ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 159

International Conference on Mathematics, Modelling, Simulation and Algorithms (MMSA 2018)

Research on Hot Words Mining Algorithm of University Network Public Sentiment

Liang Hu[1,2,*] and Hongmei Yu[1]

[1]Department of humanities and management, Jiangxi Police College, Nanchang City, Jiangxi Province, P.R. China
[2]Collaborative Innovation Center for Economics crime investigation and prevention technology, Jiangxi Province, P.R. China
[*]Corresponding author

***Abstract*—Hot words are the topic of concern of the netizens, which can help the management department to monitor the public opinion of the network. Due to the large degree of freedom, irregular syntax and immediacy of data, it is difficult for data engine to grasp text hotspots accurately through traditional text analysis. This paper deals with the text features of a time span, and determines whether it is a hot word by adding the time information of the span. Through experiments, the hot words excavated correspond to the corresponding hot events. This shows that the method proposed in this paper has a good effect and can be further studied.**

Keywords—network public opinion; hot words; data mining; university

I. INTRODUCTION

With the rapid development of computer technology and network technology, great changes have taken place in the form and way of receiving information. The traditional way of receiving information is television, newspapers, books and so on. At this stage, web sites, WeChat, micro-blog, QQ and other new media environment, such as mobile Internet, have become the main way for people to obtain information. Therefore, when major events occur, the process and orientation of public opinion emerge in the new situation and new changes. The traditional ideological and political education work is remedied after the outbreak of public opinion, which has fallen behind in time. College students are in close contact and users of new media technology, plays an important role in the spread of the Internet public opinion in Colleges and universities, therefore, educators should closely follow the development trend of network, the establishment of early warning mechanism of network public opinion, control of College Students' ideological trend, to provide a stable learning environment for college students, so as to promote the smooth development of the ideological and political work in colleges.

Micro-blog is a platform based on user relationship to achieve information sharing, dissemination and acquisition. Its character is limited to 140 words, which is real-time, convenient, concise, fast and interactive. It is widely favored by users. Hot topics in chat texts generally refer to topics that cause much attention in a span of time, and can correspond to hot events that have occurred nowadays. They have short time span and high content diffusion. In order to help users quickly understand and participate in the topic of interest, the hot topics in the society are excavated from mass chat text. However, due to the huge volume of texts, it is difficult to analyze massive chat text data timely and efficiently based on artificial methods, and find out the hot topics related information in the current text. Therefore, it is necessary to use a variety of different content mining algorithms and information analysis techniques to accurately locate the hot topics in the current chat text.

II. CHARACTERISTICS OF PUBLIC OPINION IN UNIVERSITY

The group of college students occupies a large proportion among the netizens, and the rich pictures, audio and images undoubtedly lead to a large number of data. The development of media, such as micro-blog, WeChat public address and so on, has provided the voice room for college students. The Internet public opinion reflects their world outlook, outlook on life and values. The diversification of the expression form and media, involving the diversification of the theme, makes the work of network public opinion more complex in Colleges and universities. Second, the emotion is stronger. Because of the large proportion of population, group behavior is very easy to infect each other. Once an ideology touches individual psychology, individuals will lose their ability in truthfulness judgment and cause irrational communication. But it is not difficult to find that the cause of irrational communication is the common psychology of college students. Using big data technology can excavate the relevance of the main body, transform the amount of reading and forwarding data, and strengthen the analysis, so as to achieve the control of public opinion in Colleges and universities. Third, "opinion leaders" are more. The network gives college students' equal right to speak, the individual can through the network media to express their views on social phenomena, the view, youth psychology and empathy agglomeration makes part of the individual from the backstage, the speech aroused in most of the students, gradually regarded as "opinion leaders", cause the cluster behavior of even the phenomenon of group polarization. In a word, the development of the era of big data has brought great challenges to the work of public opinion control in Colleges and universities.

III. TOPIC MODEL OF MICRO-BLOG PUBLIC OPINION

This paper selects micro-blog as the research platform, and constructs the network public opinion theme model through 4 dimensions of information time, information content, user relationship and user behavior.

A. Information Time Dimension

Public opinion event life cycle has been selected as a time granularity, which is to analyze all stages of public opinion events in order to achieve dynamic monitoring of public opinion propagation process.

B. User Relationship Dimension

Users are the main body of Internet public opinion, and Internet public opinion events are formed by users' cognition, attitudes and opinions on the Internet.

C. Information Content Dimension

The content of this article refers to user's emotional content, which is published by micro-blog or micro-blog reviews. The micro-blog content directly released by users is called "original micro-blog", and the comment content is called "comment micro-blog".

D. User Behavior Dimension

Views usually refer to users' opinions or emotional tendencies towards an event or thing, rather than the basic elements of public opinion events, but based on content summarization and extraction, that is, views are highly generalized and summarized from users' published contents.

