ATLANTIS PRESS
Advances in Engineering Research (AER), volume 135
2nd International Conference on Civil, Transportation and Environmental Engineering (ICCTE 2017)

Analysis of micro-blog public opinion based on complex network

Lixia Yao

LinYi University China

yaolixia@lyu.edu.cn

Keywords: Co-word network matrix, Complex network, Microblog, Public opinion analysis

Abstract. With the popularity of social media applications, micro-blog has become the main carrier to reflect social public opinion. On the one hand, the extensive application of micro-blog has a positive effect on the development of national social and cultural aspects, on the other hand, it will produce some negative effects. Therefore, it is necessary to find and analyze the hot spot public opinion information in a timely and accurate manner. In this paper, we use the method of complex network and co word network to construct the co word matrix of micro-blog, which provides a method for the analysis of micro-blog's public opinion.

Introduction

At present, micro-blog has become an important research topic of public opinion. While in the traditional analysis of public opinion research, the main research object based on Web links between different sites, but this association established mostly by managers to establish web site operators, it is difficult to show the characteristics of the free flow of information to the user oriented, and this link is of non frequent. Whether it is public opinion topic link network based tracking, or using PageRank or HITS algorithm mining potential information web links based on public opinion, there are some disadvantages: firstly, there will be users cheating or advertising the use of a large number of links rather than content to obtain attractive behavior; secondly some posts has timeliness, link analysis takes time the accumulation to get into the chain, and this post may have lost time. On the other hand, micro-blog's unique features are mostly limited to the user's information released in each of the 140 characters within the scope of the short text based public opinion analysis challenges. Micro-blog in most of the information is scattered, high noise, random and fragmented, resulting in high dimension of document vector space matrix is constructed, but the data is sparse, so the traditional text clustering method based on VSM (Vector Space Model) on the micro blog hot spot detection in poor. In order to overcome the shortcomings of the link based method and short text clustering in the analysis of micro-blog public opinion, this paper attempts to apply the co-word network and complex network method to the analysis of micro-blog's online public opinion.

Research methodology

For the same website users, such as Sina micro-blog in a large number of users, or between them because of different interests, or because of a common topic, or because of the interaction between users Reply of subjective or objective, this link is large, frequent and independent. With respect to the limitation of the traditional Internet links, characteristics of network links will bring great convenience to the network public opinion data analysis, by using between text analysis and user node connection can be established co-word network and complex network model, and the characteristics of the situation and on the basis of micro-blog.

Co-word network

At present, most of the research methods of web public opinion research are based on Web page text clustering, which is mainly focused on the text information such as longer news reports. For short text clustering in micro-blog, although there are also some feature extraction and factor analysis method for dimensionality reduction, but the factor analysis are often the generalizability of the original data, lost the ability to explain, not the clustering results will be reduced to public opinion hotspot. Some scholars try to use external resources to solve the problem of short text clustering, such as each short text submitted to the search engine, with the help of the retrieval results of extended clustering, or concept with Wikipedia's cluster expansion. Although it has achieved good results, but the process is complicated and the efficiency needs to be improved, it is difficult to apply in the analysis of micro-blog's public opinion with massive data.

Based on the above analysis, in order to overcome the shortcomings of short text clustering, this paper attempts to solve the problem of micro-blog public opinion hotspot detection by using the co-word network method of Co-word analysis. Because the word network is a kind of cognitive meaning of knowledge network, so we can use the word co-occurrence relations to achieve the purpose of reducing the focus of public opinion, through co-word links between strength and co-word network diagram of the subgroup analysis, hot division, public opinion hot words set segmentation by clustering, representation of public opinion hotspot.

Complex networks

The Internet is also one of the most complex networks, which can be regarded as a complex network structure. As a typical representative of the Internet, the network of public opinion information is in line with the typical topological characteristics of complex networks. The main rules of distribution of nodes in the complex network is small world and scale-free, small world characteristics, micro-blog community social networks in complex networks with scale-free characteristics and high degree of aggregation features. Between the small world topology index of public opinion information dissemination network and scale-free network, the scale-free property can measure the breadth of the spread of public opinion, reflect the influence degree of the spread of public opinion by the core node, the node "authority" release the information more easily by its neighboring nodes attention and forwarding.

The characteristics of small world network can measure the depth of the spread of public opinion, indicating that the distribution of public opinion communication between nodes of the equilibrium, the number of hops will not be unlimited extension. Therefore, we can measure the complex network characteristics of public opinion information by the two indexes of small world effect and scale-free property.

