Network Public Opinion Analysis and Research were Reviewed

Qun Gu[1], Yifan Jia[1], Xiaohong Hao[1,a*]

[1]School of Computer and Communication, Lanzhou University of Technology, Lanzhou, 730050, China

[a]lzgq66@163.com

Keywords: network public opinion, public safety, method

Abstract. With the development of computer science and technology, the internet has gradually come into all aspects of society, the popularity of all kinds of social software and the development of network media makes the information transmission speed is greatly increased and creates a network of public opinion events. Public opinion events involved in all aspects of society, the public opinion monitoring and analysis of incidents in public safety, economy, politics and management has an important significance on various aspects. In this paper, relevant technologies of the network public opinion analysis are summarized, citing the problems and solutions in the literature, and some of these methods are evaluated. Then on the basis of the existing network, public opinion analysis techniques for autonomous regional public security in our country is prospected.

Introduction

Public opinion is short for "public opinion", refers to the certain social space, surrounding the occurrence and development of the intermediary social events and change, as the main body of people as the object of social management, enterprises, individuals and other kinds of organizations and their political, social and moral orientation and hold social attitudes. It is more about the various phenomena in the society, the public expressed beliefs, attitudes, opinions and emotions, and so on the sum of the performance. Network public opinion is a kind of public opinion. Is carrier of the generation, transmission, network, network public opinion is the sum of the attitude of netizens, comments and opinions.

In the world today is in the rapid development of network information science as the representative of the third industrial revolution in the new era, the Internet has gradually become one of the main channels of information dissemination. By the end of 2015, the global Internet users will be close to 2015, the Internet has become the people to express personal subjective view of one of the main channel. In this context, the network public opinion analysis technology arises at the historic moment, especially in the public security, finance and other aspects can play an important role. Government agencies, monitoring, analysis and research of network public opinion and correct guidance of public opinion has significant social and political significance.

1 network public opinion analysis method

In the field of network public opinion analysis, study abroad started earlier, a series of achievements, mainly through the review meeting in the form of giving impetus to the development of the whole field. Results obtained are: conference topic detection and tracking, information retrieval for professional meeting and text information retrieval. In domestic, the research of network public opinion analysis starts relatively late, but have also made some achievements. Network public opinion analysis of relevant technology mainly covers the following aspects: topic detection and tracking (TDT) technology, network public opinion data acquisition technology, mass data storage technology, web page text extraction technology, web based text mining technology, the analysis and prediction of public opinion in related technology.

1.1 topic detection and tracking technology. Topic Detection and Tracking technology is a major means of network public opinion monitoring, first began the research of this technology in the United States. The TDT technologies into story segmentation and topic detection, topic tracking detection and correlation detection is reported five tasks, runs throughout the network public

opinion analysis and prediction of the whole process, in the hot public opinion and emergency public opinion play a crucial role in the discovery process. The research process of this technology is in the public opinion, on the basis of information collection, the reports were collected from the Internet, shard to report data flow after the split into separate reports, topic detection model is established through the text clustering algorithm, to judge the semantic relationships between input reported correlation detection, and found that the semantic relationship stronger, into the same topic focus as reports have found [1]. This process can be understood as a data filtering process, the key lies in data acquisition and text clustering and data storage. In the network public opinion analysis and prediction of abundant information storage technology, can use text clustering process can use ant clustering algorithm and its improved algorithm, similarity calculation can be through the establishment of the traditional calculation method of vector cosine vector space model.

1.2 the network public opinion data acquisition technology. Internet search engine search, effective methods of gathering data from the Internet, search engine is one of the core technology of web crawler technology. Working principle is through a web crawler technology called web crawler program automatically, automatic individual nodes constantly crawling on the Internet, start from one or several initial URL scraping of the page, and constantly new URL can be drawn from the current page to join the queue, until meet certain conditions[1]. the web crawler technology is very suitable for network public opinion automatic collection of data.

