Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2370745, 15 pages
https://doi.org/10.1155/2022/2370745

Research Article

The System of the Dissemination Characteristics of Internet Public Opinion Big Data Based on Artificial Intelligence

Xiaobo Wu[1] and Sitong Liu[2,3]

[1]*School of Business, Lingnan Normal University, Zhanjiang 524048, Guangdong, China*
[2]*School of Journalism and Communication, Nanjing University, Nanjing 210023, Jiangsu, China*
[3]*School of Management, Guilin University of Aerospace Technology, Guilin 541004, Guangxi, China*

Correspondence should be addressed to Sitong Liu; liusitong@guat.edu.cn

Received 21 May 2022; Revised 22 July 2022; Accepted 29 July 2022; Published 26 August 2022

Academic Editor: Jun Ye

In the era we live in today, the network is often used to analyze a large number of complex systems. With the development of the information society, there are more and more ways to disseminate public information through social networks. Public opinion dissemination refers to the process of disseminating public opinion information through social networks. Because the dissemination of public opinion is the basis for the exchange of ideas among multiple communicators of public opinion, the network community will certainly have an impact on the dissemination and development of public opinion. This article is based on artificial intelligence to study the network public opinion big data dissemination characteristic analysis system, introduces the network public opinion analysis system based on the characteristics of the network public opinion, introduces in detail multiple methods and clustering algorithms for extracting the text information of Internet public opinion, and proposes the Kmeans + Canopy + semantic similarity algorithm, and uses the A event to compare the parameters of the network clustering coefficient, the correlation measure and the degree centrality measure, and the performance of the Kmeans + Canopy algorithm and the Kmeans + Canopy + semantic similarity algorithm. The results of the experiment found that the clustering coefficient of "People's Daily" in the network dissemination of A event was 0.038, which was the highest among all nodes. It shows that 3.8% of the nodes established by the "People's Daily" can interact one-to-one to deliver information and intelligence resources. Although the complexity of the algorithm has increased and the time consumed by the system has increased, the accuracy of clustering has been improved, especially for cultural articles, the accuracy rate has been as high as 75%, and entertainment articles can reach up to 70%, and stabilize at around 70%.

1. Introduction

1.1. Background. Public opinion can always be defined as the attitudes and ideas of a person or group that touches and connects the emotions and ideas of social groups to each other in a fixed time and angle for many public events. From a sociological point of view, citizenship itself is a full response to the will of the people. The complex network we want to study, as a network close to real life, plays an important role in the research of public opinion communication, and has attracted the attention of various fields. The traditional big data dissemination of public opinion on the Internet has always been in need of improvement in terms of expansion and accuracy, and the research results in the combination with artificial intelligence are not ideal. Artificial intelligence realizes intelligent resource management for wireless communication through powerful learning and automatic adaptation capabilities. On the other hand, in wireless communication resource management, the use of artificial intelligence requires new network architecture and system models and standardized interfaces/protocols/data formats to promote the large-scale deployment of artificial intelligence in future B5G/6G networks.

1.2. Significance. With the advent of the Internet, more and more netizens pay attention to public opinion events through online media. With the spread of comments by netizens on emergencies, negative public opinion topics and confrontations related to these incidents are also widely spread on the "self-media" network. Gradually, this phenomenon has aroused the attention of academia as well as the attention of all sectors of society. People search all kinds of hot news on the Internet to express their thoughts. However, due to the openness and authenticity of the Internet, the spread of negative events can even affect social stability. Due to the large number of public opinion events that occur at the same time, and the public opinion events also contain a large amount of information, management departments must quickly and regularly guide the fermentation of public opinion events on the Internet, which is of great significance for determining the emotional intensity of Internet users. Public events are hot topics of social concern, which can easily arouse enthusiastic discussions among viewers and netizens, thereby forming general public opinion on the Internet. Public opinion on public events has a major impact on the development and maintenance of public events, and the wrong direction will affect the implementation of public events.

1.3. Related Work. The "self-media" network is real time and open. This kind of network lacks a gatekeeping system. Starting from the characteristics of the "self-media" network public opinion elements, Wang G constructed a multidimensional network model oriented to the "self-media" network public opinion topology. Based on the real process of generating and disseminating multiple topics of the same event, he designed a topic detection algorithm suitable for these multidimensional public opinion networks. The research results are mainly summarized from the following three aspects: (1) The multidimensional network model can effectively describe the communication characteristics of multiple topics on the "self-media" network. (2) Using a multidimensional topic detection algorithm, 70% of public opinion topics related to case study events have been effectively detected, which shows that the algorithm is effective in detecting topics in the information flow on the "self-media" network. (3) By defining the psychological scores of single and paired Chinese keywords in public opinion information, using topic detection algorithms to determine the emotional tendency of each topic, revealing the negative topics discussed on the "self-media" network, but the research results have not yet been obtained wide range of applications [1]. Metadata is the information about the organization of data, data domains, and their relationships. Traditionally, information science has been the center of metadata research. Mayernik MS introduced 5 key sociotechnical characteristics of metadata in digital networks. He said that the nature and importance of metadata in network communication systems are rarely discussed in the information science community. Without such metadata, network communications cannot exist. If we want to have meaningful discussions about our digital traces, or make wise decisions about new policies and technologies, then we must develop theoretical and empirical frameworks for explaining digital metadata. But the practicality is not strong [2]. As Internet users use tags when posting and searching for information on social media, it is important to understand who built the tag network and how the information is spread across the network. Kim J unleashes the potential of the AlphaGo tag network by solving the following problems. His current research examines whether traditional opinion leadership (that is, the influence hypothesis) or the grassroots participation of the public (that is, the interpersonal relationship hypothesis) promotes the dissemination of information in the tag network. Finally, it tested the correlation between the attributes of key users (i.e., the number of followers and followers) that have a significant impact on the distribution of information and their central position in the network. The results show that the main participants in the network actively receive information from their followers, rather than acting as an intermediary between participants, but the experimental process is too complicated [3]. Cyberspace is reshaping the way companies manage sales and marketing assets. Gang LI's research in these network applications and services found that some public relations companies hire staff to post product reviews on different online communities and social networks, and the hired has not even consumed these services or products. Although online paid posters can be used as an effective e-marketing strategy, they can also conduct malicious behavior by spreading rumors or negative information about competitors. More specifically, a set of paid posters can cooperate with good marketing to produce the expected results of positive or negative opinions to attract attention or stimulate curiosity. However, how to use social media to sell better still remains to be considered [4]. Modern malware uses advanced techniques to hide static and dynamic analysis tools. Therefore, investigating how to use general indicators (such as the energy consumed by the device) to reveal the presence of malware is very important. From this perspective, Caviglione L uses two detection methods based on artificial intelligence tools, such as neural networks and decision trees, to find malware that hides the exchange of data. In order to verify its effectiveness, seven covert channels have been implemented and tested on the measurement framework using Android devices. Experimental results show that this method can effectively detect hidden data exchanges between applications, but it requires a lot of initial investment [5]. How to fully tap the potential of artificial intelligence (AI) technologies in future wireless communications, such as beyond 5G (B5G) and 6G, is a very popular interdisciplinary research topic worldwide. Lin M reviewed the latest developments in resource management authorized by artificial intelligence from a framework perspective to a methodology perspective, not only considering the management of radio resources (such as spectrum), but also other types of resources, such as computing and caching. He also discussed the challenges and opportunities faced by AI-based resource management in the widespread deployment of AI in future wireless communications, but further research is needed to achieve intelligent resource management [6]. However, because of the relative virtual nature of the network, it is of great practical significance to quickly and effectively dig out important information from the massive network information and to understand the public opinion dynamics of the people and control them in time.

1.4. Innovation. This paper studies the analysis system based on artificial intelligence on the dissemination characteristics of network public opinion big data. Aiming at the characteristics of Internet public opinion, this article introduces in detail multiple methods and clustering algorithms for extracting text information of Internet public opinion. In order to improve the clustering convergence and the accuracy of network information detection, this paper considers that the results of calculating similarity between Chinese words are still not ideal. Because there are many synonyms in Chinese words, this paper proposes a method of combining Kmeans + Canopy and semantic similarity algorithm, and compares it with Kmeans + Canopy, and the effect is very good.

