A Method of Public Opinion Analysis in Big Data Environments

Cuiyuan YU[1]

1Binzhou Vocational College
Binzhou 256603,China

***Abstract* — With the rapid development of online social networks, such as social networking services, microblog blog, etc., a real new media age has gradually arrived. Everyone can create their own web content and spread out quickly through online social networks. Massive data has brought severe challenge to public opinion monitoring. We will introduce a microblog public opinion monitoring system based on the Hadoop. It can mine and analyze large scale collected data, realize detection and tracking of hot topics perform social network analysis on the microblog. Network public opinion analysis focuses more on the huge amounts of data collection, storage, cleaning and text clustering, so the traditional public opinion research method only based on data mining is no longer held. This paper summarizes the related literature of the research, sums up the basic process of network public opinion analysis framework, and how to use the distributed computing in all stages of text clustering, thus improving the accuracy and timeliness of the network public opinion analysis. The proposed system will provide automated, systematic, and scientific information support for party and government organizations, enterprises and other units and organizations to detect sensitive information timely, grasp the hot points and the trend of public opinion.**

Keywords- Big data; Hadoop; HBase

I. INTRODUCTION

Public opinion is a subjective reflection of social reality in a certain period of time [1]. It is a comprehensive expression of the group's attitude, thought, emotion and demand. Public opinion [2] is in a certain social space around the intermediary social event occurrence, development and change. Internet openness, the characteristics of the virtual make the speech reach an unprecedented level of activity. Analysis of network public opinion under the big data has become a hot issue of the current government and research institutions. With the rapid development of the Internet, the worldwide Internet users are rapidly expanding; while there have been a number of excellent online social networks, which are the most active social networking sites and microblogging. With the massive microblogging messages constantly being created, how to tap these vast amounts of data, analysis, and dissemination to achieve sustained track sensitive information and public opinion trends hot topic judgments has become an important research direction and challenges. The traditional method of public opinion analysis based on data statistics is no longer applicable [3].

Traditional public opinion monitoring systems are based on expensive workstation or server cluster. In the face of massive data, it is often shown as high cost, poor scalability, single point communication failure, etc. Coupled with the traditional database is difficult to manage and batch processing hundreds of millions of data records. We give an implementation of public opinion hot topic discovery and tracking.

This paper summarizes the relevant literature on the analysis of large data network public opinion, and summarizes the basic process framework of public opinions analysis. We put forward the solution of each stage of network public opinions mining based on the big data environment, and construct the basic model of public opinion mining for social networks to improve the accuracy of mining public opinions.

II. RELATED WORKS

From the existing public opinion analysis of the relevant literature, it is not difficult to find that social public opinion analysis has gone through several stages. It includes simple l public opinions mining for large data environments. Simple social public opinion analysis mainly analyzes the relationship between the current hot issues, the government decree and social public opinions mining. Simple social public opinion analysis mainly is taken through the questionnaire survey to obtain the original data analysis. MacLennan et al [4] studied the attitudes of New Zealand people to alcohol policy by sampling survey. Alan et al [5] took the Gallup world opinion survey data to study the relationship between terrorist attacks and people's attitudes. Network public opinion analysis is accompanied by the rise of Facebook, micro-blog, WeChat, Twitter and other social networking platforms and applications.

First of all, because of the openness of the network, a lot of information is generated every day. Second, the development of multimedia makes the data have a variety of forms such as text, video, pictures, audio, etc. Based on the analysis of the characteristics of large data network public opinion, there are several network public opinion analysis methods: public opinion analysis based on web log data mining, based on the social network analysis of the relationship.

Few studies is on the technical aspects of large data network public opinion analysis.

