Research on the Monitoring and Guidance of Internet Public Opinion by Big Data Mining Technology from the Perspective of the Internet

Ji Sun

School of Art and Design, Liaoning Communication University, Shenyang 110013, China

9841184425@qq.com

Abstract. Based on the Internet technology, this article proposes a leader guidance model to intervene and guide the public opinion super network, and conducts model simulation to explore the influence mechanism of the model and the influence of model parameter changes on the guidance effect. The simulation results show that the use of the leader guidance model for public opinion guidance and intervention can not only quantify the guidance strategy, but also obtain the role and feedback law of the guidance strategy through the changes and linkage mechanisms of various elements in the hypernetwork, providing a new way for guiding network public opinion Quantitative methods.

Keywords: Internet; Big data mining; Network public opinion monitoring; Public opinion guidance.

1. Introduction

With the popularization of the Internet, the Internet has increasingly become an important channel for people to publish, exchange, and obtain information. It has also become an important platform for netizens to comment on social events, express their attitudes and vent their emotions, and even become an important force to promote the development of events [1]. The monitoring and management of online public opinion has increasingly become an important aspect that affects national interests and is related to national security. This makes the research on the guidance and intervention of online public opinion a hot spot.

More and more netizens tend to express their opinions and spread ideas through this channel, and the Internet has become the birthplace and center of information. If the guidance is not good, negative online public opinion will pose a greater threat to social public safety [2]. How to strengthen timely monitoring and effective guidance of online public opinion, and actively resolve the crisis of online public opinion, has important practical significance for maintaining social stability and promoting national development, and it is also the connotation of creating a harmonious society.

2. Overall evolution analysis

2.1 Overall quantitative evolution

Analyze the evolution of the total amount of event e over time, and obtain the changes in the amount of discussion of event e by netizens over time, as well as the emergence, growth, decline, and disappearance of the event. The overall evolution sequence EVL_e of event e can be expressed as the following formula:

$$EVL_e = \{t_1^{num(e)}, t_2^{num(e)}, ..., t_n^{num(e)}\} \quad (1)$$

Among them, t_n represents the current time slice, and $num(e)$ represents the number of events. The evolutionary analysis of the total amount of sub-topics ST_i contained in each event is carried out to obtain the emergence, growth, decline and disappearance of a single sub-topic. The evolution sequence EVL_{ST_i} can be expressed as the following formula:

$$EVL_{ST_i} = \{t_1^{num(ST_i)}, t_2^{num(ST_i)}, ..., t_n^{num(ST_i)}\} \quad (2)$$

Among them, t_n represents the current time slice, and $num(ST_i)$ represents the number of subtopics.

2.2 Evolution of subtopic sequence

Analyze the number of sub-topics $sNum(e)$ of event e, and obtain the growth of $sNum(e)$ overtime and the sequence of sub-topics of event e in each time slice t_i. The subtopic growth sequence e_{t_i} of event e can be expressed as the following formula:

$$e_{t_i} = \{ST_1^{t_i}, ST_2^{t_i}, ..., ST_n^{t_i}\} \quad (3)$$

Among them, t_i represents the current time slice, and ST_n is the n subtopic.

3. Subtopic association model

A directed acyclic graph (DAG) is constructed by calculating the correlation value between the sub-topics, which we call the sub-topic correlation graph [3]. Use it to describe the relationship between the various sub-topics of the event, and intuitively see the basic process of event evolution and the evolution relationship between the sub-topics from the diagram.

3.1 Subtopic association graph

Articles organize data according to three granularities: events, subtopics, and posts. The sub-topic association diagram of an event is shown in Figure 1 below:

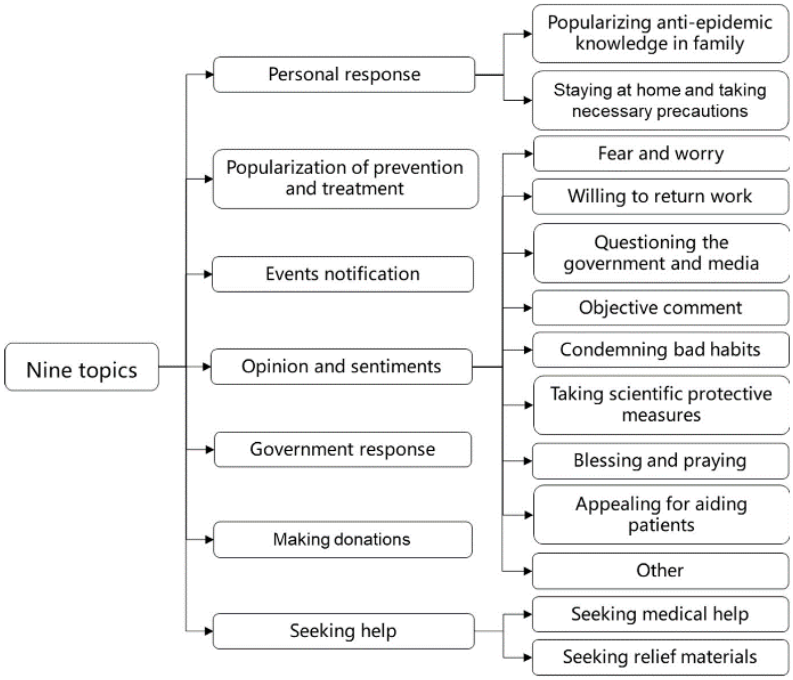


Figure 1. Association diagram of subtopics

The two sides A and B in the above figure indicate that: Subtopic 1 has triggered Subtopic 2 and Subtopic 3.

3.2 Timing similarity

This article assumes that each subtopic ST occurs at a specific time, denoted as: $start\ (ST)$. And it ends at a specific time, denoted as: $end\ (ST)$. The duration of the subtopic is the difference

between the end time and the occurrence time, denoted as: $len\ (ST) = end(ST) - start(ST)$. Because the sub-topic association graph shows the triggering relationship between two sub-topics, only when the occurrence time of one sub-topic ST_i is before the other sub-topic ST_j, can ST_i trigger ST_j, and it is possible between the two Create an edge [4]. If the occurrence time of the two subtopics is the same, the two subtopics do not have a triggering relationship. The temporal similarity $TSS(ST_i, ST_j)$ between subtopic ST_i and subtopic ST_j is defined as:

$$TSS(STi, STj) = e^{-\frac{|end(STi) - start(STj)|}{len(STi)}} \tag{4}$$

3.3 Content similarity

Content similarity is used to separate several subtopics from an event. Assuming that the post sets contained in two subtopics ST_i and ST_j are D_{ST_i} and D_{ST_j} respectively, then the content similarity refers to the similarity calculation of the posts in D_{ST_i} and D_{ST_j}, denoted as:

$$SimCS(pi, pj) = \frac{\sum_{k=1}^{n}(Vi_k, Vj_k)}{\sqrt{\sum_{k=1}^{n} Vi_k^2 \times \sum_{k=1}^{n} Vj_k^2}} \tag{5}$$

pi, pj are the posts in the post collection D_{ST_i} and D_{ST_j}, respectively, and Vi, Vj, are the vector representations of the posts pi and pj, respectively [5]. Content similarity comprehensively considers the similarity between all posts under one subtopic and all posts under another subtopic, and requires sufficient similarity between posts on two interrelated subtopics.

3.4 Feature similarity

Feature similarity is also used to separate subtopics from an event. The feature of a sub-topic refers to extracting those feature items that represent the sub-topic to the greatest extent from the post collection to which it belongs to represent the sub-topic, and the selection of feature items is based on TF-IDF. Therefore, the feature similarity calculation is performed only once for the two subtopics [6]. Suppose the feature vectors of two subtopics ST_i and ST_j are Fi and Fj respectively, then their feature similarity is recorded as:

$$SimFS(STi, STj) = \frac{\sum_{k=1}^{n}(Fi_k, Fj_k)}{\sqrt{\sum_{k=1}^{n} Fi_k^2 \times \sum_{k=1}^{n} Fj_k^2}} \tag{6}$$

Feature similarity is different from content similarity. It does not require a high degree of similarity between the posts of two subtopics, but only judges the similarity based on the characteristics of the two subtopics, and requires two interrelated subtopics to be guaranteed. The feature vocabulary is sufficiently similar.