In addition, for users, subject-oriented queries the general web crawler program cannot meet the requirements, focused crawler arises at the historic moment. Different from traditional web crawlers, focused crawler is not the pursuit of large coverage, and the target as the grab is associated with a particular topic content page, data resources for subject oriented user query, focused more on the accuracy of web scraping. Focused crawler work process is relatively complex, need according to certain web filtering analysis algorithm has nothing to do with the theme of links, retain the useful links and put it in waiting for fetching the URL of the queue, according to certain search strategy choice to grab the next web URL from the queue, and repeat the process, until it reaches an system of a certain conditions when to stop. In addition, all the spiders crawling web pages will be storage system, a certain amount of analysis, filtering, and indexing, so that after the query and retrieval. For focused crawler, the process of analysis results can also be for the next phase of the process give feedback and guidance.

1.3 mass information storage technology. The current background, the network public opinion data into order of magnitude of geometric growth, and public opinion of the data collected into a geometric growth [1]. And these data needs to be stored in the local form putted forward and long-term preservation for future use. In the face of such a large data storage capacity is not the general relational database based on SQL can afford got[1]. So, need more large capacity storage equipment and database to finish[3]. Massive information storage technology is the baidu, Google, one of the core technology such as amazon. The technology at the same time can be applied to public security, frontier defense and other government agencies and public opinion information security department store. The current mass information storage technology mainly divided into two aspects, storage mode and large-scale database.

Facing massive information storage model mainly adopts distributed storage mode, the current distributed storage mode are divided into direct attached storage mode and network storage mode. The network storage is divided into network attached storage and storage area network. Direct attached storage through the server and the network connection, can cause additional load to the server, network storage is directly connected to the network, avoid to cause server load[3].

Large-scale database management system to meet the demand of the high concurrency, speaking, reading and writing of the database, the demands for efficient storage and access of data, as well as to the database of high scalability and high availability requirements. The main types of database to Oracle, SQL Server as a representative of traditional relational modle database (RMDB), as well as with BigTable, HBase and Mongo DB as a representative of the non-relational database (No SQL, not only SQL), both of which has realized the database data distributed storage and parallel processing operation. Relatively, the traditional relational database data structure used to store some

formatting, although easy to connect between table and table and operation, but is hampered by its performance. Rather than a relational database to the key value store, not limited to a fixed structure, the connection does not exist in the operation, such as the actual part reduces the time and space of overhead. Enter the web2.0 era, a large amount of unstructured data appears on the Internet, the SQL operations on data is significantly weakened, this objectively prompted the non-relational success[3].

1.4 text mining technology based on web. Based on web mining technology, is a branch of data mining technology field, is a special mining technology for Internet environment. Web mining is a cross discipline, covers the statistics, machine learning, artificial intelligence, bioinformatics, and other research areas, its purpose is in the Internet data found potential, valuable information or model. According to the different Mining object oriented, web Mining can be divided into three categories: web Usage Mining, web Structure Mining and web Content Mining[3]. In network public opinion analysis and prediction of the related research, because of its object oriented is the body of the message content web text, and therefore belongs to the category of the web content mining [3].

With general data mining technologies, web content mining is the basic technology of classification and clustering technology, its process generally includes web mining and web preprocessing, text model representation and text classification or clustering step [3].

Text clustering analysis is a key step of data processing is more, can be divided into space concept semantic similarity degree, based on the combination of Support Vector Machine (SVM) and unsupervised clustering method.

1.5 information filtering technology. Information filtering technology is mainly used in public security network public opinion information acquisition system, used to according to the needs of users to filter and dynamic information feedback to the user[4].

Public opinion information collection, public security intelligence officers through Boolean logic based on keyword matching to collect data, tend to get beyond the imagination and absorptive capacity on the number of collecting results, but also contains a great deal of data and user needs far and noise data, literature[4] proposed information filtering technology can effectively solve this problem. It is through the intelligent constant learning master user demand information extraction technology, extract of interest to the user information, filter the useless information[4].