DOI 10.5013/IJSSST.a.17.10.16

ISSN: 1473-804x online, 1473-8031 print

III. BIG DATA TECHNOLOGIES

The arrival of the era of big data to the existing data processing technology has brought great challenges. Based on the characteristics of the current diversity of large data, data storage and data processing have raised the corresponding solutions. In the aspect of data storage, the data storage method of network public opinion analysis is still in the traditional SQLServer, ORACLE, and Sybase and so on.. Traditional structured database cannot meet the current rapid large variety of data storage requirements. At present, there are three kinds of big data storage technology: distributed file storage system for massive unstructured data. In terms of data processing, the current parallel processing and cloud computing is a more efficient way to solve large data computing. Hadoop [6] is a major technical means for the current academic and enterprise to solve large data storage and analysis. It is composed of HDFS, MapReduce and HBase.

(1) HDFS

HDFS is the most basic position in the whole Hadoop architecture. It includes the client, the name node and the data node. Name node is a manager of a distributed file system. It is mainly responsible for the namespace of the file system, the configuration information of the cluster and the information of the data block.

(2) HADOOP

Hadoop [7] is an open source distributed system based on the Apache architecture. Due to its great performance, it has been widely used in many areas. Hadoop is mainly composed of HDF, MapReduce and HBase. HBase is taken as a massive collection of data storage database. It includes microblogging information collection, micro-blog public opinion monitoring and user interaction three layers.

(3) MapReduce

MapReduce [8] is a distributed computing model, which package the details of parallel computing, fault tolerance processing, localization computation, load balancing and so on. Through this interface, it can automatically calculate the amount of large data by parallel and distributed execution.

(4) HBase

The HBase [9] Hadoop Database is built on HDFS, and column-oriented open source distributed database system is the open source implementation with Google Big table. HBase provides a simple API interface for storing and managing data.

IV. MICROBLOGGING PUBLIC OPINION MONITORING SYSTEM STRUCTURE

Microblogging public opinion monitoring system consists of information collection module, information analysis module, the index storage parts, monitoring and mining parts, interactive parts.

(1) Microblogging contents collection part: getting contents and information.

(2) Microblogging information analysis module: information extraction, web page elimination, text segmentation.

(3) Index memory module: Provides Hadoop distributed data operation interface.

(4) Public opinion monitoring and analysis modules: a text representation of the index database, HBase cluster analysis, social network analysis.

(5) Interaction module: Based on the J2EE architecture for user interaction.

Microblogs are collected by the method API and web analytic solutions based on literature proposed by [].Open API refers to a micro-blogging service offer their own service package into a series of API Interface. We can access these data interface to get micro-blog contents, comments, users, relations and other information. To balance server load, micro-blogging service providers a method for different users to set different API interface to call the query frequency range. We used a combination of API and web based analysis program to collect micro-blog information.

(1) Acquirer: by calling the API to return JSON format interface information for collecting bloggers.

(2) Crawler: grab micro-blog content through a distributed crawler, and use Dom to parse HTML and extract information.

The N access devices and crawlers are running on the N slaver machines, the scheduler runs on master machines.

This system uses the word frequency, inverse document frequency TFIDF vector representation feature vector to represent the content of micro-blog. This method is based on all the words in micro-blog's content without considering the context and the structure of the text. So in essence it is also a kind of micro-blog content word set representation. A micro-blog content TFIDF feature vector to some extent reflects the content characteristics of the micro-blog. The column in the matrix is the feature set, and the row set is the collection of micro-blog content that has been crawled.

Public opinion monitoring and analysis as the main module of the system, it contains the latest news, hot topics, hot bloggers, active bloggers tracking, geographical tracking, communication path analysis, trend analysis, social network analysis.

Hot topic is based on the structure of the above mentioned feature matrix to do clustering analysis.

(1) Read the feature element from the feature matrix;

(2) Get the data center based on the Canopy algorithm by MapReduce;

(3) K-Means algorithm based on MapReduce algorithm is taken to calculate the distance between data objects and cluster centers;

(4) Clustering results of each center and sub key are written into the analysis library.

