International Journal of Web Based Communities

ISSN online: 1741-8216 - ISSN print: 1477-8394
https://www.inderscience.com/ijwbc

A method for tracing big data of network public opinion based on data mining algorithms

Shumin Zhi, Lin Yu

DOI: 10.1504/IJWBC.2024.10061794

Article History:

Received:	19 May 2023
Last revised:	14 July 2023
Accepted:	10 October 2023
Published online:	04 November 2024

A method for tracing big data of network public opinion based on data mining algorithms

Shumin Zhi* and Lin Yu

Department of Health Management,
Zhengzhou Shuqing Medical College,
Zhengzhou, 450064, Henan, China
Email: Zhimin066@163.com
Email: 15639718219@163.com
*Corresponding author

Abstract: In order to achieve accurate traceability of massive public opinion data, this study carried out a study on the traceability method of network public opinion big data based on data mining algorithm. First of all, the network public opinion data is cleaned up and its data characteristics are mined. Then, the extracted public opinion features are taken as the input of the recursive neural network, which is used to construct the attention model and output the prediction results of the network public opinion. Finally, determine the network public opinion information to be tracked. Support vector machine is used to improve the probability packet tagging tracking algorithm and output the tracking results of public opinion information. The experimental results show that the implementation efficiency of this method is higher than 99%, and the average error of data tracing is less than 0.1, which has great application value.

Keywords: data mining algorithms; online public opinion; big data; traceability methods; kernel fuzzy clustering; probability packet labelling.

Reference to this paper should be made as follows: Zhi, S. and Yu, L. (2024) 'A method for tracing big data of network public opinion based on data mining algorithms', *Int. J. Web Based Communities*, Vol. 20, Nos. 3/4, pp.245–262.

Biographical notes: Shumin Zhi received her Master's in Computer Technology from the Zhengzhou University in 2015. She is an Associate Professor in the Department of Health Management of Zhengzhou Shuqing Medical College. Her research interests include computer network technology, information security and data mining technology.

Lin Yu graduated from the School of Computer Science of Henan University with a Master's degree in 2009. Currently, she is working in the Department of Health Management of Zhengzhou Shuqing Medical College, and her main research fields are computer network and web front-end.

1 Introduction

At present, big data technology has become an important technology for transmitting and sharing resources in many industries. Different fields apply big data technology to achieve efficient data transmission and sharing. With the growth of network data volume,

how to mine useful information from massive data is a research difficulty in the field of data mining. With the rapid development of internet technology, online media has become an efficient way of information dissemination (Karimi et al., 2021). People are accustomed to using the internet to express public emotions, and the internet has become an important channel for people to communicate and express emotions. As an important medium, it has the characteristics of openness, freedom, and interaction (Li et al., 2021), which more attracts the public to use it as a means of information dissemination. Users are good at using online channels to express their opinions on urgent situations and hot topics in the network. In this context, news comments, Weibo, and other means have become the main ways for online users to express their opinions. The development of online public opinion has a significant impact on society. When negative information appears in the network, if network users are unable to correctly guide or identify negative public opinion in the network, it will have a very serious impact on society. Therefore, tracing the source of online public opinion and clarifying the source of public opinion formation is very important, which helps network public opinion management personnel make specific management actions based on the content of public opinion (Ma and Lang, 2023; Yang, 2020; Lu et al., 2020).

When dealing with big data of online public opinion, traditional data processing methods cannot meet the needs of massive data usage and cannot provide users with the information they need quickly. Data mining technology is an important technology commonly used in big data processing, which can provide technical support for the processing of massive data. At present, using data mining technology to analyse online public opinion has become a hot topic in dynamic monitoring of network information. Data traceability is an emerging technology derived from the continuous development of big data technology, which involves finding the source of data and the process of forming data. Applying data traceability to the traceability of online public opinion is of great significance for controlling the development of online public opinion. At present, many researchers have studied the traceability of online public opinion data, and Zhao et al. (2020) have studied the traceability of online public opinion big data. This method uses cryptography to process online public opinion data, sets data objects as the operating unit of data processing, and uses the data object version number calculation method to complete the generation and traceability of online public opinion data. This method can trace the source of online public opinion data, but its accuracy and applicability are poor; Xu et al. (2020) applied the K-means clustering method to the traceability of online public opinion, effectively collecting indicator data of online public opinion events, obtaining sets of different categories of online public opinion events, and using the recognition and classification results of public opinion to achieve public opinion traceability. This method can provide a basis for tracing the sources of different types of online public opinion and provide effective basis for accurate analysis of online public opinion. However, it has the characteristic of strong subjectivity and is easily affected by other factors in the network, which reduces tracing performance; Liu et al. selected the link text to propose the network method (Liu et al., 2020) to achieve accurate positioning of online public opinion text information, selected the multi granularity potential Dirichlet distribution method to extract topics associated with online public opinion and achieve accurate traceability of online public opinion. This method can achieve traceability for different types of data such as images, text, and videos in the network, but it has high computational complexity and poor real-time traceability.