2. Method of Network Public Opinion Big Data Dissemination Characteristic Analysis System Based on Artificial Intelligence

Following the three major media, newspapers, radio, and television, the Internet has been given an important mission as the "fourth media" and has penetrated into all aspects of human life. Through social networks, people can check some popular news and social events that interest them and express their opinions. In recent years, due to the development of the Internet era, the number of mobile Internet users has increased significantly, as shown in Figure 1 [7]. As one of the most convenient Internet products, mobile phones have become an indispensable tool for Internet users. At the same time, because the Internet is so open and free, more and more netizens express their attitudes, ideas, opinions, and problems quickly, directly, and honestly through the Internet. The Internet has become an important platform for the development and dissemination of public opinion [8].

Public opinion refers to an individual's expression of social reality events in a certain period of time and is an objective expression of group attitudes, thoughts, feelings, and problems. Public opinion is exposed and spread through the Internet. To a certain extent, people can use these media to express their reactions and emotions to events and express their thoughts, and even personal remarks can easily trigger the fermentation of certain events [9]. The way to start forwarding without understanding the comprehensiveness of the event also belongs to the category of online public opinion, which will have a negative impact on economic development and social stability. Therefore, timely detection of online public opinion, and appropriate response measures based on its development and changes, and effective monitoring of it have become important tasks for government departments.

The social network public opinion communication process is divided into 4 stages, as shown in Figure 2 [10].

Because most of the Internet public opinion information appears in Chinese texts, the relevant processing of Chinese Internet public opinion texts has become a research focus. The following mainly introduces various knowledge of processing public opinion text, such as Chinese preprocessing, feature extraction and representation, text similarity calculation, and text clustering knowledge [11]. Figure 3 is the network public opinion analysis system.

The network-based public opinion analysis system is a web-based application system. For big data processing technology, Hadoop is a distributed system infrastructure developed by the Apache Foundation. Users can develop distributed programs without understanding the underlying details of the distribution and make full use of the power of the cluster for high-speed computing and storage. Hadoop's MapReduce parallel computing framework is a computing model running on HDFS distributed storage system. The main purpose of this design is to achieve sharing and overcome the problems of the overall system, which can greatly reduce the initial communication consumption during data transmission [12]. In the processing of network public opinion data, the subsystem preprocesses the collection of public data for public opinion analysis and analysis and prediction of the basic system. The data collected by the collection system is mainly stored locally in the form of webpage text and cannot be used directly [13]. A subsystem that pregenerates each automatically encoded information and recognizes the received web page text, then removes network noise, extracts text information, and provides Chinese word segmentation to the public opinion analysis and prediction subsystem for further processing.

The public opinion analysis and prediction subsystem is an important part of the entire public opinion analysis and planning framework and is used to perform the key functions of the system. The public opinion prediction and analysis system analyzes the public network information previously processed through public opinion data. The main content usually includes the following aspects:

(1) Subject search and tracking, from general network information such as news and blogs on the Internet, through data mining, automatically search and retrieve the public opinion topics contained therein, and identify each follow-up report of the subject, to achieve the completeness of tracking the birth, development, and demise of the theme [14].

(2) Hot topic recognition, for most of the titles found, according to their news source sequence, visit volume, number of comments, number growth rate, and comment time intensity, to identify the headline information on the Internet hotspot

(3) Sensitive topic recognition, based on detailed analysis of various sensitive words found on the Internet, using keywords to manage, identify "suspicious" topics, and generate relevant warnings

The public network database is essentially composed of a network public opinion information database, a network public opinion analysis database, a system management database, a sensitive word database, and a user database. The specific situation is shown in Figure 4 [15]. According to the characteristics of the quantity, size, growth rate, and type of data stored in different databases, different data

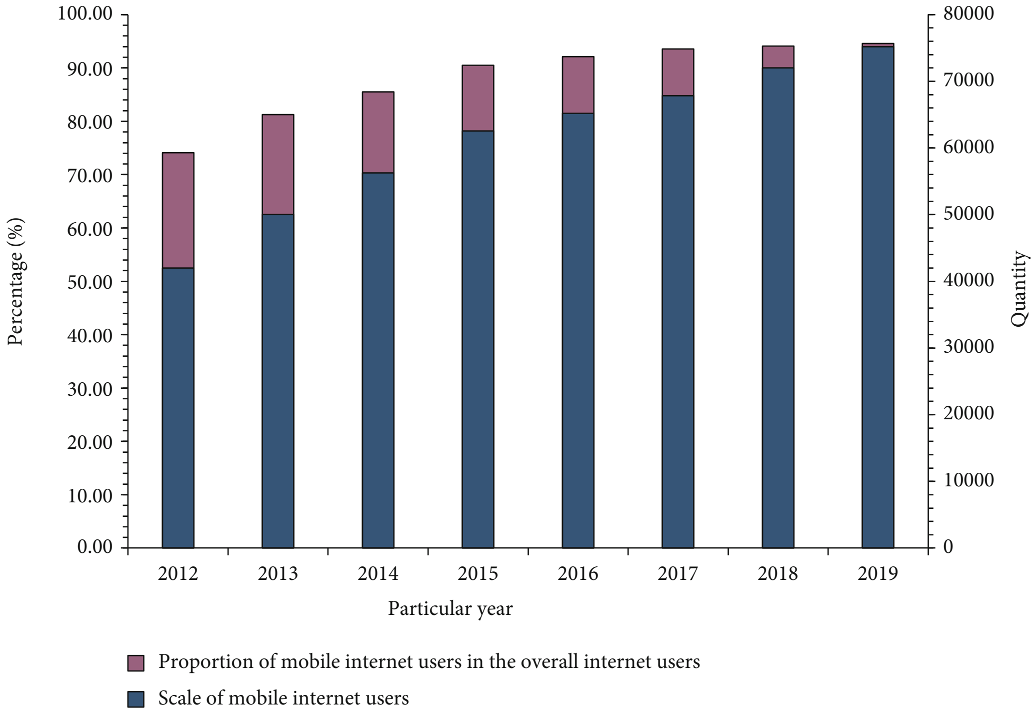

Figure 1: Scale of Internet users and Internet penetration in China.

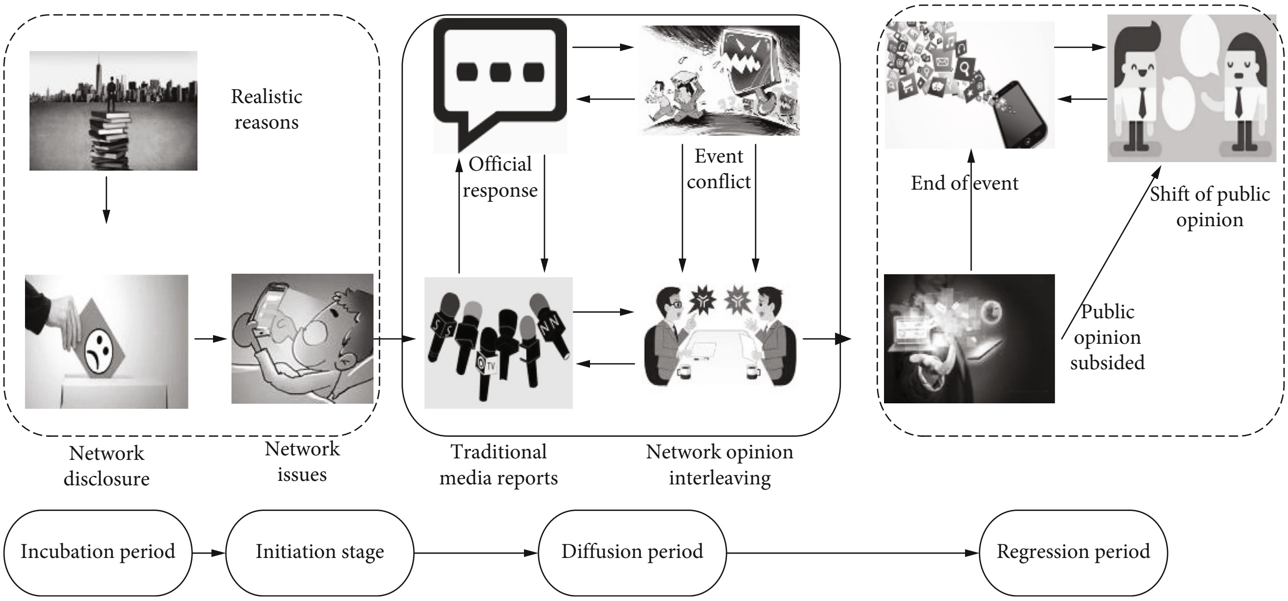

Figure 2: Network public opinion dissemination model diagram of social events.

systems are used to determine the database and control. Network public opinion data storage and management are collected through a network data collection system, which provides preconditions for the processing of network public opinion data [16].