Complex network is a real network which is discovered by people in recent years on the basis of the rapid development of computer processing and computing power. It has many statistical characteristics, which are different from regular network and random network. One of the most important three features is the characteristics of small world, scale-free characteristics and clustering coefficient. First of all, the concept of the network G = (V, E) is used to describe the relationship between the object. Secondly, the main methods of studying complex networks include social network analysis methods. It is a kind of special network in the research field of complex network [10]. In this paper, the node degree distribution and clustering coefficient are used to analyze the no-scale and small world characteristics of micro-blog network.

Each node in the microblogs represents a user. For a user, the number of other users is called friends. The in-degree of a user refers to the number of its followers. And the out-degree refers to the sum of its followings. The node degree distributions throughout the social networks can help understand the structure of social networks.

For the social network, the group form is an important feature. The social network analysis method used in this system is to calculate the number of fans in the main information and the number of the degree.

V. Network public opinion analysis technology

Network public opinion analysis mainly involves the data collection, text segmentation, text vector representation, text clustering classification.

(1) Vector space model

Vector space model is a model that can be expressed as a vector form in unstructured text.

$$v(d)=(t_1,w_1(t_1);\ldots;t_i,w_i(t_i);\ldots t_n,w_n(t_n)) \qquad (1)$$

Here, the t_i is the keyword of documents. The $w_i(t_i)$ is the weight of keywords.for documents.

How to determine the weight of keywords and key words is the key factor for the analysis of network public opinion. The TF-IDF index is taken in the paper to determine the key words and weights of the network text. A normative representation is generally divided by the number of documents that contain the word.

(2) Text similarity computing

Similarity computing method includes the Hamming distance, Euclidean distance, and cosine distance.

$$C(A,B)=1-\frac{\sum_{i=1}^{n}a_i*b_i}{\sqrt{\sum_{i=1}^{n}a_i^2}\sqrt{\sum_{i=1}^{n}b_i^2}} \qquad (2)$$

Here, a_i and b_i are the items of the documents *A* and *B*.

(3) Text clustering algorithm

There are a lot of texts clustering algorithm, including hierarchical clustering, density-based clustering, grid-based clustering, and division-based clustering. K-Means algorithm [10] is used in this paper.

(4) Sensitive events and clues extraction

Event discovery module uses centralized document corpus, identify where the event involved, and in accordance with the event groups documents. The module can recognize and extract the events related to the time and place, character elements from the document. Fig. 1 gives the module design class diagram. Event discovery module consists of nine categories, their respective functions and relationships are detailed below:

The NewsReport class is a wrapper class in the corpus of the document, which in addition to save the document text, but also the subsequent storage of intermediate results, such as the segmentation results is stored in the sequence container in class.

The TimeDateFile class is the date and time format configuration file management tools, it is the function of management.

The TimeDatePattern class is the date and time pattern, the pattern for time and date format recognition in the document.

The TimeDateMark class is classes of data, each instance of the class are recorded in the relevant information of an entity date and time.

The TimeDataAnalyser class is a time analyzer, generally only generate an instance. The TimeDataAnalyser class will match text generated over the course of time and date pattern in the NewsReport class, extract, date and time instances, and generate the TimeDateMark object. We will fill to the corresponding domain of the object information of the time and date of the instance, perfect object information, such as at the beginning of the text in the end position, time and date and time format, date literals etc. The TimeDateAnalyser class is to the intersection of adjacent and TimeDateMark combined treatment, using the greedy strategy. The TimeDateAnalyser class is taken by the relationship between the reference times, every time the date instance into its absolute time representation.

The Entity object is used to store information about places and people, organizations and other entities.

Recognize Named Entity Analyzer is the base class, the implementation will be taken by Proxy mode related method named entity parser instance by calling the base class, named entities extracted from the text NewsReport object.