Public security intelligence information needs and preferences are relatively stable, you can see information filter with the aid of information technology, the public security intelligence can be constant and rapid increase of information from the Internet and complicated in more accurate access to the required information technology through a growing demand for intelligence personnel to understand learning mechanism, and the User Profiles in the form of a record, and according to the record to the collected information filtering[4]. And intelligence can decide for themselves whether to send feedback to the filter to indicate what information in accordance with their needs, after such repeated study, adjustment, will eventually make filters filter rules root used more and more close, forming a virtuous circle.

2 network public opinion analysis and forecast the key problems and solutions

2.1 network public opinion analysis and prediction problems. At present, the research of network public opinion analysis and detection at home and abroad have achieved fruitful results, but there are still some problems unsolved. Literature[5] the question:

(1) under the Internet environment, information spread fragmentation caused Internet users is difficult to obtain comprehensive information. Under the Internet environment, Internet users at the same time, the pursuit of personality in the dissemination of information, Internet culture structure, age structure, regional structure, and the difference of social class structure, causing them to focus on the same events, select the spread of information and information are also different, eventually led to the spread of information fragmentation. Information spread fragmentation caused Internet users is difficult to obtain the comprehensive information, but get is discrete, partial bits of information. One-sided information also provides people with ulterior motives with wanton play, the space of false facts, under certain conditions can easily evolve into network public opinion.

(2) the plane topic model to describe the semantic relations between seed events and derivative events that lead to related event detection difficult. Existing related event detection method using the core words as the basis of event correlation, the scheme to detect the seed events and the relationship between the derivative events, and whether related between derivative events is hard to detect. Particularly when the derivative events are closely related with seed events, seed traits became derivative correlation detection of interference between events. How to eliminate the disturbance of seed events, event detection seeds and the relationship between derivative events, is proposed to solve the key problems.

(3) the existing methods ignore the impact of Internet users emotional evolution and spread of events, results in the decrease of incident detection accuracy information dissemination process is essentially the netizen emotional communication and collision process. Sudden incident, Internet users to get information from the Internet can spread information at the same time, in the process of information transmission they tend to attach panic, anger, or other negative emotions. In the process of information transmission, Internet users emotion is amplified and empathy, eventually led to the group of phenomenon, stimulate the network public opinion incident broke out.

2.2 solution. According to the above problem, literature [1] proposes the following solutions:

(1) in this paper, a class based on the improved ant clustering algorithm of partitioning strategy Devoid of the algorithm by constructing recycle cosine vector space Angle to calculate the average similarity method, effectively avoid the characteristics between semantic missing problem.

Average similarity calculation data object and its surrounding objects can be thought of as the process of clustering decision process, can be described as for fragments of reports set D一 {$d_1,d_2,d_3,...,d_n$}, giving a fragment d_i, calculate the average between d_i and its coverage around the d_j similarity, the greater the average similarity, the greater the chance that it belongs to the category. Based on this, the definition of loss function f (d).

$$\text{cost}(d_i, d_j) =$$
$$\min\{I(C_{before}, d_i) - I(C_{after}, d_j)\} \tag{1}$$

Reports will be involved in clustering fragment mapping in the feature space, Assuming that fragments of d_i report feature space with W^d, said the loss function into

$$\text{cost}(d_i, d_j) = \min\{I(C_{before}, W^d) - I(C_{after}, W^d)\}$$
(2)

For formula (2) is derived, by deduction, eventually form the average similarity function

$$f(o_i) = \max\{0, 1 - D_{kl}(f(d_i, W_i)||f^{medi}(d_i, W_i)\}$$
(3)

After the average similarity function is deduced. The sigmoid function is selected as a probability transformation function. Then a random moving ants "picked up" the probability of a data is:

$$P_p = 1 - sigmoid(f(o_i))$$
(4)

A randomly moving ants "put down" the probability of a data is:

$$P_d = sigmoid(f(o_i))$$
(5)

$$sigmoid(f(o_i)) = \frac{1 - e^{-cf(o_i)}}{1 + e^{-cf(o_i)}}$$

(6)

c for convergence parameters.