2.1. Preprocessing of Online Public Opinion Information Text. With the advent of the big data era, the data on the Internet is showing a trend of doubling. How to efficiently obtain and use the content of interest on the Internet is an important direction of value-added in the current data mining field. For this purpose, web crawlers have ushered in a new wave of revitalization and have become a hot technology that has developed rapidly in recent years. Web crawler is a program that automatically extracts web pages according to certain rules. It can automatically crawl web pages on the Internet through the network [17]. Figure 5 shows the design form of the web crawler. The various text information on the webpages obtained by the web crawler contain a large amount of unorganized original webpage data, and there are

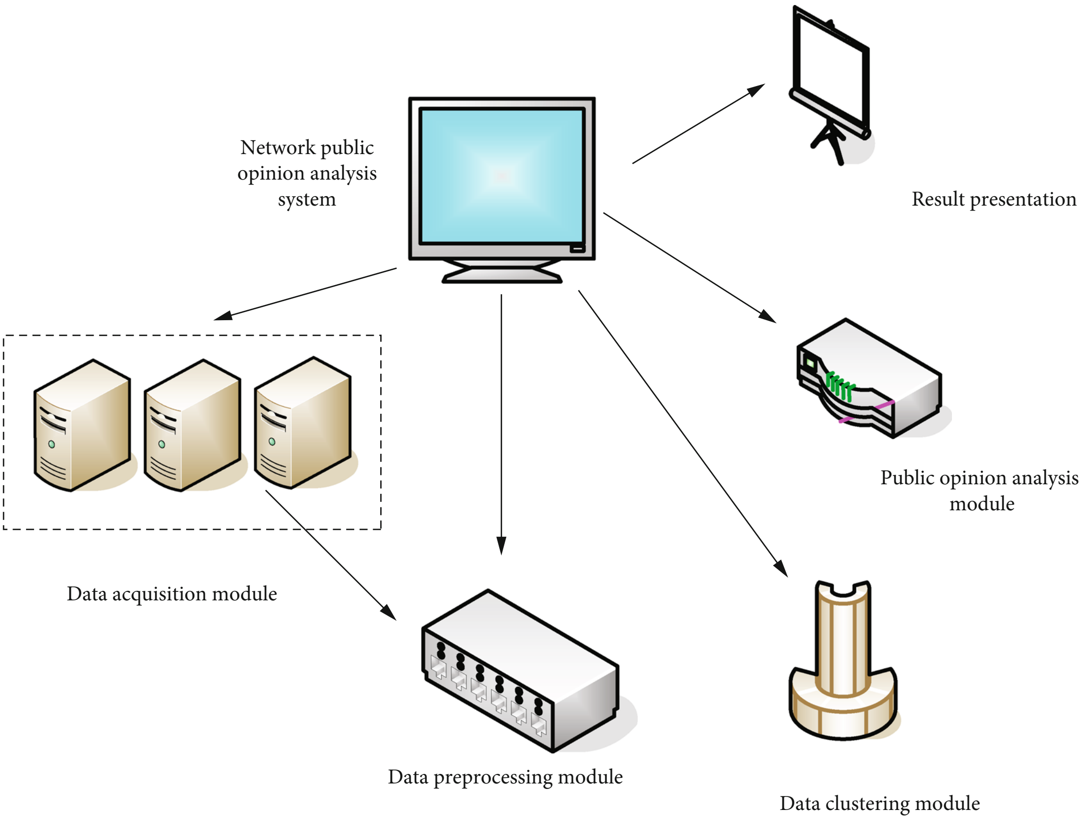
Network public opinion analysis system
Result presentation
Public opinion analysis module
Data acquisition module
Data preprocessing module
Data clustering module

Figure 3: The composition of the network public opinion analysis system.

many repetitive downloading webpages and information with little relevance. Therefore, it is very important to preprocess the text of the network public opinion information.

Text information is composed of many words, punctuation marks, spaces, and important characters. It belongs to the unstructured data. Word segmentation is the first step of preprocessing, the process of combining rules into words [18]. English word segmentation is relatively concise, based on basic unit words, with obvious divisions, which is convenient for computer processing. In our understanding, Chinese word segmentation depends on the word as a meaningful language. It is more complicated and difficult due to the simplicity of the combination of words and the diversity of meanings. But in order to extract relevant keywords from the text, word segmentation is a very important step in generating information. With the development of Chinese information processing, related technologies have been significantly improved, including Chinese word segmentation technology, and many word segmentation algorithms have appeared [19]. Existing word segmentation algorithms include string-based word segmentation, understanding-based word segmentation, statistics-based word segmentation, and semantic-based word segmentation. This article will perform word segmentation processing on the network public opinion information text. The system includes functions such as Chinese word segmentation, part-of-speech tagging, and new word recognition and supports user dictionaries. After the word segmentation process, only part of the text is planned, since there is no unified standard for the Chinese stop word list; in this case, a corresponding stop word list is established according to the feature set. As some modal auxiliary words, conjunctions, and prepositions have no practical meaning and have little effect on the vocabulary, select an effective set of feature items, thereby reducing the feature space dimension, which can save storage space and calculation time, which can improve performance and accuracy [20].

2.2. Text Feature Extraction

2.2.1. Mutual Information Method. Mutual information is mainly used to measure text features and categories. The main idea is as follows: If a word has a higher frequency in a certain category and is higher than other categories, then the amount of common information between the word and the corresponding category is greater [21, 22]. The mutual information method can improve the efficiency of classifier training and application by reducing the effective vocabulary space. The operation process is relatively simple, which is convenient for everyone to understand and use. Mutual information can be used as a measure to measure the correlation between text feature w and category l, which can be defined as $\mathrm{OR}(w, l)$, and the formula is shown in

$$\mathrm{OR}(w, l) = \log p(w, l)/p(w)p(l), \tag{1}$$

where $p(w)$ represents the proportion of the number of texts containing the feature w to the total number of texts, $p(l)$

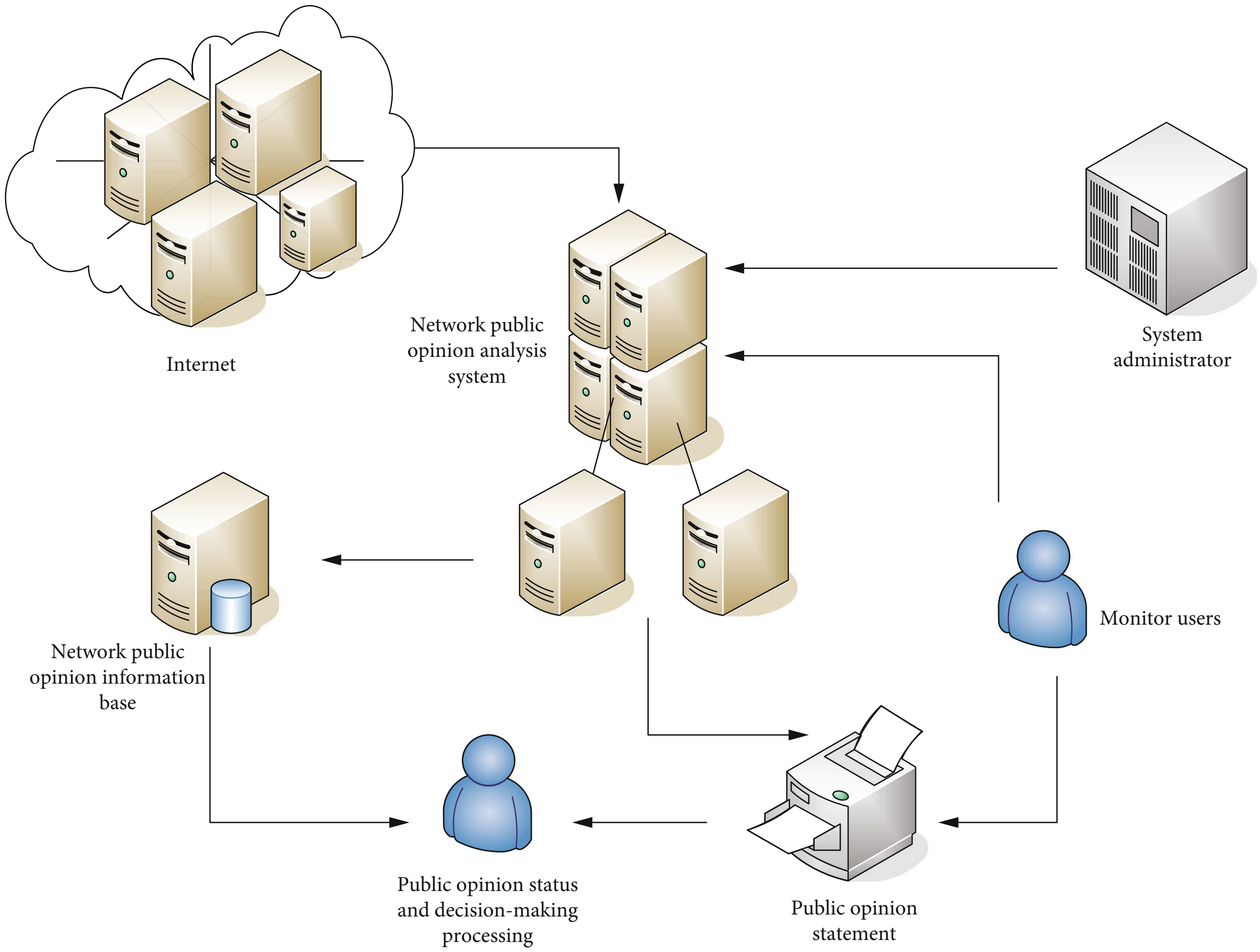
Internet
Network public opinion analysis system
System administrator
Monitor users
Network public opinion information base
Public opinion status and decision-making processing
Public opinion statement