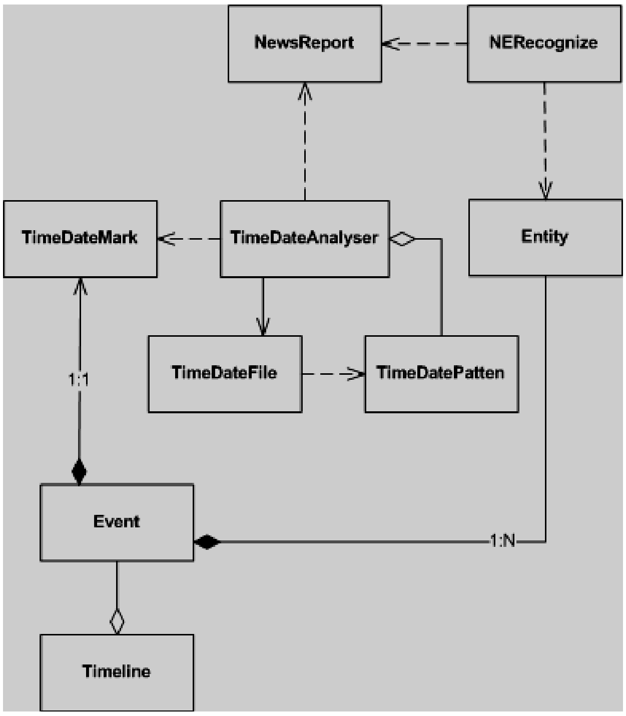

Figure 1. The class diagram of event discovery

Event class is the event of the data class. Experiments show that using the time expressions, namely the date and time as the center of the entity instance attachment type combination is most reasonable.

When caching, the event object is stored in the persistent media using the serialization mechanism of JAVA. Because the output produced by the module can be used as event detection results show to the user, but also can be used as the intermediate results of further analysis, so the module will save the output to the database event information table.

The event identification and evolution analysis module has two main functions: 1) in the event that extract module events in the real world is a projection of the events occurring in the space in the text, called reference for event instances or events. It is possible that an event instance that appears in multiple documents refers to the same objective event. One of the functions of this module is to identify the common finger of these event instances. 2) After obtaining enough objective events, this module can analyze the relationship between the events. Fig. 2 gives the module design class diagram.

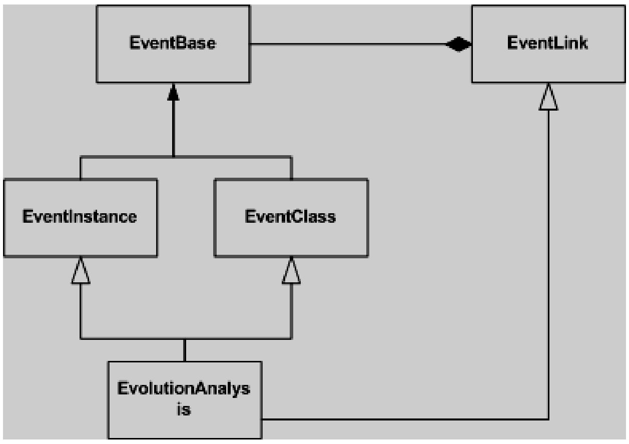

Figure 2. The class diagram of the event identification and evolution

The EventBase class is the event base class, which maintains the common characteristics of the event, such as the occurrence of the event, the entity, the context of the event, and so on.

EventInstance is an instance of the event, which inherits from the EventBase class. This kind of data is the result of the analysis in the time discovery module, which is obtained from the database.

EventClass is an objective event class, which is also inherited from EventBase. An EventClass object may be a superposition of data in one or several EventInstance objects.

Inter event relations is EventLink. For that event instances, there will be an EventLink object pointing two refers to a total of event instances, and in the object that two is reference; similarly, for two objective events before and after the relationship will have corresponding to objects in the EventLink that who is who's who is who of.