In response to the problems of the above methods in tracing online public opinion, a big data tracing method for online public opinion based on data mining algorithms is studied to achieve secure management of online public opinion. This method first uses data classification induction to achieve data cleaning, and uses incremental kernel fuzzy clustering algorithm to achieve data mining. Then, based on the attention model, it predicts the development trend of network public opinion, confirms the network public opinion information to be traced. Finally, based on the support vector machine, it improves the probability packet label tracking algorithm, and achieves network public opinion information tracing based on the improved probability packet label tracking algorithm. I hope that this study can provide a theoretical basis for the design and research of traceability methods for online public opinion data in the future.

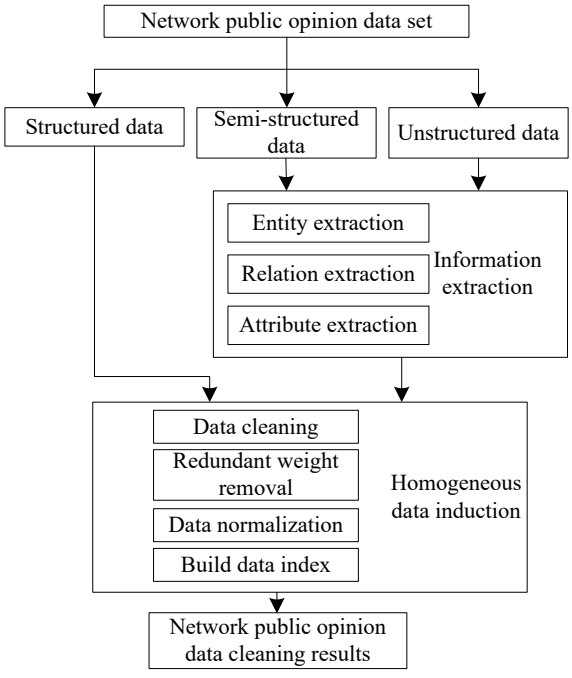
2 Material method

2.1 Cleaning of network public opinion data

The data volume of online public opinion data is enormous, and when tracing its source, it is necessary to first classify, merge, and label the data. The data in online public opinion mainly includes unstructured data and semi-structured data, so it is necessary to clean the big data of online public opinion in advance (Mukunthan, 2021). Preliminary extraction of valuable data through information cleansing. This study selected the data classification and induction method to clean online public opinion data, and preliminarily extracted the big data of online public opinion. The cleaning process is shown in Figure 1.

From the flow chart of cleaning online public opinion big data in Figure 1, it can be seen that this study first divides online public opinion big data into different types of data, then extracts information from all online public opinion data, and finally classifies and processes it layer by layer based on its purpose. Access records and access logs in the network belong to structured data, while service background error records, service operation records and other data in the network belong to semi-structured data. Data such as network operation status and network original message belong to unstructured data. For redundant data in online public opinion big data, the same type of data induction method is used for deduplication (Jiang et al., 2021). Through the above processing process, the network public opinion big data is processed into valuable high-quality inductive data, and redundant data is deleted to avoid interference with public opinion data traceability. After completing the induction and cleaning of data of the same category, extract and classify the big data of online public opinion based on the purpose of the data. Based on the data characteristics of online public opinion big data such as sending time, sending frequency, geographical location, and sending status, further classification is carried out. Based on the data classification results, a data index is constructed to provide a basis for tracing the source of subsequent online public opinion big data.

Figure 1 Flowchart of cleaning network public opinion big data



2.2 Incremental kernel fuzzy clustering algorithm for mining big data of network public opinion

After completing data cleaning, mine the data to extract valuable information. The big data of online public opinion mainly refers to the collection of attitudes and cognition towards events that occur in life and society, through online communication channels and utilising online communication methods. When users publish information about public opinion big data online, they can anonymously publish it, making it easy for many online users to use the internet as an important way of communication and venting. Different types of speech targeting different events may appear on the internet, forming a massive amount of data information. Data information spreads rapidly in the network, and the group effect of network users accelerates the spread of online public opinion, gradually expanding the scope of public opinion influence in the network. Therefore, this study is based on the changes and dissemination characteristics of online public opinion big data in the network, and based on the principle of fairness, analyses the characteristics of online public opinion big data.

The characteristics of online public opinion big data are shown in Table 1.

The transmission characteristics of big data of network public opinion constructed in Table 1 provide a good foundation for the follow-up big data traceability of network public opinion. The specific trend and development trend of network public opinion can be intuitively presented through the big data communication characteristics table of network public opinion in Table 1. The incremental kernel fuzzy clustering algorithm allows data to be processed step by step without the need to load the entire data set at once, which is very useful for dealing with large-scale data sets and can reduce memory

consumption and computational complexity, and the algorithm also allows incremental learning in the data stream, which can handle new data in the dynamic data environment without retraining the entire model. Optimise the training process and improve the mining efficiency. Therefore, in order to mine useful information in big data of online public opinion, this study selects incremental kernel fuzzy clustering algorithm to perform cluster analysis on big data of online public opinion (Feng et al., 2020).