Figure 4: Business flow chart of network public opinion analysis system.

represents the proportion of the number of texts contained in the category l to the total number of texts, and $p(w, l)$ represents the probability of the occurrence of the feature w in the texts contained in the category l. It can be understood that if the probability of feature w appearing in a certain category l is significantly higher than that of other categories, it means that the probability of using feature w as a feature item of category l will be greater. Set the set of all classes of text to s; then, the average value of mutual information $\mathrm{OR}(w)$ can be defined as:

$$\mathrm{OR}(w) = \sum_{i=1}^{s} p(l_i)\mathrm{OR}(w, l_i). \quad (2)$$

If the mutual information value is equal to 0, it means that the feature w does not belong to type l at all. Similarly, when the mutual information value reaches the maximum, it means that the feature w belongs to type l only. However, in the mutual information, the frequency of the feature w is not considered, so it is likely to ignore the high-frequency vocabulary with obvious characteristics, and instead select the low-frequency vocabulary as the best feature item of the text. Therefore, the extracted feature information can be guaranteed by increasing the dimensionality of the feature space, but this will increase the calculation time and space utilization.

2.2.2. χ^2 Statistical Method. χ^2 Statistics can also be used to measure the correlation between features and categories. The formula of χ^2 statistical method between feature w and category l is:

$$M = \left[p(w, l)p(\bar{w}, l) - p(\bar{w}, l)p\left(w, \bar{l}\right)\right]^2, \quad (3)$$

$$\chi^2(w, l) = \frac{\mathrm{NM}}{p(w)p(\bar{w})p(l)p\left(\bar{l}\right)}, \quad (4)$$

In the above formula, when the value of $\chi^2(w, l)$ is smaller, it means that the correlation between feature w and category l is greater, and the distinguishing ability of feature w in category l is more obvious. Therefore, we must pay attention to the size of the value. Usually, the average statistics and maximum statistics of χ^2 statistics can also be calculated. If the total number of text categories is s, it can be defined as:

$$\chi^2_{\mathrm{avg}}(w) = \sum_{i=1}^{s} p(l_i)\chi^2(w, l_i), \quad (5)$$

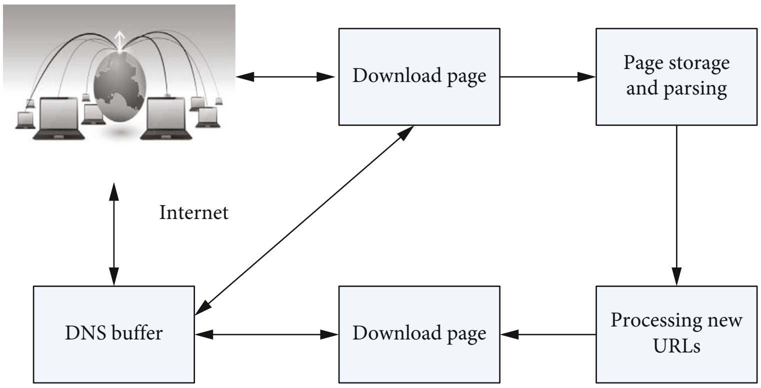

FIGURE 5: The design form of the web crawler.

$$\chi^2_{\max}(w) = \max \left\{\chi^2(w, l_i)\right\}. \tag{6}$$

The accuracy of χ^2 statistics in feature selection is very high, and the sensitivity to the training set is relatively low, so it is relatively stable.

2.2.3. Information Gain Method. The metric of information gain is the amount of information that a feature item carries in the clustering process. The greater the amount of information, the more important the feature is. Information gain is a statistical method based on entropy, and this entropy is the amount of information the feature carries in the clustering system [23].

Assuming that the feature is w and $\bar{w}$ represents the feature other than w, the information gain formula for the feature and category l is:

$$\begin{aligned} I(w) = {} & \sum_{i=1}^{s} p(l_i) \log \left(1/p(l_i)\right) - p(w) \sum_{i=1}^{s} p(l_i \wedge w) \log \left(1/p(l_i \wedge w)\right) \\ & - p(\bar{w}) \sum_{i=1}^{s} p(l_i \wedge \bar{w}) \log \left(1/p(l_i \wedge \bar{w})\right), \end{aligned} \tag{7}$$

where $p(l_i)$ represents the proportion of the number of texts contained in category l to the total number of texts, $p(w)$ represents the proportion of the number of texts containing feature w to the total number of texts, and $p(l_i \wedge w)$ represents the probability of feature w belonging to category l.

Information gain only considers the effect of features on the entire classification,

$$\mathrm{IDF}_i = \log \frac{S}{s_i}. \tag{8}$$

Formula (8) does not consider the relationship between features and categories, so it is very restrictive, which will greatly reduce the efficiency of feature extraction.

2.2.4. Term Frequency-Inverse Document Frequency (TF-IDF). When the importance of a word or phrase in a document needs to be determined to extract the features of the text, the TF-IDF method is used to calculate the vector density of the feature to establish a field vector model. The TF-IDF method is the most commonly used method in data update. The main idea of TF-IDF is: If a word or phrase appears in the article with a high frequency of TF, but rarely appears in other texts, it can be considered that the word or phrase has good distinguishing ability, and then, it is beneficial to classification. TF (word frequency) is used to indicate the frequency of occurrence of the word in the text of the document. The main idea of IDF (inverse document frequency) is as follows: If only a few documents contain the keyword, that is, the smaller s, the IDF is larger, indicating that the keyword is easy to distinguish. Therefore, the TF-IDF weight can be calculated to reflect the importance of keywords in a specific document.

In the TF-IDF method, the total number of documents is set to S, the keywords appear in each document, $n_{i,j}$ is the number of times the keyword appears in the document, and then, the frequency of the keyword appearing in the document is $\mathrm{TF}_{i,j}$. As shown in Equation (8), the denominator of this equation represents the sum of the number of occurrences of all words in the document. The calculation process of IDF_i of the keyword is expressed in Equation (8). Equation (9) expresses the importance of keywords in the document through a weight of $w_{i,j}$.

$$\mathrm{TF}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}, \tag{9}$$

$$\mathrm{IDF}_i = \log \frac{S}{s_i}, \tag{10}$$

$$w_{i,j} = \mathrm{TF}_i \times \mathrm{IDF}_i. \tag{11}$$

2.3. Text Representation. After extracting text feature items, in order to facilitate quantitative expression of the relationship between the texts, the text can be expressed as a text vector composed of feature items to describe and replace the text. The commonly used text representation models are as follows:

2.3.1. Vector Space Model (VSM). The concept of the vector space model is easy to understand. It is used to express the meaning of the document in text, and only needs to represent the feature item representing the document and the frequency of its appearance, without considering other things such as position, order, and factor.

The main idea of using the space vector model to represent text is to use a vector in the vector space to represent a single text, and the dimension of the vector space is represented by the feature items representing the corresponding text. And the dimension value corresponds to the weight of the feature vocabulary, which can be calculated by the TF-IDF method.