Module drive is EvolutionAnalysis, this class generally only need one instance, it first from the database read events module that generates results are found, and generate EventInstance object, encapsulation of the data; and detection EventInstance object is refers to the situation, on the formation of CO reference event generation eventclass types of packages, and eventclass reference pointer to an instance of an object in the original event, necessary to create EventLink class identifier that; finally, the eventclass set in each of the two object detection, judge whether there is a before and after the reference to the relationship between, such as the presence of is created EventLink class using the connection.

In addition, the interface of the module is as follows: the input and output of the module is completed in the event information table in the database. As an off-line analysis module, this module is called the analysis process, which is not directly coupled with other modules.

(5) Community discovery

Community discovery module found discuss specific event or topic and the relationship between the members and target topic list of attributes, analysis the topic of network communication between the members of communication relationship and communication content, to generate topic community structure $G(T,V,E)$ and community evolution. Fig. 3 gives the module design class diagram.

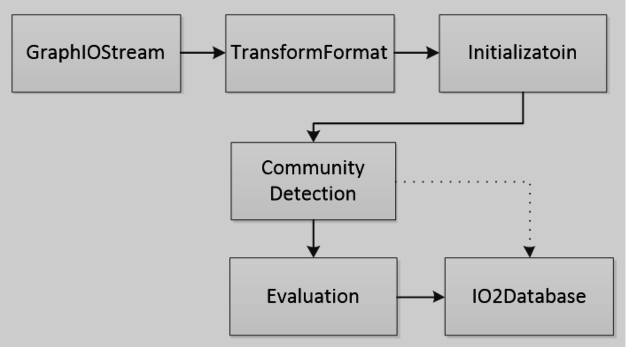

Figure 3. The class diagram of the community discovery.

GraphIOStream is responsible for reading the characteristics of the topic of the document and the author.

TransformFormat is responsible for the original data into the data format processing algorithm

Initializatoin is responsible for the initialization of complex algorithm parameters and the assignment of input.

CommunityDetection is the core algorithm for community discovery, is responsible for finding communities.

Evaluation is to evaluate the results and save the effective results.

IO2Database will be the results into the database for subsequent analysis and interface display.

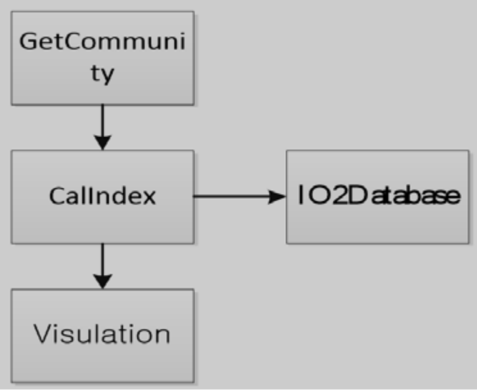

Figure 4. The class diagram of the display

Community indicators display module is to calculate the community related indicators, graphical interface based on the reality of community relations and the various indicators. Fig. 4 gives the module design class diagram.

Interactive GetCommunity interface is to analyze the community to eradicate.

CallIndex is a calculation of each indicator.

IO2Database save the results to the database.

Visualization will be displayed on the interface.

(6) Opinion leaders' discovery

The influence analysis module measures the influence of each user based on the multi relational network and user attributes.

The interface module: module directly from the database access to document the latest data as input, and deal with the historical data in the cache node module. Due to the generation module output can be as opinion leaders found the results displayed to the user, but also can be used as intermediate results for further analysis and processing, so the module can output saved to the database, the influential individuals in the table.

Opinion leader mining module can analyze the influential Internet users to a particular public opinion topic, which makes quick access to public opinion analyst opinion leaders; opinion leaders divert public opinion according to the instant.

VI. NETWORK PUBLIC OPINION ANALYSIS TECHNOLOGY

The system is connected by a Gigabit Ethernet switch on a Master (ubmaster) and 3 Slaves (ubslavel, ubslave2, ubslave3). IP were 192.168.102.230 ~ 192.168.102.233. Master runs on NameNode and JobTracker.