The scale-free nature of the cause of edge connections most of its nodes only have less complex network public opinion field, but there are a few nodes with Hub large number of edges, these Hub nodes of the complex network communications hub, so it can pass through the index node center to measure and find public opinion leaders in micro-blog. In addition to the visual network, this paper uses the following quantitative indicators to measure and discover the leaders of public opinion.

This paper studies the basic parameters of complex network topology, which is used to describe the direct influence of nodes in the network. The ability to establish a direct link between the node and its surrounding nodes can be used to measure the influence of the leader of the public opinion on the person who is directly in contact with it. If the formula (a), D (x) is the number of edges directly connected to the node x.

$$C_d(x) = d(x) \quad \text{(a)}$$

Micro-blog public opinion analysis

Micro-blog of a plate under a large number of users of micro-blog news released different definitions of the set $M_i=\{U_i,D_i,f_i,c_i\}$, where U_i is the user ID, micro-blog D_i for text messages, f_i is the number of forwards, c_i is the number of comments. In addition, all micro-blog text messages constitute a text message set D. The word set is W. XML data is used to store the results of data extraction, and in order to construct the co word matrix, the document is analyzed and processed. In order to assure the validity of the results it is necessary to use some standard of data extraction from D document set and the word segmentation results set W word set are selected and cleaned with different hot performance words and construct weighted co word matrix.

(1) Although micro-blog message length is generally short, but usually a useful information is unable to use a few words clear, so we set the corresponding threshold θ_1, excluding the text length is less than θ_1 characters Length (Di) $<\theta_1$, is commonly used to shield but for public opinion hotspot detection meaningless message or micro-blog reply to posts and non-text information, pictures, video and other information.

(2)Based on information entropy theory. The smaller the uncertainty of the source, the less the amount of information transmitted, in order to exclude the impact of spam, advertising information, can be used to eliminate the impact of the simplified information. The amount of information γ for the total number of simplified definition of a number of different words in a micro-blog news in proportion, such as formula (b). In order to ensure the accuracy and validity of the result, the value γ is very small.

$$\gamma_i = \frac{|distWords \in D_i|}{|allWords \in D_i|} \quad \text{(b)}$$

(3) Set the corresponding threshold θ_2, excluding the results of the word whose frequency meets Frequency(ω_i)> $<\theta_2$, such as only one or two words is obviously not representative. And then the results of the word segmentation in the frequency is too high does not have the information value of the results of artificial rejection.

(4) Each micro-blog message will have a different amount of forwarding and comment. Comment on the volume of the spread of a message, the more comments, indicating that the more people see the news, the wider the scope of its impact. The forwarder can send the message to his circle of friends by forwarding the message, so that the first level down. Thus, the weight of the same word that appears in different messages has different effects on public opinion. So in the construction of co-word matrix, for each message to give the corresponding weight, the weight is also a common word in the (W_i,W_j), the global weight. In this paper, the values α and β are taken as 0.6.

$$\text{Weight} = \frac{2(\alpha f_i + \beta c_i)}{\max(f_i)+\max(c_i)}, \alpha + \beta = 1 \quad \text{(c)}$$

(5) Finally, construct the co-word matrix, where each element in the matrix of W_{ij} W_i and W_j word co-occurrence frequency weighted value, namely $W_{ij} = \sum Weight(W_i, W_j)$, when i=j said W_i alone the weights of the accumulated value of record for W_{ij}. In order to eliminate the influence of the difference of data order on the result, the data of each atom in the matrix is normalized. Since the co-occurrence relation between words is undirected, the co-word network is a weighted

undirected network. So far, the construction of data cleaning and co-occurrence matrix is finished.

References

[1]Adoption of different strategies in diversity-optimized populations promotes cooperation[J] . Lei Sun,Dong-Ping Yang. Physica A: Statistical Mechanics and its Applicat . 2014

[2]Effects of payoff-related velocity in the co-evolutionary snowdrift game[J] . Zhihu Yang,Zhi Li,Te Wu,Long Wang. Physica A: Statistical Mechanics and its Applicat . 2014

[3]The influence of local majority opinions on the dynamics of the Sznajd model[J] . Nuno Crokidakis. Journal of Physics: Conference Series . 2014 (1)

[4]Modeling the Forming of Public Opinion: An approach from Sociophysics[J] . Serge Galam. Global Economics and Management Review . 2013 (1)

[5]Social selection of game organizers promotes cooperation in spatial public goods games[J] . Yongkui Liu,Xiaojie Chen,Lin Zhang,Fei Tao,Long Wang. EPL (Europhysics Letters) . 2013 (5)