Assuming that in the text set A, there are m feature items t for each text a and the feature items are independent of each other, then:

$$a = (e_1, e_2, \cdots, e_m). \tag{12}$$

Assuming w_i is the weight of feature e_i in text a, with $(e_1, e_2, \cdots, e_m)$ as the coordinate axis of the m-dimensional space and $(e_1, e_2, \cdots, e_m)$ as its corresponding coordinate value, then:

$$a = (t_1 w_1, t_2 w_2, \cdots, t_m w_m). \tag{13}$$

The weight of the feature item is used to construct the feature vector of text a, and then, the feature vector $\bar{V}_d$ of text a can be expressed as:

$$\bar{V}_d = (w_1, w_2, \cdots, w_m). \tag{14}$$

Therefore, the text set A can be formally expressed as:

$$A = \begin{bmatrix} \vec{V}_{d_1} \\ \vec{V}_{d_2} \\ \cdots \\ \vec{V}_{d_n} \end{bmatrix} = (e_1, e_2, \cdots, e_m) \times \begin{pmatrix} w_{11} & \cdots & w_{1m} \\ \cdots & \cdots & \cdots \\ w_{n1} & \cdots & w_{\mathrm{nm}} \end{pmatrix}. \tag{15}$$

It can be seen from the above that the vector field model is very useful in the field of information retrieval. This method of word processing reduces the complexity of the problem and improves the performance of statistical data. However, the vector field model also has certain limitations. If the words are independent of each other, it is difficult to achieve literal association. The weight of the structured text is read according to TF-IDF, and a certain amount of data in the storage library is required to reflect the calculation effect; the text vector produced by the field vector model is high and sparse, which will increase the storage capacity [24, 25].

2.3.2. Probability Model. The probability model uses the probability relationship between distinguishing words and distinguishing words and between distinguishing words and text to retrieve information. The first is to divide the content in the text into related text information, and the rest are less important content. Then, by selecting the probability value corresponding to the distinguishing word, the probability of the distinguishing word appearing in the two types of text is displayed, and then, the correlation between the terms is calculated. The probability model is largely based on theory. If it has a different understanding of correlation, there will be different probability models and different ranking results. The basis of the probability recovery model is the basis of relevance, and the technical problem of the probability recovery model is the data source and probability calculation. Because of this, there will be unpredictable situations and corresponding evaluation criteria for each probability model, which is not conducive to specific use and requires a large-scale calculation.

2.3.3. Boolean Model. Boolean model is a simple information retrieval model based on Boolean algebra theory. In the standard Bolianian model, the text information is expressed as:

$$d = (w_{i1}, w_{i1}, \cdots, w_{\mathrm{in}}). \tag{16}$$

The model also represents a single text as a vector, and the weight w_{ik} of the feature item in the vector is represented by 0 or 1, respectively, indicating whether the feature word appears in the text. It can be seen from this that the Boolean model is a special form of the above-mentioned space vector model.

In the Boolean model, it is assumed that the relationship between the feature words and the text only includes appearance and non-appearance, so the Boolean model vector is composed of all the different feature words in the text. That is, the dimensionality of the vector is determined by the number of different feature words, and the corresponding weight value is 1 or 0.

The structure of the Boolean model is simple, easy to understand, and highly maneuverable. However, there are many limitations. For example, the matching conditions are too harsh, which will affect the accuracy of text retrieval; the Boolean model can only be used in information retrieval to calculate the relevance of a query and a document, but it cannot calculate the similarity between documents in depth.

2.3.4. Clustering Algorithm. For processing large data sets, clustering algorithms are relatively scalable and efficient, which will be of great help to the processing of online public opinion text information. Text clustering is the core step of the public opinion analysis system, and the selection of the clustering algorithm directly affects the result of topic detection. The function of text clustering is to divide the mixed text collection into several categories. The text similarity in each category is very high, but the text similarity between the categories is very low. The main clustering algorithms include Kmeans clustering algorithm and Canopy algorithm.

Kmeans algorithm: randomly select k data points as the initial value of each cluster center; remove the initial center point; calculate the distance from all other points to the center of each cluster; and then, classify the data point closest to the center point into the cluster; according to the new cluster, find the point with the smallest distance from other points in the cluster as the new cluster center point; repeat the above steps until convergence. The Kmeans algorithm is widely used in real-life clustering applications due to its simple and effective analysis method. The specific process is shown in Figure 6.

Canopy clustering algorithm: The parallelization of Canopy clustering algorithm is similar to Kmeans clustering algorithm. First, it needs to set two thresholds T1 and T2, T1 > T2; one of the clusters is called the Canopy set, which is empty at the beginning; reading the data as a Canopy in the collection, then read the data, calculate the distance between it and all Canopy in the collection, if it is less than T1, divide the data into this Canopy; when the space is less than T2, the data cannot be considered as the center of other Canopy; otherwise, a new Canopy will be introduced.

Therefore, the Canopy algorithm is usually used in combination with the Kmeans algorithm, the original data is clustered through the Canopy algorithm, the parameter K and the first cluster center are determined, and then, the Kmeans algorithm is used for clustering to obtain the final result.

For the evaluation index of the effect of network public opinion text clustering, the most commonly used index is to use the accuracy rate and the recall rate when obtaining information and calculate the break-even point P to make a judgment. Assuming that R is the number of texts that are classified correctly, W is the number of texts that should be classified, and C is the number of texts that are actually classified in this category, and then, the accuracy and recall and P values are:

$$\text{Precision} = \frac{R}{C} \times 100\%, \tag{17}$$

$$\text{Recall} = \frac{R}{W} \times 100\%, \tag{18}$$

$$Q = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}, \tag{19}$$

$$P = \frac{2}{Q}. \tag{20}$$

2.3.5. Analysis of the Dissemination Characteristics of Online Public Opinion. In the process of spreading online public opinion, there are two ways of spreading, namely, linear spreading and decentralized spreading. Because of the differences between the media and the content of the communication, the mode of communication will have different effects. In real life, the information exchange between individuals belongs to linear communication, and even the exchange of opinions between small groups can also be classified as linear communication. Although it seems that the spread speed is not fast, it can also spread widely, and the growth effect of the exponential function is the same. At the same time, it can meet high requirements on the accuracy and completeness of information dissemination. Linear propagation can be regarded as a single-line connection between individuals, with high correlation and good information connectivity, so the scope is relatively small, information update is more convenient, and the next level of individuals is prevented from getting erroneous information. There is a big difference between decentralized communication and linear communication. Decentralized communication usually occurs in groups, with a wider spread and greater intensity. Therefore, attention should be paid to the correctness and function of the control information at the source of the dissemination. Because the feedback will take a long time, and the probability of feedback will be extremely low, therefore, there will generally be cases where relevant departments directly respond to the incident.

3. Analysis Experiment on the Dissemination Characteristics of Internet Public Opinion Big Data

3.1. Clustering Coefficient Measurement. The clustering coefficient represents the topological characteristics of the network and reflects the relative scale of the social network. The clustering coefficient represents the average size of all participants in the network, which is equal to the actual number of cores in the neighborhood/the maximum number of cores in the neighborhood. Therefore, its coefficient is always greater than 0 and less than 1, and its size indicates the degree of integration of the entire network. This article extracts a hot event A for network analysis.

3.2. Relevance Measure and Degree Centrality Measure. Relevance measurement: The measurement represents the relationship between members in the network and reflects the breadth of information dissemination. Common indicators are diameter and average path. The diameter refers to the shortest line distance between any two nodes in the network, that is, the shortest distance between all nodes. The larger the diameter, the more links the entire network passes through, the smaller the diameter, the fewer the links, and the faster the spread of information. The smaller the number of segments, the greater the breadth of information dissemination, and vice versa.

Degree centrality demonstrates the ability of an individual to interact and connect with other individuals in the network. The larger its value, the closer to the center of the network, the greater the weight, and the more it will affect others. The degree centrality is divided into degree centrality and degree centrality. The degree centrality is the number of connections to the nodes in the network. The larger its value, the greater the number of connections. The degree centrality indicates the degree of integration and overall continuity between the members of the network and represents the overall consistency of the network center. The larger its value, the more compact and consistent the network system.