Table 1 Feature table of big data of online public opinions

Index feature	*Index name*	*Quantitative form of index*
Normal characteristics	Event subject	LDA document
Anomalous feature	Topic sensitivity	Evaluation result
Information feature	Information type	Degree of socialisation
Narrative formal features	User attention	Cumulative number of publications
Characteristics of transmission mode	Velocity and flux of public opinion	Duration

According to the importance of data points in the big data of network public opinion, this algorithm is applied in this paper to assign different weights to different data points to achieve data clustering (Song et al., 2020). Use $X = \{x_1, x_2, \ldots, x_n\}$ to represent the dataset of online public opinion big data, and use w_i to represent the corresponding weights of data points within the dataset. The incremental kernel fuzzy clustering algorithm randomly selects k initial clustering centres from the network public opinion big data set. The membership expression from the data points in the big dataset of online public opinion to the clustering centre of the dataset is as follows:

$$u_{ji} = \sum_{l=1}^{k} \left(\frac{d_{ij}^2}{d_{li}^2} \right)^{\frac{1}{m-1}} \tag{1}$$

After calculating the membership degree of each data point to the cluster centre, the cluster centre is updated in real-time.

The expression for updating the network public opinion big cluster analysis centre is as follows:

$$v_j = \frac{\sum_{i=1}^{n} w_i u_{ji}^m x_i}{\sum_{i=1}^{n} w_i u_{ji}^m} \tag{2}$$

In the above formula, v_j and m represent the cluster centre and fuzzy coefficient, respectively, d_{ji} and u_{ji} represent the distance from data point x_i to cluster centre v_j, and the degree to which data point x_i belongs to cluster centre v_j. The calculation formula for the distance d_{ji} from data point x_i to cluster centre v_j in formula (1) is as follows:

$$d_{ji} = \left\| x_i - v_j \right\| \tag{3}$$

By using the clustering update formula of formula (2), update the clustering centre of network public opinion big data mining, obtain a new clustering centre (Li et al., 2020), and recalculate the membership degree of each data point to the new clustering centre.

Repeat the above process until it meets the convergence conditions set by Big data traceability of network public opinion. The conditions are as follows:

1 All u_{ji} are fixed and will no longer change.

2 All v_j are fixed and will no longer change.

3 s and ε represent the iteration steps and convergence accuracy of data mining in the traceability of network public opinion big data, respectively. $J(U, V)$ represents the criterion function, and the formula for calculating the criterion function of network public opinion big data mining is as follows:

$$J(U,V)=\sum_{j=1}^{k}\sum_{i=1}^{n}w_i u_{ji}^m d_{ji}^2 \tag{4}$$

The convergence criterion expression for setting data mining is as follows:

$$\left|J^{s+1}(U,V)-J^s(U,V)\right|<\varepsilon \tag{5}$$

When the above conditions are met, terminate the mining of network public opinion big data using incremental kernel fuzzy clustering algorithm.

The dataset of online public opinion big data belongs to a linearly indivisible state. When using an incremental kernel fuzzy clustering algorithm for clustering, a kernel function is used to map the samples of online public opinion big data (Zhang et al., 2021) to a high-dimensional space, converting the data into linearly separable. The kernel function expression for mapping low-dimensional network public opinion data to high-dimensional space is as follows:

$$k\left(x_i,x_j\right)=\delta\left(x_i\right)^T\delta\left(x_j\right) \tag{6}$$

In formula (6), δ represents the mapping function.

Gaussian kernel function and polynomial kernel function are commonly used kernel functions in incremental kernel fuzzy clustering algorithms. The expression for the operational distance of the incremental kernel fuzzy clustering algorithm after kernel function optimisation is as follows:

$$d_{ji}^2=\left\|\delta\left(x_i\right)-\delta\left(v_j\right)\right\|=\delta\left(x_i\right)^T\delta\left(x_i\right)+\delta\left(v_j\right)^T\delta\left(v_j\right)-2\delta\left(x_i\right)^T\delta\left(v_j\right) \tag{7}$$

According to formula (7), convert the membership formula of formula (1) as follows:

$$u_{ji}=\sum_{l=1}^{k}\left(\frac{\left\|\delta\left(x_i\right)-\delta\left(v_j\right)\right\|^2}{\left\|\delta\left(x_i\right)-\delta\left(v_l\right)\right\|^2}\right)^{\frac{1}{m-1}} \tag{8}$$

Convert the clustering centre expression of formula (2) as follows:

$$\delta\left(v_j\right)=\frac{\sum_{i=1}^{n}w_i u_{ji}^m\delta\left(x_i\right)}{\sum_{i=1}^{n}w_i u_{ji}^m} \tag{9}$$

When $A_t = \{a_1^t, a_2^t, \ldots, a_k^t\}$ represents time $t-1$, the network public opinion big data is clustered and mapped to obtain a set of data points. Assuming that $a_j^t = \sum_{i=1}^{n} a_{ji}^t \delta(x_i^t)$ and V_{t-1} are mapped to obtain A_t, the expression for A_t is obtained by solving the optimisation problem as follows:

$$O = \min \left\| \delta\left(v_j^{t-1}\right) - a_j^t \right\| \tag{10}$$

Introduce formula (10) and use incremental kernel fuzzy clustering algorithm to mine network public opinion big data. When step t is used, the clustering centre expression obtained through clustering is as follows:

$$\delta\left(v_j^t\right) = \frac{\sum_{l=1}^{k} w_l^t \left(u_{ij}^t\right)^m + \sum_{s=1}^{k} a_{sl}^t w_s^{t,a} \left(u_{iS}^a\right)^m}{\sum_{s=1}^{k} w_s^t \left(u_{iS}^t\right)^m + \sum_{s=1}^{k} w_s^{t,a} \left(u_{iS}^a\right)^m} \delta\left(x_l^t\right) \tag{11}$$