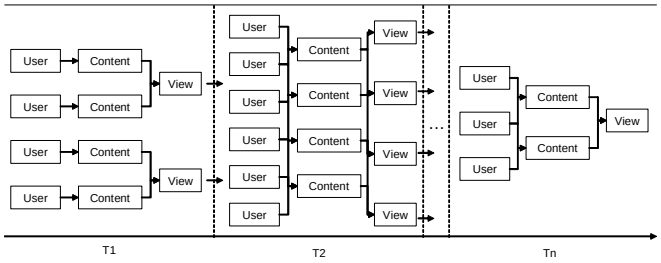

FIGURE I. TOPIC MODEL OF MICRO-BLOG PUBLIC OPINION

IV. HOT WORDS MINING

Hot words have a high frequency in a time window. Some words are pseudo hot words, although the frequency of high frequency, possibly even more than real hot words, but these words have very high frequency in every text, distribution and daily on the longer average, so it can not be used as hot words. The steps to get hot words in this paper are as follows:

1) The *N* words are counted according to the daily word frequency to form a vector $V_j = (a_{j1}, a_{j2}, a_{j3}, \ldots, a_{jm})$, $V_N = (a_{n1}, a_{n2}, a_{n3}, \ldots, a_{nm})$ form representation, where a_{jk} indicates the number of times *j* words appear in the text of day *k*, and *M* indicates that the time span of the text used for analysis has a *m* day.

2) Compute the average number of *N* words in the chat text, respectively:

$$Avg(w) = \frac{(a_{j1} + a_{j2} + \cdots + a_{jm})}{m} \tag{1}$$

and record the maximum frequency.

3) All words are calculated using the next formula, and the first *L* words are used as the candidate hotspots:

$$S = Avg(w) \times Std(w) \tag{2}$$

Word frequency is reduced to reduce the influence of high frequency ordinary words, of which *Std(w)* is a vector standard deviation.

4) The *L* words are further screened using the next expression, in which n depends on the correct rate and recall of the hot words that need to be extracted. The hot words after screening are the final hot words of text mining.

V. EXPERIMENTAL RESULTS

This paper extracts 50 thousand chat text data, the first word of the daily text for the day statistics words the number of words, each word to construct a multi-dimensional vector, and calculate the mean and standard deviation of each word, the final score is calculated for each word, the extraction of high 2000 words as before the candidate words, and then screened, the final results are shown as follows:

TABLE I. SORTING OF HOT WORDS

Keywords	Avg	Std	Score
College student	81	54	134
Tuition	58	36	125
CET	30	49	122
Scholarship	48	31	121
Job-hunting	49	24	119
Graduation	30	32	117
Civil service examination	44	20	115
Entrepreneurship	31	14	109
Entertainment	19	19	107
Part-time job	13	26	107

The experimental results show that CET is one of the current hot spots, and the graduation is also a hot spot. It can be seen that the algorithm proposed in this paper has a good effect in the mining of chat text.

VI. DISCUSSION AND FUTURE WORK

This paper presents a data for a large chat text hot words mining algorithm, the experimental results show that the algorithm has good mining effect, can more accurate positioning of the text in the short term hot words appear, and the high frequency of daily use of non hot words have a good filtering effect, can hot spot mining from massive words in. However, due to the fact that the theme of the text is very

dispersed in the big data environment, many hot words are not necessarily accompanied by large word frequency changes.

ACKNOWLEDGMENT

This author's work is supported by Jiangxi Science and Technology Research Project of Education Department (GJJ151193), Jiangxi University Party Building Project(16DJQN065), Jiangxi Police College Scientific Research Project(2016JGZB008) and Jangxi Science Education Planning Project(17YB244).

This author's work also is supported by the Opening Project of Collaborative Innovation Center for Economics crime investigation and prevention technology, Jiangxi Province.

REFERENCES

[1] Wang Xuemei. Study Of Improved K-Means Clustering Algorithm[J]. Computer And Digital Engineering, 2013, 41 (11): 1717-1719. (in Chinese)

[2] Chen Fuji. The Network Public Opinion Event System Dynamics[J]. Journal Of Information Communication Research, 2015, 34 (9): 118-122. (in Chinese)

[3] Liu Hong. Study Of Hot Spot Of Network Public Opinion[J]. Science And Technology Bulletin, 2011, 27 (3): 421-423. (in Chinese)

[4] Xie Haiguang. Internet Content And Public Opinion Analysis [J]. China Youth Political College 2006, 25, 3:95~100. (in Chinese)

[5] Li Gang. Research And Empirical Analysis Of The Communication Process Of Network Public Opinion Under The Environment Of Web2.0[J]. Information Science, 2011, 29 (12): 1810-1814. (in Chinese)