3.3. Intermediate Centrality Measure and Close Centrality Measure. The measure of intermediate centrality represents the centrality of the individual in the whole of all networks. There are intermediate centrality and intermediate centrality. The middle centrality indicates the degree of control over resource information. In other words, if a node is on the shortest route of other parties as a “bridge,” individuals must use this “bridge” to obtain information and resources, and this route will master the sources of other components. With some control, other individuals can even be guided to

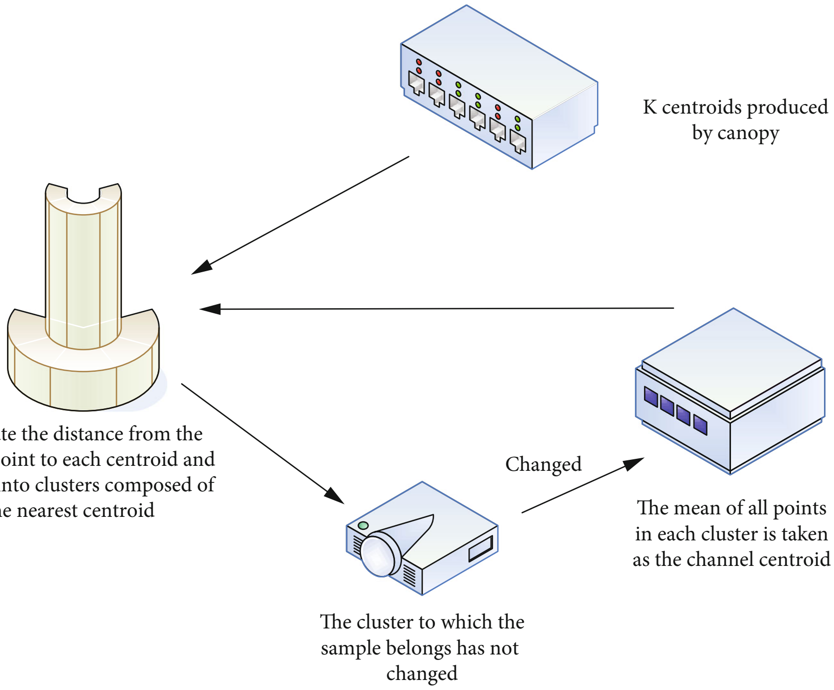
K centroids produced by canopy
Calculate the distance from the sample point to each centroid and divide it into clusters composed of the nearest centroid
Changed
The mean of all points in each cluster is taken as the channel centroid
The cluster to which the sample belongs has not changed

Figure 6: Flow chart of Kmeans algorithm.

obtain wrong information to achieve their goals. The intermediate central potential refers to the difference between the center of the highest node in the network and the centers of other nodes.

Proximity centrality measure: In social networks, neighboring nodes will distribute information sources more flexibly. It represents the degree to which one individual in the network is not restricted by other individuals. These include near central potential and near centrality. The proximity centrality is the sum of the shortest distances between an individual and all other individuals in the network. If it is farther from the center, it means that the restriction between this individual and other groups of individuals is higher. That is to say, the closer to the center, the deeper the core position in the network, the smaller the effect on information retrieval and retrieval sources, and the lower the weight.

3.4. Network Density Measurement. The general network depends on the structure of the entire network. According to the types of social network components and the diversity of their relationships, there are many general networks. The communication relationship between general networks can be considered as a "virtual interpersonal network" between individuals. The number of actors in this network is 100, which means there are 100 segments. The size of the network affects the relative structure of the network and the distance between nodes. As one of the key indicators of network structure, network density represents the approximate level of connections between individuals in social network relationships. The density is directly related to the relative density between nodes. If there are X connections and n nodes, there are $n(n-1)$ connecting lines, and the calculation formula for network density Y is:

$$Y = \frac{X}{n(n-1)}. \tag{21}$$

Generally speaking, the magnitude of the Y value and the relative density between the segments of the network are directly related to the impact on the individual. The larger the value of Y, the more useful it is in providing sources of information for individuals, but because the relationship between individuals is so close, the same individual may receive feedback from different individuals on the same information, which leads to the superposition of sources and information, increases the tedious degree of work, and is not conducive to the development of individuals in the network. In order to explain the closeness of the social network members in event A, the network density will be tested.

In the experiment, a cluster system consisting of 2 hosts was used. Each host is connected to a switch through a 100 M network card to form a local network and is connected to the Internet through the current stable Gigabit Ethernet card. Hadoop-2.6.0 was selected as the test platform, and designed a general network public opinion analysis system, and compared the performance of the Kmeans + Canopy algorithm and the Kmeans + Canopy + translation algorithm. Two of the hosts are configured as shown in Table 1.

4. The Experimental Analysis of the Dissemination Characteristics of Big Data of Online Public Opinion

Table 2 is the clustering coefficient measurement of the network public opinion of the A event.

Table 1: Configuration information of two hosts.

Project	Host 1	Project	Host 1
CPU	Intel Xeon (TM) 2.8 GHz	CPU	Intel Xeon (TM) 2.8 GHz
Memory	2 G	Memory	2 G
Hard disk	100 G	Hard disk	100 G
Host IP	192.168.1.211	Host IP	192.168.1.212

Table 2: Cluster coefficient measurement of event a network public opinion.

Node	Clus Coef	nPairs
Today's headlines	0.021	732
People's daily	0.038	602
Surging news	0.034	596
Tencent news	0.028	852
NetEase news	0.024	623

It can be seen from Table 2 that in the spread of this incident, the clustering coefficient of "People's Daily" is 0.038, which is the highest among the news software surveyed. It shows that 3.8% of the associated nodes have established contact with the "People's Daily" and can communicate with each other, provide information and understand the source. The nPairs size of "Tencent News" is 852, which is the largest among all nodes, indicating that "Tencent News" is the actor with the most network segments and connections. It can be seen from the results that the average size of the entire network is much higher. Some of them, such as "Tencent News" and "People's Daily," have greater relative proximity and are more likely to establish relationships with other parts.

The results of the relevance measurement results of the network public opinion dissemination of the A event and the degree centrality of the network public opinion dissemination data are shown in Table 3.

There are no unconnected nodes in the network public opinion dissemination of the A event; that is, all the nodes spreading in the network of this event can be directly connected or established by other participants. The average network path of "People's Daily" and "Netease News" is 4.9, and the power of information dissemination in public networks is stronger, access to the network is better, and information can be widely disseminated.

The degree centrality of "Tencent News" is 39, which is the highest among all nodes on the entire network. It reflects the great influence gained through media communication in the information network, and its high central position in the communication of online public opinion, which also confirms the role and appeal of active media and popular online actors.

Proximity to the center describes the closeness of participants in the network to other participants and refers to the information exchange capability of the entire network. The size of the proximity to the center is inversely proportional to the degree of information dependence between network participants. The results are shown in Table 4 for calculating the intermediate centrality of the public opinion propagation of the A event and the close centrality of the network public opinion.

Table 3: Correlation measurement results of network public opinion communication and degree centrality results of network public opinion communication data.

Node	Average path	Degree centrality
Today's headlines	5.2	36.00
People's daily	4.9	33.00
Surging news	5.1	32.00
Tencent news	5.0	39.00
NetEase news	4.9	37.00

Table 4: Calculation results of intermediate centrality of event a public opinion communication and proximity centrality of network public opinion.

Node	Intermediate centrality	Near centrality
Today's headlines	482.32	1462
People's daily	512.35	1528
Surging news	500.49	1546
Tencent news	493.26	1501
NetEase news	481.52	1623

It can be seen from the results that in event A, "People's Daily" and "The Paper" have a high degree of centrality. They have a higher degree of control over information sources and have a greater role as a "bridge." Obviously, their power and influence are enormous. In the network dissemination of event A, "Today's Toutiao" has a close relationship with other nodes, is highly independent, has a low degree of dependence on other parties, and is not easy to control. "Tencent News" and "Netease News" followed closely behind. These network segments are close to the core of the network and have advantages in receiving information and capacity.

For the measurement of "individual network density" in the general network, the public opinion data of event A in Ucinet is shown in Figure 7.

