinion. Because micro-blog text brief expression of freedom and flexibility, are often omitted and refers to the situation, the existence of grammatical phenomenon is not standardized, it is difficult to study micro-blog text, that there is a big gap between the theoretical research and practical application. In the further research of the future, the establishment of large scale corpus and the improvement and innovation of the algorithm combined with micro-blog's text feature is very necessary.

ACKNOWLEDGMENT

This author’s work is supported by JiangXi Research on teaching reform of higher education (JXJG-15-19-3), JiangXi Science and technology research project of Education Department(GJJ151193), JiangXi Social Science Planning Projects during the 12th Five-Year Plan(14TQ05), JiangXi University Party Building Project(JXGXDJKT.QN-201665) and JiangXi Police College Scientific Research Project (2014QN001).

REFERENCES

[1] Sun Wei. Micro-blog user interest mining and modeling research[D]. Dalian: Dalian University of Technology, 2012. (in Chinese)

[2] Field. Event trend analysis and prediction research based on micro-blog platform[D]. Wuhan: Wuhan University, 2012. (in Chinese)

[3] Duan Tingting. Research on the generation mechanism of the influence of micro-blog[D]. Ji'nan: Shandong University, 2012. (in Chinese)

[4] Lv Wenna. Study on the population structure of micro-blog group[D]. Beijing: Beijing Jiaotong University, 2011. (in Chinese)

[5] Luo Yaping. Study on the sentiment orientation of Chinese text sentiment analysis based on network public opinion[D]. Dalian: Dongbei University of Finance and Economics, 2010. (in Chinese)

[6] Zhang Jing. Research on the model and platform of network hot spot discovery based on micro-blog[D]. Wuhan: Huazhong University of Science and Technology, 2010. (in Chinese)

[7] Zhang Shaojie. Analysis model and Realization of public opinion based on the social network of micro-blog[D]. Guangzhou: South China University of Technology, 2011. (in Chinese)

[8] Zhang Jianfeng, Xia Yunqing, Yao Jianmin. A review of research on text processing in micro-blog[J]. Chinese Journal of information and information technology, 2012,26 (4): 21-27. (in Chinese)