Set the weight of data sample $x_i^t \in X_t$ in the big data of online public opinion to 1. For the transfer point a_j^t of the incremental kernel fuzzy clustering algorithm, the weight $w_j^{t,a}$ calculation formula is as follows:

$$w_j^{t,a} = \sum_{i=1}^{n} u_{ji}^t w_i^t + \sum_{s=1}^{k} u_{js}^a w_s^{t-1,a} \tag{12}$$

From this, the final membership degree u_i^t of network public opinion data point x_i^t can be calculated as follows:

$$u_i^t = \sum_{l=1}^{k} \left(\frac{\left\| \delta\left(x_i^l\right) - \delta\left(v_j^t\right) \right\|^2}{\left\| \delta\left(x_i^l\right) - \delta\left(v_l^t\right) \right\|^2} \right)^{-\frac{1}{m-1}} \tag{13}$$

At this point, the formula for calculating the membership $u_j^{t,a}$ of transfer point a_j^t is as follows:

$$u_j^{t,a} = \sum_{l=1}^{k} \left(\frac{\left\| \delta\left(a_j^t\right) - \delta\left(v_j^t\right) \right\|^2}{\left\| \delta\left(a_j^i\right) - \delta\left(v_l^t\right) \right\|^2} \right)^{-\frac{1}{m-1}} \tag{14}$$

Using incremental kernel fuzzy clustering algorithm to continuously update and re cluster the newly added data blocks and clustering results obtained from network public opinion big data, using the same clustering steps to effectively mine network public opinion big data.

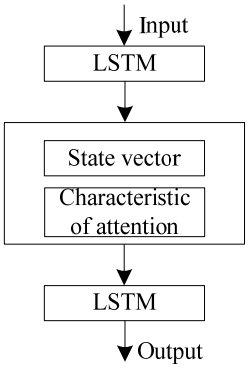
2.3 Prediction of the development trend of online public opinion based on attention model

After using the incremental kernel fuzzy clustering algorithm to mine the features of network public opinion big data, based on the mined features, predict the development trend of network public opinion and determine the network public opinion information to be traced. Firstly, construct an attention model using a recurrent neural network to predict the development status of online public opinion big data. Recurrent neural networks can effectively remember the relationships between network public opinion sequences (Yang et al., 2020), and can still accurately process input of variable length sequences. Use $x = \{x_1, x_2, \ldots, x_T\}$ to represent the transmission sequence of network public opinion big data, and use h_t to represent the operation of step t in the hidden layer of the recurrent neural network. The update formula for the hidden layer is as follows:

$$h_t = f\left(h_{t-1}, x_t\right) \tag{15}$$

In formula (15), f represents the nonlinear activation function. $y = \{y_1, y_2, \ldots, y_t\}$ represents the output of a recurrent neural network, where both the input and output are variable length information sequences. The recurrent neural network obtains the probability distribution on the network public opinion information sequence through the training process. Use the conditional distribution $p(x_t|x_{t-1}, \ldots, x_1)$ to represent the output of the running steps of the recurrent neural network. During the gradient descent process of recurrent neural network operation, gradient explosions and other situations are prone to occur. In response to the traceability application requirements of online public opinion big data, an attention model structure diagram for predicting the development of online public opinion is constructed using a recurrent neural network, as shown in Figure 2.

Figure 2 Attention model of online public opinion development prediction

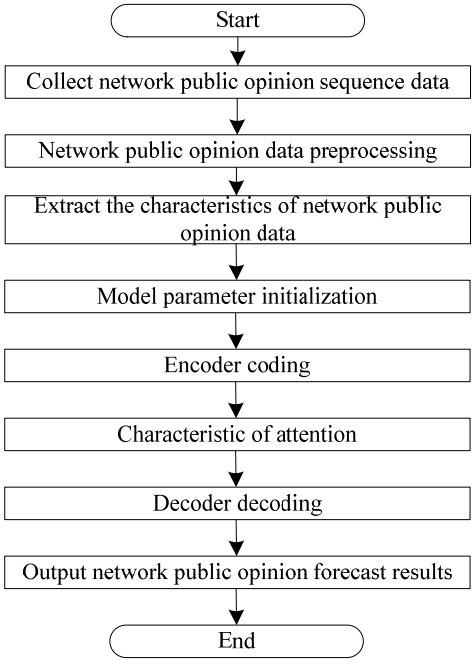


The recurrent neural network is used as the basic unit of the encoder and decoder of the network public opinion information prediction model, and the state vector containing the network public opinion sequence information is used as the input, that is, the feature combination vector x_t is used as the input of the encoder in the attention model, and the hidden layer variable is represented by h_t. When time is t, the prediction sequence obtained is network output, named y_t, which is related to attention characteristics, state vector and historical output. To improve the predictive performance of attention models for online public opinion information, a local attention mechanism is introduced, and the

pooling layer is used to transmit the output results to the fully connected layer, reducing the dimensionality of online public opinion big data. Transfer the attention features of network public opinion output by the recurrent neural network to the decoder, and output the predicted results of network public opinion development.

The flowchart of using the attention model to output the predicted results of network public opinion development is shown in Figure 3.

Figure 3 Big data prediction flow chart of online public opinion of attention model



Analysing Figure 3, the prediction process of big data for online public opinion is as follows:

1 By utilising the feature mining results of big data on online public opinion and predicting the development needs of online public opinion, the online public opinion information is divided into training sets and testing sets.