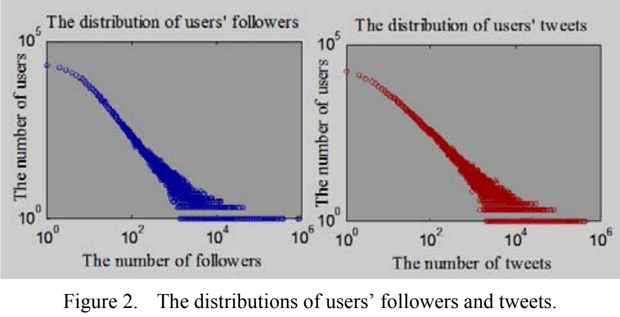

Figure 2. The distributions of users' followers and tweets.

At the start of Hadoop, HBase, Hive before the need to configure the Hadoop environment configuration file Hadoop.env.sh, core-site.xml, mapred.site.xml, hdfs-site.xml, hbase.env,sh, regionservers, hbase-site.xml, hive-env.sh, hive-site.xml. HBase database stores the collected information, micro-blog blogger content and focus information.

In the system implementation phase, we collected a total of nearly one million personal information and 220,000 micro-blog content. The user reads through the interaction analysis module in the library and the cluster center. Then we use the force directed algorithm and JavaScript to visualize.

The three most important characteristics of complex network are small world characteristics, scale-free characteristics and clustering coefficient. The system uses the node degree distribution and clustering coefficient to analyze microblogging network scale-free and small world. We calculate MapReduce of bloggers information attention and fans relationship. The experimental result is shown in Fig. 2. The distributions of users' followers and tweets are approximately power-law, implying that the vast majority of users have a small quantity of followers and tweets, while a small fraction have a large number of followers and tweets.

VII. CONCLUSION

With the rapid development of the Internet, every day will produce huge network data, how fast and effective analysis and processing of data rather than the data become a disaster, is in large data environment is an urgent need to address the problem, in the massive network data quickly obtain accurate geographical situation information is the hotspot and focus of current research. In this paper, the basic theory and large data processing technology are combined with the hot spot of network public opinion. This paper puts forward the solution of distributed public opinion analysis under the environment of large data.

REFERENCES

[1] Christopher P. Borick and Barry Rabe. Belief in Global Warming on the Rebound: National Survey of American Public Opinion on Climate Change. SERIES: Issues in Governance Studies, vol. 2, No. 08, pp. 14-15, 2012.

[2] BI Page, RY Shapiro. Effects of public opinion on policy. American political science review, Cambridge Univ Press, vol. 5, No. 07, pp. 45-56, 1983.

[3] Michael K,Miller K W.Big Data :New Opportunities and New Challenges. Computer, vol. 46, No. 06, pp. 22-24, 2013.

[4] Maclennan B,Kypri K,Langley J, et al.Public Sentiment Towards Alcohol and Local Government Alcohol policies in New Zealand.International Journal of Drug Policy, vol. 23, No. 01, pp. 45-53, 2014.

[5] Alan B.Krueger and Jitka Malecková.Attitudes and Action:Public Opinion and the Occurrence of International Terror ism. Science, , vol. 325, No. 5947, pp. 1534-1536, 2009.

[6] Shvachko K,Kuang H,Radia S,et al.The Hadoop Distributed File System . Mass Storage System and Technologies(MSST),2010 IEEE 26th Symposium on.IEEE, 2010: 1-10.

[7] White T Hadoop:The Definitive Guide:O'Reilly Media 2009.

[8] Dean J;Ghemawat S MapReduce:Simplified data processing on large clusters 2004.

[9] George L HBase:The Definitive Guide:O'Reilly Media 2011.

[10] Kanungo T. et al. An efficient K-means clustering algorithm: analysis and implementation. IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 24, No. 7, pp. 881-892, 2002.