It can be seen from Figure 7 that "Tencent News" has 39 network node relationships, which is the largest in the network scale. This shows that 39 nodes have established an interactive relationship with it. There are 49 connections established with "Tencent News," and the average network can have up to 190 relationships, and the ratio of the total number of associations to the maximum number of possible relationships is 4.2. On the whole, the personal network

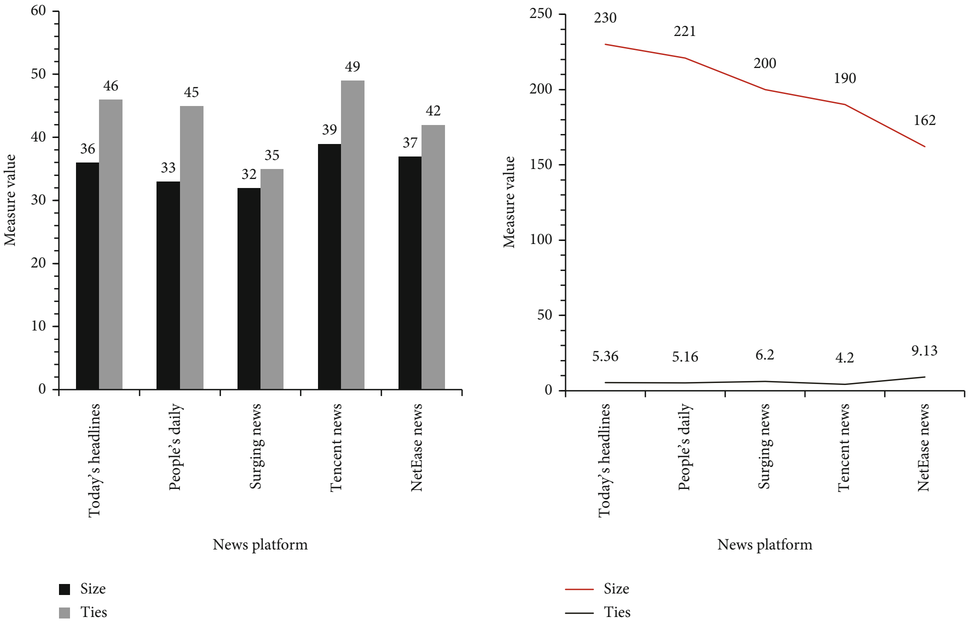

Figure 7: Measurement of "individual network density" in the overall network of public opinion of event a network.

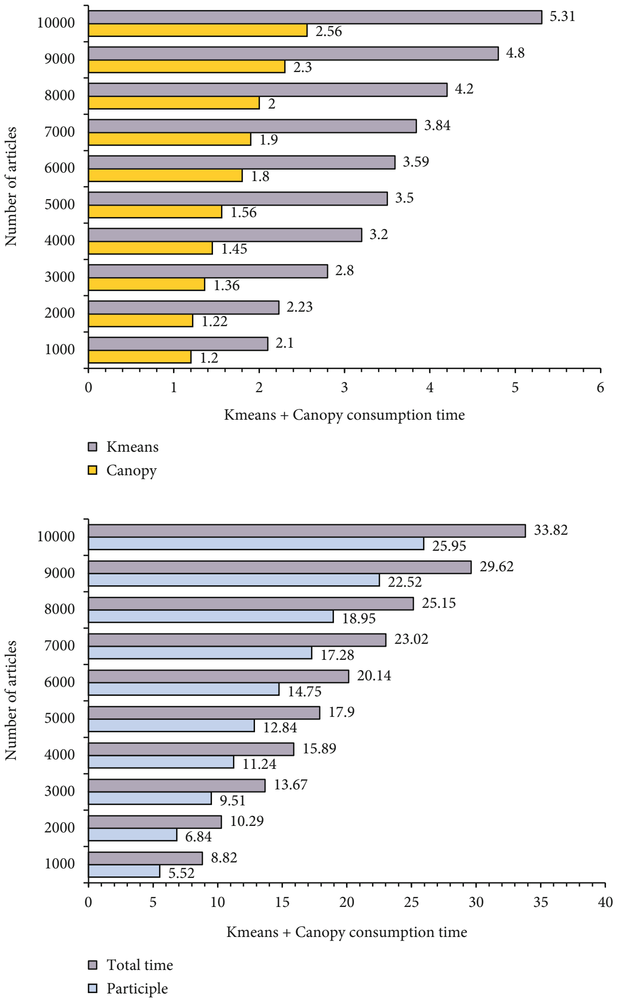

Figure 8: Time consumption of Kmeans + Canopy in each stage.

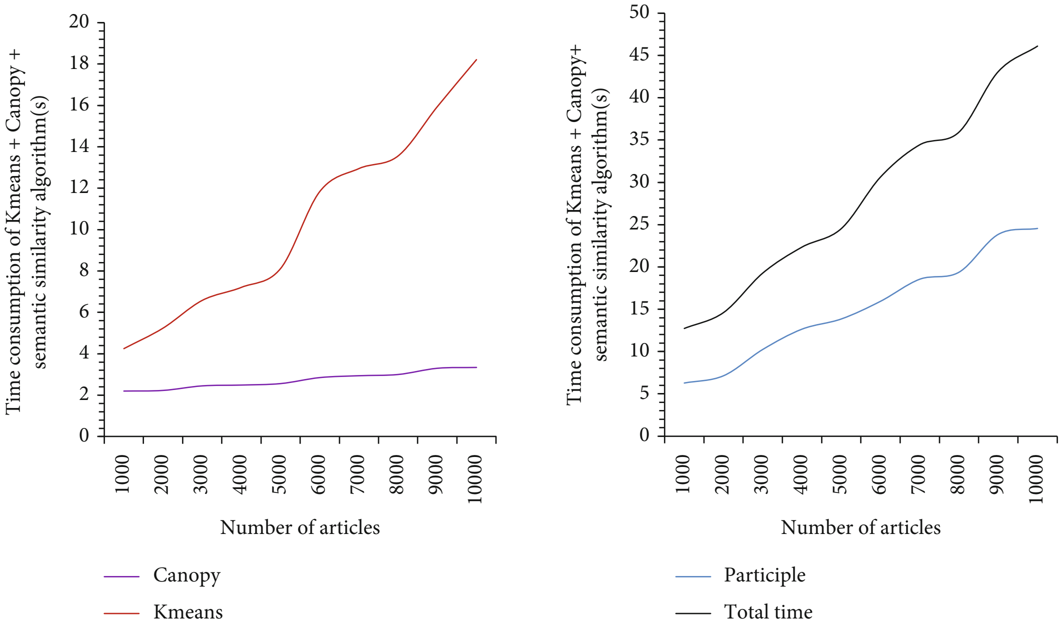

Figure 9: Time consumption of Kmeans + Canopy + semantic similarity algorithm.

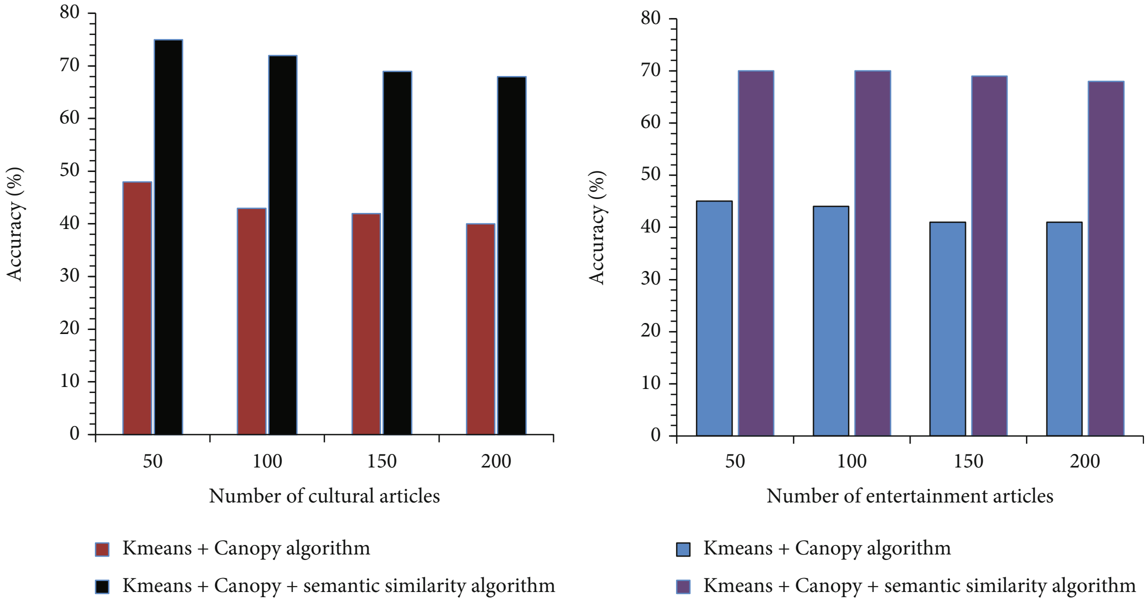

Figure 10: Accuracy comparison results of the two algorithms.

density of "Tencent News" is relatively high. The network density among all participants on the network is closely related to other nodes. This reflects the important role of authoritative media and online opinion leaders.

The Kmeans + Canopy algorithm and the Kmeans + Canopy + semantic similarity algorithm process 1000-10000 articles, respectively. The time consumed in each stage of the two clustering algorithms is shown in Figures 8 and 9.