2 Analyse the characteristics of online public opinion information and determine the appropriate features of online public opinion information. Normalise and standardise the identified network public opinion features, and input the processed network public opinion information into the encoder.

3 Initialise the parameters of the improved attention mechanism model.

4 Encode the time series of network public opinion information using an encoder, generate the state vector of network public opinion, and use a decoder to decode and process the network public opinion information.

5 Transfer historical data related to online public opinion information as key nodes to a recurrent neural network. Utilise the fully connected layer to reduce the

dimensionality of network public opinion data, and encode the output results through the encoder of key nodes in the recurrent neural network. Use the fully connected layer to obtain the attention features of the final network public opinion information.

6 Transfer the obtained attention features to an improved attention mechanism decoder to provide historical information for predicting network public opinion information;

7 Train and improve the attention model, using the completed attention model to output network public opinion prediction results.

2.4 Traceability of network public opinion big data based on improved probability packet labelling tracking

Based on the prediction results of the development of online public opinion, determine the network public opinion information to be traced, and perform traceability processing on the determined network public opinion information. Using probability packet labelling tracking algorithm to trace the source of online public opinion big data. The so-called probability packet labelling algorithm is also known as the basic packet labelling method. The probability packet marking algorithm searches for specific areas in the packets selected by the router and records information containing IP addresses. After capturing and sending network public opinion data packets, analyse network public opinion data, search for effective data from labelled information, reconstruct the transmission path of network public opinion data packets, and complete network public opinion information traceability. If you need to obtain the complete output path of network public opinion, you need to sort the IP addresses sent by the network information according to fixed rules. Use φ to represent the probability of marking when sending network public opinion information, and l to represent the distance between the router sending public opinion information and the tracking information router. The probability of receiving routing information for sending network public opinion is related to the communication distance. Personnel receiving public opinion information can sort the probability of the received network public opinion router information based on fixed rules and order, and obtain the actual transmission route of network public opinion.

The probability packet labelling traceability algorithm can achieve traceability only through routers, and it does not need to label all network public opinion information when labelling the probability of network public opinion information. This study combines the probabilistic packet labelling traceability algorithm with the support vector machine algorithm to avoid the problem of high false alarm rate caused by overly complex calculations during path reconstruction, and also to avoid the impact of excessive labelling on the routing and forwarding function of network public opinion information. Optimising it using support vector machines can improve the judgment time for the source of network public opinion and reduce the router load for network public opinion transmission. The setting of marker space in probability packet labelling tracking algorithms is extremely important. To accurately trace network public opinion information, it is necessary to select an appropriate traceability probability. When the probability of using the probability packet labelling tracking algorithm to successfully trace the network source is φ, and the probability of using a router with a random distance as the final labelled data packet is $1/\varphi$, the minimum number of data packets is required for reconstructing the network public opinion information transmission path.

To determine the label space of data packets, setup network public opinion transmission under the IPV4 network. IPV4 is a commonly used communication protocol in mobile communication networks, occupying an important position in internet communication. The header field settings of IPV4 networks are shown in Figure 4.

Figure 4 IPV4 network packet header field

Version	Differentiated service		Total length
Identification	DM	MF	Piecewise offset
Survival period	agreement		Head check
Source address			
Destination address			
Option			

The IPV4 network is a 32-bit address with a maximum length of 65,535 bytes. In IPV4 networks, DF and MF represent non-segmented and more segments, respectively. Use the segmented offset field to determine the specific position of the field in the data message. The IP packet header in IPV4 network consists of three parts: DF term, identification term, and offset term. The MF domain and DF domain are used to determine the state of packet sharding.

When using packet labelling algorithm to trace the source of network public opinion big data, boundary routers still filter and label the transmitted information that is not related to network public opinion. When data packets with the same public opinion characteristics are forwarded, as they have already been labelled and analysed by the router, it will consume a large amount of router computing resources and increase the communication load of the network. Therefore, this study utilises an improved packet labelling strategy to filter and process big data that does not belong to online public opinion. A feature model is constructed for already labelled online public opinion data packets, and the network public opinion transmission path is reconstructed to achieve traceability of online public opinion big data.

Using the support vector machine method to improve the packet labelling traceability algorithm, the improvement process is as follows:

1 Initialise: Using data packets from online public opinion big data as input samples for the support vector machine algorithm.

   Due to the high uncertainty in the number of online public opinion hotspots, it belongs to a multi classification problem. The network public opinion information is transformed from low dimensional space to high dimensional space by using nonlinear functions. Construct a discriminant function for network public opinion information in high-dimensional space, and use the constructed discriminant function to achieve nonlinear discrimination of the original network public opinion information. The expression of the support vector machine output network public opinion classification decision function is as follows:

$$f(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} y_i k(x_i, x) + b^*\right] \tag{16}$$

In formula (16), $k(x_i, x)$ represents the radial basis function of the support vector machine, and b represents bias.

Using the radial basis function as the inner product kernel function, its expression is as follows:

$$k(x_i, x) = \exp\left(\gamma \left\| x_i - x_j \right\|^2\right) \tag{17}$$

In formula (17), γ represents the kernel parameter.

Obtain the distribution of public opinion hotspots over a fixed time period using support vector machines.