3.5 Constructing an association graph

After obtaining the above-mentioned similarity calculation indicators, we define an association value R to measure the degree of association between two subtopics, and use the value of R as the basis for establishing an edge between two nodes [7]. According to the indicators defined in the previous sections, the correlation value of the two sub-topics is affected by the following three factors: the time-series similarity of the two sub-topics; the distance between the two sub-topics, which is the trigger period; and the post similarity of the two sub-topics. Therefore, the correlation value $R(ST_i, ST_j)$ between the two subtopics ST_i and ST_j is recorded as:

$$\begin{cases} 0,(start(ST_i)\geq start(ST_i)) \\ a\times TSS(STi,STj)+a'\times(b\times SimCS(STi,STj)+b'\times SimFS(STi,STj)) \end{cases} \quad (7)$$

Among them, a , a' , b and b' are adjustment parameters used to balance sequence similarity, content similarity, and feature similarity, and satisfy the following relationships:

$$a+a'=1,b+b'=1 \quad (8)$$

When the correlation value $R(ST_i,ST_j)$ exceeds the threshold r , an edge will be established between the subtopics ST_i and ST_j , indicating that ST_i causes ST_j . The computer of the associated value and the display process of the associated graph are shown in Figure 2 below:

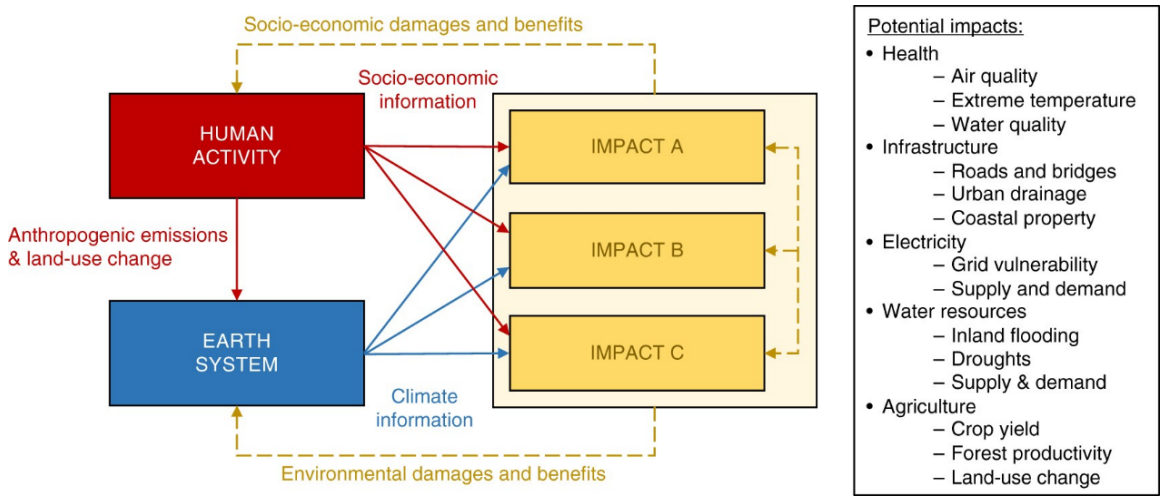


Figure 2. Subtopic association relationship processing flow

In the above figure, the first step is to distribute the posts to each event and obtain the event ID; the second step is to distribute the posts under each event to each sub-topic and obtain the sub-topic ID; the third step is to calculate the sub-topic correlation value and construct the sub-topic Topic association graph.

4. Analysis of experimental results

We carried out a correlation analysis on each subtopic under each event. Let's take the event 1: "360VS Tencent" after the first clustering as an example to analyze the experimental results [8]. This chapter uses the method of Chapter 3 to cluster the collection of posts contained in the "360VS Tencent" event to obtain 6 effective clusters, each of which represents a subtopic and contains more than 10 posts. The 6 subtopics are shown in Table 1 below:

Table 1. The second clustering results (take the No. 1 cluster category "360VS Tencent" as an example)

Cluster number	Time of occurrence	Number of members	Cluster class ID
1	2020-10-14	47	Tencent 360 sues
2	2020-10-30	31	360 buckle bodyguard
3	2020-11-03	64	360 Tencent deducts bodyguards and recalls one of the two
4	2020-11-08	19	Government compromise
5	2020-11-10	12	360 compatible products Tencent
6	2020-11-21	23	360 Ministry of Industry and Information Technology apologize and criticize

From the above table, we can see the 6 subtopics of the "360VS Tencent" event. Calculate the associated value of these subtopics. When calculating the correlation value, the threshold r is set to

0.5, and the parameters a and b are 0.3 and 0.3, respectively. Get the correlation diagram shown in Figure 3 below:

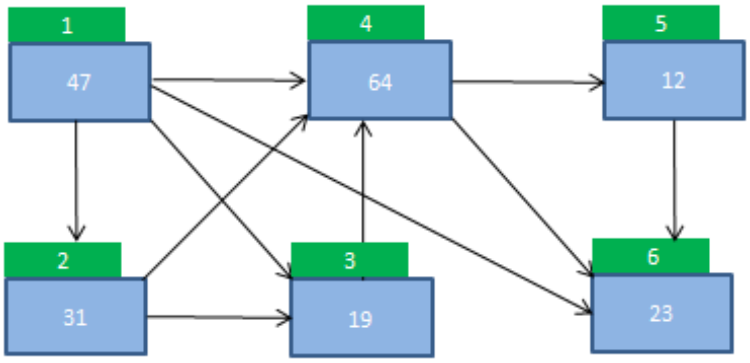


Figure 3. The sub-topic association diagram of the "360VS Tencent" event

The green block represents the cluster category (subtopic) number, and the number in the blue block represents the number of posts contained in the subtopic. Due to the large degree of some nodes, the above figure is optimized, the threshold is increased to 0.55, and the maximum out degree and in degree of the node are limited to 2 after optimization, the following figure 4 is obtained:

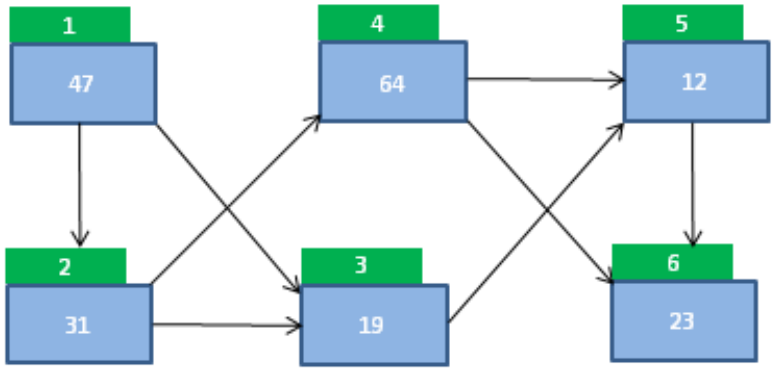


Figure 4. The sub-topic association diagram of the "360VS Tencent" event (after optimization)

After getting the real sub-topic association graph, we use the F1 value composed of the accuracy rate P and the recall rate R to evaluate the experimental results. Accuracy and recall are the two most basic indicators in the fields of information retrieval, classification, recognition, and translation, and they are also the most commonly used indicators. Accuracy refers to the ratio between the number of correct edges in the experimental result graph and the total number of edges in the experimental result graph. The recall rate refers to the ratio between the number of correct edges in the experimental result and the total number of edges in the real graph. The accuracy rate and the recall rate influence each other. Ideally, we must try to achieve both as high as possible, but in general, if the accuracy rate is high, the recall rate will be low, and if the recall rate is low, the accuracy rate will be high. Therefore, the F1 value can be used to measure when both requirements are high. The relationship between F1 value and P and R is as follows:

$$F1 = \frac{2PR}{P+R} \tag{9}$$

It is calculated that the accuracy rate of the optimized sub-topic association graph is 0.83, and the recall rate is 0.74. This article adjusts the values of parameters a and b to make the accuracy and recall rate reach the relative maximum. Figure 5 below shows the effect of different combinations of values of parameters a and b on accuracy and recall:

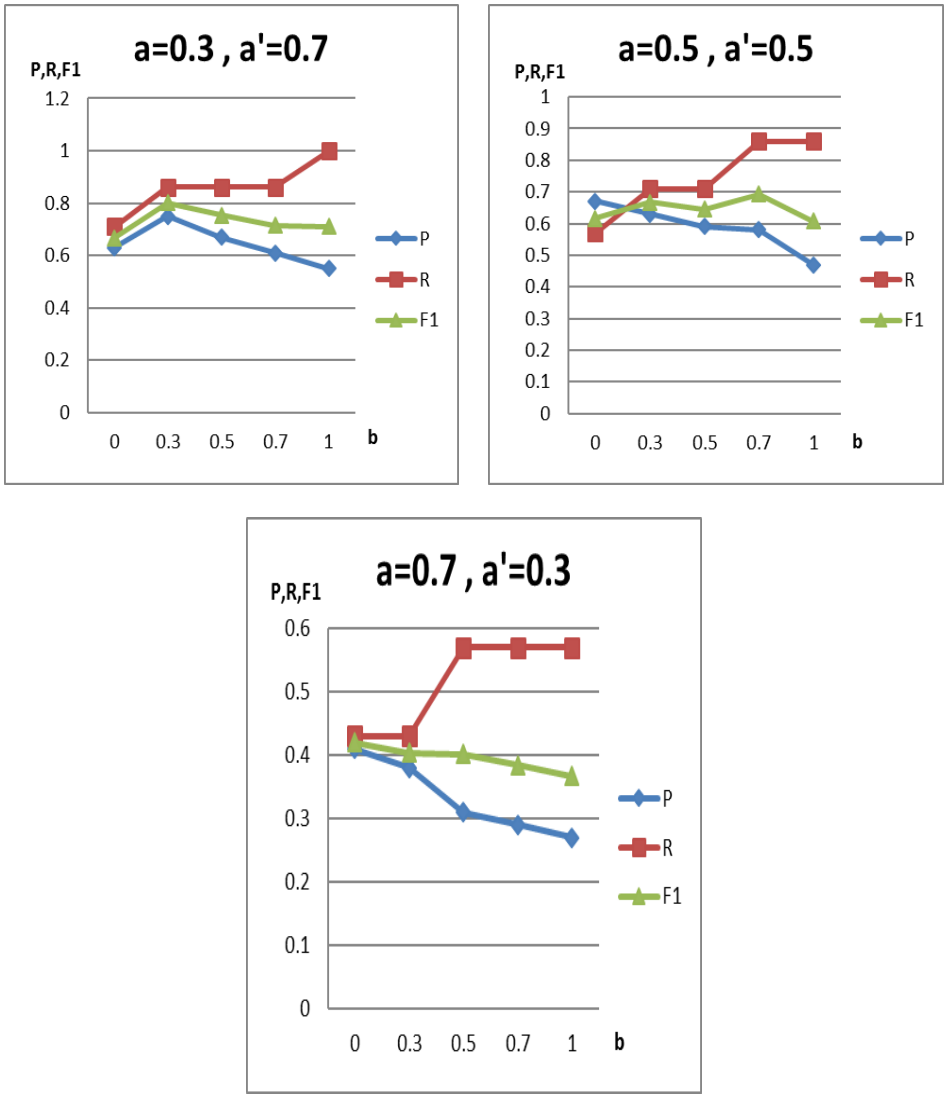


Figure 5. The effect of different parameter combinations on accuracy and recall

5. Conclusion

This paper establishes a super-network model composed of social sub-networks, environmental sub-networks, psychological sub-networks, and opinion sub-networks, and applies the hyper-link prediction algorithm to the model, revealing the reasons and driving mechanisms for the formation and evolution of network public opinion, which is a guide the selection of objects provides a theoretical basis. Introduce the "leader" super edge to implement guiding measures on the public opinion super network to quantify the guiding strategy, and through the changes of various elements in the super network and the linkage mechanism, the function and feedback law of the guiding strategy are obtained, so as to formulate a completer and more effective the intervention and guidance strategies provide a reliable basis.

References

[1] Luo Ping, Wu Bin. Network public opinion big data dissemination feature mining system based on artificial intelligence. Modern Electronic Technology, vol. 43, pp. 4-11, April 2020.

[2] Cao Ruijuan, Jiang Rengui, Xie Jiancang, et al. Public opinion monitoring and evolution mechanism of urban waterlogging network based on big data. Journal of Xi'an University of Technology, vol.36, pp. 8-12, February 2020.

[3] Hu Xinjie, Lu Yunan, Lu Chuan. Research on the tendency analysis of Internet public opinion based on cluster analysis. Acta Sichuan Armamentary Engineering, vol. 40, pp. 115-118, May 2019.

[4] Xiong Xianlan, Luo Guangyuan. Discussion on Social Public Opinion Governance Countermeasures for Food and Drug Emergencies in the Context of Big Data. Journal of Hunan University of Science and Technology: Social Science Edition, vol.23, pp. 8-11, March 2020.

[5] Li Jie, Chen Siyu, Zhang Jingwen, et al. Research on the Spatiotemporal Relationship between the Geographical Spread of the Epidemic and Internet Public Opinion Based on Big Data. Geographic Information World, vol. 27, pp. 5-12, March 2020.

[6] Bai Rong, Wang Dongxin. Design of Internet Public Opinion Monitoring System Based on JavaEE. Electronic Design Engineering, vol.28, pp. 5-9, February 2020.

[7] Zhang Silong, Wang Lancheng, Lou Guozhe, et al. Frontier evolution and development trend of online public opinion research in the era of big data. Chinese Journal of Medical Library and Information, vol.28, pp. 50-55, January 2019.

[8] Wang Duo, Wang Xiwei, Jia Ruonan, et al. Big data-driven social network public opinion ecological evaluation and empirical research. Information and Documentation Services, vol.41, pp. 8-12, February 2020.