Figures 8 and 9 analyze the time consumed by the Kmeans + Canopy algorithm and Kmeans + Canopy + semantic similarity in the three stages of word segmentation, Kmeans, and Canopy. Experimental results show that the two algorithms generally spend the most time in the word segmentation stage, and the Canopy stage spends the least time. The complexity of the Kmeans + Canopy + semantic similarity algorithm has increased, and the time spent in each stage of word segmentation, Canopy, and Kmeans has increased accordingly. Among them, the increased time in the word segmentation and Canopy phases is not obvious. The time consumed in the Kmeans phase is roughly twice that before the improvement, but the total time consumed does not exceed twice the time before the improvement.

The Kmeans + Canopy algorithm and the Kmeans + Canopy + semantic similarity algorithm process 50, 100, 150, and 200 cultural articles and 150 entertainment articles, respectively, and the accuracy comparison results are shown in Figure 10, respectively.

From the experimental results obtained in Figure 10, it can be seen that the clustering algorithm of Kmeans + Canopy is used to cluster the mixed two types of articles into multiple topic categories, the clustering convergence is poor, and the effect is very poor; using the clustering algorithm combining semantic similarity and Kmeans + Canopy, the two types of mixed articles are clustered into two topic categories, and the clustering effect is very ideal. Comparing the two algorithms, it can be seen that the clustering effect of the clustering algorithm using the combination of semantic similarity and Kmeans + Canopy is greatly improved, and the effect of corresponding public opinion analysis is also improved.

5. Conclusion

Artificial intelligence is a branch of computer science, a science that studies intelligent technology, that is, the use of artificial intelligence technology to develop smart devices or smart systems to simulate and expand human cognitive behavior. With the development of society, the spread of public opinion on the Internet has become an unstoppable trend. At the same time, more and more netizens pay attention to public opinion events through online media, and they actively comment on online public opinions, so proper regulation of online information is an inevitable trend. Therefore, based on the development trend of online public opinion, this article analyzes the collection method of online public opinion information. This article first analyzes the concept of online public opinion, introduces an online public opinion analysis system, and focuses on the development of textual information on online public opinion. In the clustering module, according to Chinese characteristics, such as synonyms and polysemous words, this paper proposes a clustering algorithm that combines Kmeans, Canopy, and semantic similarity and compares the performance of Kmeans + Canopy algorithm and Kmeans + Canopy + semantic similarity algorithm. Experiments have found that although the algorithm complexity increases and the system use time increases, the accuracy of clustering is improved. Especially for cultural articles, the accuracy rate is as high as 75%, and entertainment articles are as high as 70%, and stabilized at around 70%, thereby improving the discovery ability of online public opinion and achieving good research results. However, due to time and personal abilities, this article may have many areas for improvement. For example, this article only focuses on articles with large differences in content styles between culture and entertainment, the experimental data may not be representative, and multiple types of articles should be analyzed together.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest

The author states that this article has no conflict of interest.

Acknowledgments

This work was supported by the Natural Science Youth Program of Lingnan Normal University QN2116 and the Youth Fund for Humanities and Social Sciences Research Project of Ministry of Education: “Research on the Audience Communication of Network Rumors in the Big Data Era” (Project number: 18YJC860021).

References

[1] G. Wang, Y. Chi, Y. Liu, and Y. Wang, “Studies on a multidimensional public opinion network model and its topic detection algorithm,” *Information Processing & Management*, vol. 56, no. 3, pp. 584–608, 2019.

[2] M. S. Mayernik and A. Acker, “Tracing the traces: the critical role of metadata within networked communications,” *Journal of the American Society for Information Science and Technology*, vol. 69, no. 1, pp. 177–180, 2018.

[3] J. Kim, “How did the information flow in the #Alpha Go hashtag network? A social network analysis of the large-scale information network on twitter,” *Cyberpsychology, Behavior and Social Networking*, vol. 20, no. 12, pp. 746–752, 2017.

[4] G. Li, W. Niu, L. Batten, and J. Liu, “New advances in securing cyberspace and curbing crowdturfing,” *Concurrency and Computation: Practice and Experience*, vol. 29, no. 20, article e4162, 2017.

[5] L. Caviglione, M. Gaggero, J. F. Lalande, W. Mazurczyk, and M. Urbanski, “Seeing the unseen: revealing mobile malware hidden communications via energy consumption and artificial intelligence,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 4, pp. 799–810, 2016.

[6] M. Lin and Y. Zhao, “Artificial intelligence-empowered resource management for future wireless communications: a survey,” *China Communications*, vol. 17, no. 3, pp. 58–77, 2020.

[7] A. Chuan, “The national geographic characteristics of online public opinion propagation in China based on WeChat network,” *Geoinformatica: An international journal of advances of computer science for geographic*, vol. 22, no. 2, pp. 311–334, 2018.

[8] E. Davidov, “Nationalism and constructive patriotism: a longitudinal test of comparability in 22 countries with the ISSP,” *International Journal of Public Opinion Research*, vol. 23, no. 1, pp. 88–103, 2011.

[9] Y. Ming, J. Yang, J. Cao, Z. Zhou, and C. Xing, “Distributed energy sharing in energy internet through distributed averaging,” *Tsinghua Science and Technology*, vol. 23, no. 3, pp. 233–242, 2018.

[10] A. Mandes and P. Winker, “Complexity and model comparison in agent based modeling of financial markets,” *Journal of Economic Interaction and Coordination*, vol. 12, no. 3, pp. 469–506, 2017.

[11] F. Wang, Y. Zhang, Q. Rao, K. Li, and H. Zhang, “Exploring mutual information-based sentimental analysis with kernel-based extreme learning machine for stock prediction,” *Soft Computing*, vol. 21, no. 12, pp. 3193–3205, 2017.

[12] H. Lu, Y. Li, M. Chen, H. Kim, and S. Serikawa, "Brain intelligence: go beyond artificial intelligence," *Mobile Networks and Applications*, vol. 23, no. 7553, pp. 368–375, 2018.

[13] K. Lin, C. Li, D. Tian, A. Ghoneim, M. S. Hossain, and S. U. Amin, "Artificial-intelligence-based data analytics for cognitive communication in heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 83–89, 2019.

[14] L. D. Raedt, K. Kersting, S. Natarajan, and D. Poole, "Statistical relational artificial intelligence: logic, probability, and computation," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 10, no. 2, pp. 1–189, 2016.

[15] S. Jha and E. J. Topol, "Adapting to artificial intelligence," *JAMA*, vol. 316, no. 22, pp. 2353-2354, 2016.

[16] S. Makridakis, "The forthcoming artificial intelligence (AI) revolution: its impact on society and firms," *Futures*, vol. 90, no. jun., pp. 46–60, 2017.

[17] R. Li, Z. Zhao, X. Zhou et al., "Intelligent 5G: when cellular networks meet artificial intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, 2017.

[18] R. Liu, B. Yang, E. Zio, and X. Chen, "Artificial intelligence for fault diagnosis of rotating machinery: a review," *Mechanical Systems & Signal Processing*, vol. 108, no. AUG., pp. 33–47, 2018.

[19] J. H. Thrall, X. Li, Q. Li et al., "Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success," *Journal of the American College of Radiology*, vol. 15, no. 3, pp. 504–508, 2018.

[20] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725-726, 2018.

[21] J. Cao, E. M. van Veen, N. Peek, A. G. Renehan, and S. Ananiadou, "EPICURE: ensemble pretrained models for extracting cancer mutations from literature," in *2021 IEEE 34th international symposium on computer-based medical systems (CBMS)*, pp. 461–467, Aveiro, Portugal, 2021.

[22] J. Zhao, D. Zeng, J. Qin, H. M. Si, and X. F. Liu, "Simulation and modeling of microblog-based spread of public opinions on emergencies," *Neural Computing and Applications*, vol. 33, no. 2, pp. 547–564, 2021.

[23] F. Xiao, "Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy," *Information Fusion*, vol. 46, pp. 23–32, 2019.

[24] Y. Zhang, H. Huang, L. X. Yang, Y. Xiang, and M. Li, "Serious challenges and potential solutions for the industrial internet of things with edge intelligence," *IEEE Network*, vol. 33, no. 5, pp. 41–45, 2019.

[25] O. I. Khalaf and G. M. Abdulsahib, "*Optimized dynamic storage of data (ODSD) in IoT based on blockchain for wireless sensor networks*," *Peer-to-Peer Networking and Applications*, vol. 14, no. 5, pp. 2858–2873, 2021.