(2) put forward a related event recognition method based on physical dependence.

In child topic classification on the basis of the characteristics of the subject is divided into seed features and derivative two dimensions, based on hierarchy event model, from two dimensions of seed features and derivative build the document the event correlation between clues.

(3) put forward a kind of emergency online identification method based on user emotions.

Literature[5] proposed a fusion of users emotional incident detection method. As the topic hierarchy model is constructed, the method and can depend on the characteristics of the temporal constant adjustment model driven mode, to achieve the goal of on-line identification emergencies. Literature[6] is calculated by the short text emotional value, emotional value as a feature to fragmentation of events, DTM evolution on the document after fragmentation analysis, but this method when calculating the emotional value with a certain subjectivity, such as if the degree adverbs in the essay, the weight value of artificially set will not be able to show words the real weight of the weight difference between different degree adverbs are difficult to control, the classification of adverbs of degree also vary from person to person, therefore, if there any error in the calculation of the emotional value will directly affect the results of accurate.

3 summary and outlook

In the current under the background of information explosion, the network public opinion analysis is not only a big Internet companies such as baidu, Google's core technology, is a core research topic in public security, frontier defense and other departments. Network public opinion monitoring has important significance for maintaining social order. For example, in xinjiang, Tibet separatist activities such as frequent area, can set up a network public opinion monitoring system based on identification of minority languages, can effectively to analysis and prevention of emergencies, and control the development, can effectively combat terrorism. At the same time can set up a corresponding network public opinion monitoring system based on a foreign language, is used to identify to the foreign public opinion information, effectively inhibit the invasion of foreign hostile forces, homeland security, the protection of state secrets.

In view of the current domestic and international situation, anti-terrorism has risen to regional and national, state and national level, the xinjiang uygur autonomous region and the Tibet autonomous region in China is facing a more serious situation, it is necessary for a few of the people's opinion and public opinion monitoring events. National minority area society, economy and the eastern and coastal areas in China is different, is a certain religious and ethnic minority areas of ethnic groups also have a certain complexity. As a result, the autonomous region public opinion report data has obvious characteristics of the improved information filtering technique may be used for data acquisition, and then the text clustering form the storage to the local case, the data storage need to use the large-scale database.

References

[1]Fan Shaoyu. Based on the mass of public opinion information network public opinion analysis system design and implementation[D]. South China university of technology,2013.

[2]Wang lei. The public security network public opinion analysis system research [D]. Beijing jiaotong university,2008.

[3] Fei Shaodong. Network public opinion incident detection and tracking related technology research[D]. Shandong normal university, 2015.

[4] He Tianxiang,Zhang Hui Li Bo,Yang Chunming,Zhao Xujian .A network public opinion evolution analysis method based on sentiment analysis [J]. Journal of software Tribune, 2015.

[5] Ding Zhaoyun. Facing the microblogging public opinion influence analysis of key technology research [D]. National university of defense technology, 2013.

[6] Wang Wei, Xu Xin. Based on the clustering of network public opinion hotspot discovery and analysis [J]. Journal of research and analysis of intelligence, 2009.

[7] xiao-bin Huang, Zhao Zhao The application of text mining in the network public opinion information analysis [J]. Journal of intelligence science, 2009.

[8] Ding Baiquan, XiaoYanyan. New media context and method of the terrorist attacks of public opinion research [J]. Journal of tianjin academy of social sciences, 2015.

[9] Wang Juan. Network public opinion monitoring analysis system to build [J]. Journal of changchun university of science and technology, 2007.

[10] Lin Xiaohua. Improve the ability of emergency public opinion guide path envisaged in national regions [J]. Journal of southwest university for nationalities, 2011.

[11] Neresini F, Lorenzet A. Can media monitoring be a proxy for pubfic opinion about technoscientific controversies? The case of the Italian public debate on nuclear power[J].Public Understanding of Science, 2014,6(10): 63一66.