2 Construct the ideal output values of the support vector machine and the termination conditions of the algorithm.

Evaluate each network public opinion information packet and use the network public opinion packet value as the ideal output value of the support vector machine, which is the termination condition of the support vector machine. Based on the actual output of network public opinion data packet values, monitor whether the network public opinion information meets the network public opinion data packet.

3 Number of refreshing packets.

4 Determine whether the output of online public opinion meets the ideal traceability results. When the ideal traceability result is met, the algorithm terminates, marks the network public opinion data packet, and uses it as the transmission path for network public opinion; when the ideal traceability result is not met, return to the second step and re iterate until the final network public opinion big data traceability result is output. Utilising the adaptive characteristics of support vector machines to cope with network public opinion, and reducing the initial labelling workload through continuous learning.

The public opinion data in this study are all transmitted in the IPV4 network. Because the single probability packet tag tracking algorithm is affected by the large amount of data, there is a problem of slow traceability operation. In order to better achieve big data traceability, this study uses support vector machine to improve the probability packet tag tracking algorithm, converting public opinion data from low dimensional space to high dimensional space, so as to improve the real-time quality of online public opinion traceability. Through the above research, the quality and efficiency of data traceability can be improved.

3 Experimental analysis

3.1 Experimental environment settings

In order to verify the traceability effectiveness of the network public opinion big data studied, experiments were carried out on a laptop with Windows 10 operating system.

Collect online public opinion data on different topics such as the two sessions, CBA, party building, and rural revitalisation, and trace the source of the collected online public opinion data. This experiment collected a total of 564.3 GB of online public opinion information, of which 300 GB was used for model training and the remaining 264.3 GB was used for testing. Select method of Zhao et al. (2020), method of Xu et al. (2020), and method of Liu et al. (2020) as comparative methods to verify the traceability performance of the method proposed in this paper for online public opinion big data.

The communication network parameter settings for verifying the effectiveness of network public opinion big data traceability are shown in Table 2.

Table 2 Communication network parameter settings

Parameter	*Numerical value*
Number of sessions	20
Terminal quantity	20
Ratio of redundant data	50%
Data transfer rate	1000 Mb/s
Server storage space	10 TB
Network transmission delay	10 ms

The storage devices and communication servers of big data communication networks have a direct impact on data processing. The description of online public opinion big data for the experimental setup is shown in Table 3.

Table 3 Description results of network public opinion big data

Index	*Content*
Data size	Memory size
Data source	Server number
Data time	Data transfer date
Data structure	Structural feature
Data linearity	Linear feature
Relevance	Associations with other data
Dimension	Big data dimension

This research first verifies the functionality of the method by comparing the duty cycle distribution changes of the network public opinion big data processed by this method, then compares the clustering execution efficiency of the three methods to verify the performance of the incremental kernel fuzzy clustering algorithm, and finally selects the average percentage error and rate of convergence after traceability processing, network overhead, routing overhead, fast tracking, tracking post-processing capability and other performance as evaluation indicators, complete a comprehensive verification of the effectiveness of the method.

3.2 Comparative performance analysis

The distribution of the network duty cycle of online public opinion big data has a high impact on the processing performance of online public opinion big data. When the duty

cycle of online public opinion big data is insufficient, it will affect the recognition and mining efficiency of online public opinion big data, and restrict the traceability results of online public opinion big data. Usually, the recognition efficiency of big data with a high proportion of characters is relatively low. In online public opinion big data, uneven distribution of the duty cycle defined by data attributes has a significant impact on the execution efficiency of online public opinion big data. The distribution of the duty cycle of online public opinion big data after processing using this method is shown in Figure 6.

Figure 5 Clustering execution efficiency is affected by the number of sessions

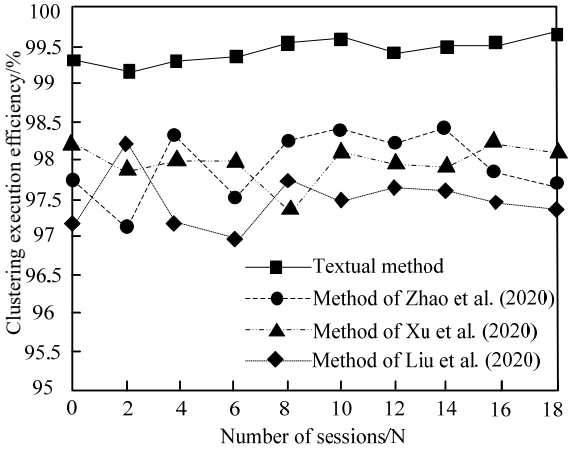
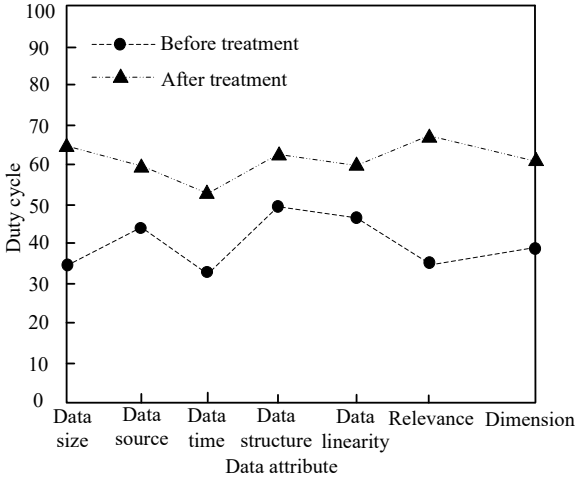


Figure 6 Change of duty ratio distribution of big data of online public opinions



Analysing the experimental results in Figure 6, it can be seen that after using the method proposed in this article to process online public opinion big data, the duty cycle of attributes such as relevance of online public opinion big data has significantly improved. The increase in the duty cycle of different attributes will lead to an improvement in the recognition and mining efficiency of online public opinion big data. After using the

method described in this article to process online public opinion big data, the uniformity of data attribute duty cycle allocation is significantly improved, proving that it can effectively improve the mining performance of online public opinion big data. By using the method proposed in this article to process online public opinion big data, the duty cycle of big data dimensions and correlation attributes has significantly increased. This method can achieve average equalisation of online public opinion big data, effectively balancing the structural and nonlinear characteristics of big data servers and communication network storage devices. Through the processing of online public opinion big data, the structural information of big data has been weakened, and the performance of multidimensional correlation in describing online public opinion big data has been improved.

This article adopts an incremental kernel fuzzy clustering algorithm to mine big data of online public opinion. The clustering execution efficiency of the three methods for different data sessions is shown in Figure 5.

From the experimental results in Figure 5, it can be seen that with the increasing number of sessions, the clustering execution efficiency of the four methods shows significant fluctuations. Using this method to mine online public opinion big data, the clustering execution efficiency of this method for mining online public opinion big data is higher than 99% for different sessions. Using other methods to mine the big data of network public opinion, the clustering execution efficiency is lower than 98.5%. The experimental results show that the method proposed in this paper has high data mining performance and can effectively mine useful information from massive online public opinion big data, providing a data foundation for tracing the source of online public opinion big data.

Using the method presented in this article, the tracing results for different online public opinion sources are shown in Table 4.

Table 4 Results of online public opinion tracing

Network public opinion topic	*Traceable IP result*	*Actual IP result*
Two sessions	58456545	58456545
CBA	84561622	84561622
Party building	105618152	105618152
Rural revitalisation	26185481	26185481
Oil price	20313135	20313135
Three gorges project	78994555	78994555
quadruplets	25413358323	25413358323
Ya Ya Panda	255852213	255852213
License plate	3321851238	3321851238
Blood sugar	184566452	184566452

From the experimental results in Table 4, it can be seen that using this method can achieve traceability of online public opinion based on data mining results. In Table 4, for online public opinion such as the two sessions, CBA, and party building, the methods proposed in this article can effectively trace the source of online public opinion. Network public opinion management personnel can determine the IP information of the sender of

online public opinion based on the results of network public opinion tracing, and accurately locate the sender of online public opinion.

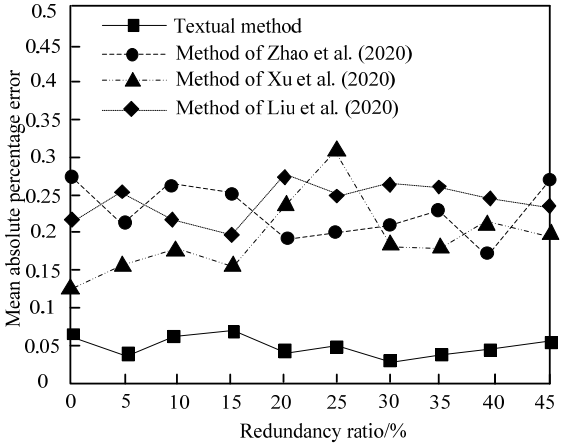
Select the average percentage error as an important indicator to measure the traceability performance of online public opinion big data, and the calculation formula for this indicator is as follows:

$$E = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - t_i}{y_i + t_i}\right| \tag{18}$$

In formula (18), y_i and t_i represent the actual source address and the source address displayed in the traceability result, respectively.

The average percentage error can measure the degree of deviation between the trend of online public opinion time series and the actual time series, and can be used to measure the traceability accuracy of online public opinion big data. Statistical methods are used to trace the source of online public opinion big data, and the average absolute percentage error of tracing is shown in Figure 7 for different redundant data ratios.

Figure 7 Average absolute percentage error of online public opinion tracing



From Figure 7, it can be seen that the method used in this article for tracing online public opinion big data has an average absolute percentage error of less than 0.1. Although the average absolute percentage error of tracing online public opinion big data using other methods is less than 0.35, it is significantly higher than the method used in this article. The experimental results verify that the accuracy of using this method to trace online public opinion sources is significantly higher than other methods, and the application performance is higher.

The computational performance of using different methods to trace the source of online public opinion big data is statistically analysed, and the comparison results of the tracing performance of online public opinion big data are shown in Table 5.

Analysis of Table 5 shows that using the method proposed in this article to trace the source of online public opinion big data significantly outperforms other methods in terms of network overhead, routing overhead, and convergence speed of operations. This proves that the method proposed in this paper has high traceability performance, strong post processing ability, and can support rapid traceability of big data of online public

opinion, achieving efficient application in online public opinion management. Looking at the intra table indicators, the convergence speed has a very high impact on the subsequent traceability time of the traceability method, and is an important indicator for measuring the big data of online public opinion. The post processing ability of tracing the source of online public opinion big data determines the effectiveness of addressing the traceability of online public opinion big data. When additional expenses arise during the traceability process of online public opinion big data, it will affect the timeliness of subsequent online public opinion big data traceability. The router overhead of communication networks also has a high impact on the router's data forwarding work, which can have an impact on the tracing effect of network public opinion in the later stage. The evaluation results of different traceability index values using this method are significantly better than other methods, verifying the traceability effectiveness of this method.

Table 5 Comparison of big data traceability performance of online public opinions

Index name	*Textual method*	*Method of Zhao et al. (2020)*	*Method of Xu et al. (2020)*	*Method of Liu et al. (2020)*
Rate of convergence	Fast	Slow	Fast	Medium
Network overhead	Low	In the	In the	High
Routing overhead	Low	In the	High	Low
Fast tracking	Support	Do not support	Support	Support
Trace post-processing capability	Support	Support	Support	Do not support

4 Conclusions

Tracing the source of online public opinion big data can intuitively describe online public opinion. With the gradual evolution of time, online public opinion big data tracing is suitable for many applications such as data recovery and data verification. In the big data environment, tracing the source of online public opinion is an urgent and important issue that needs to be addressed. Therefore, this study proposes a network public opinion big data traceability method based on data mining algorithms. This study first cleaned and processed the big data of online public opinion, and then used the incremental kernel fuzzy clustering algorithm as a mining method for tracing the source of online public opinion big data. Then, the mining results of online public opinion big data were used to trace the source of online public opinion big data. Finally, the effectiveness of the proposed method was demonstrated through experiments. This method can achieve effective traceability of online public opinion big data, with an average absolute percentage error of less than 0.1 for online public opinion traceability. It has high traceability reliability and can be applied in practical applications of online public opinion management.

References

Feng, Q., Chen, L., Chen, C. and Guo, L. (2020) 'Deep fuzzy clustering – a representation learning approach', *IEEE Transactions on Fuzzy Systems*, Vol. 28, No. 7, pp.1420–1433.

Jiang, H., Li, L., Xian, H., Hu, Y. and Wang, J. (2021) 'Crowd flow prediction for social internet-of-things systems based on the mobile network big data', *IEEE Transactions on Computational Social Systems*, Vol. 36, No. 99, pp.1–12.

Karimi, S., Shakery, A. and Verma, R. (2021) 'Online news media website ranking using user-generated content', *Journal of Information Science*, Vol. 47, No. 3, pp.340–358.

Li, J., Mi, Y., Shi, Y., Liu, W. and Yan, M. (2020) 'Fuzzy-based concept learning method: exploiting data with fuzzy conceptual clustering', *IEEE Transactions on Cybernetics*, Vol. 42, No. 1, pp.1–12.

Li, Y., Shyamasundar, R.K. and Wang, X. (2021) 'Special issue on computational intelligence for social media data mining and knowledge discovery', *Computational Intelligence*, Vol. 37, No. 2, pp.658–659.

Liu, R.Q., He, X.S., Nan, Y.F. and Wang, B. (2020) 'Mining method of public opinion related topic in network multimedia data', *Journal of Shenzhen University (Science & Engineering)*, Vol. 37, No. 1, pp.72–78.

Lu, J., Zhang, X., Liu, X., Hu, T. and Cao, Y. (2020) 'An equalisation control method for network big data transmission based on parallel computing', *International Journal of Internet Protocol Technology*, Vol. 13, No. 1, pp.32–41.

Ma, Y.J. and Lang, W. (2023) 'Trend prediction of internet public opinion based on fusion attention mechanism LSTM', *Computer Simulation*, Vol. 40, No. 1, pp.493–498.

Mukunthan, B. (2021) 'Efficient synergetic filtering in big data set using neural network technique', *International Journal of Computer Applications in Technology*, Vol. 65, No. 2, pp.134–139.

Song, Y., Lu, J., Lu, H. and Zhang, G. (2020) 'Fuzzy clustering-based adaptive regression for drifting data streams', *IEEE Transactions on Fuzzy Systems*, Vol. 28, No. 3, pp.544–557.

Xu, Y.Z., Xiao, J.Y., Yang, L. and Deng, C.L. (2020) 'Research on classification of network pseudo-public opinion based on K-means clustering in the environment of big data', *Journal of Xiangtan University (Natural Science Edition)*, Vol. 42, No. 6, pp.119–126.

Yang, H. (2020) 'Feature mining method of equipment support data based on attribute classification', *Ordnance Material Science and Engineering*, Vol. 43, No. 6, pp.124–128.

Yang, M., Liu, S., Chen, K., Zhang, H., Zhao, E. and Zhao, T. (2020) 'A hierarchical clustering approach to fuzzy semantic representation of rare words in neural machine translation', *IEEE Transactions on Fuzzy Systems*, Vol. 28, No. 5, pp.992–1002.

Zhang, Z.W., Liu, Z., Martin, A., Liu, Z.G. and Zhou, K. (2021) 'Dynamic evidential clustering algorithm', *Knowledge-Based Systems*, Vol. 213, No. 4, pp.1066430–1066435.

Zhao, L.M., Li, H.J. and Yi, J.Y. (2020) 'An analysis of data object traceability guarantee method in big data environment', *Information and Documentation Services*, Vol. 41, No. 2, pp.83–92.