



Cross-Dimension Mining Model of Public Opinion Data in Online Education Based on Fuzzy Association Rules

Li-xuan Li¹ · Ying Huo¹ · Jerry Chun-Wei Lin²

Accepted: 20 April 2021
© The Author(s) 2021

Abstract

The multi-dimensional characteristics of public opinion in online education lead to the difficulty of data cross-dimensional mining. To solve this problem, this paper designs a cross-dimensional data mining model of public opinion in online education based on fuzzy association rules. Based on the public opinion subject, object, and ontology to analyze the characteristics of public opinion in online education, Yaahp software is used to calculate the influence factor weight of public opinion in online education. According to the weight analysis results, the relationship between the dimensions of various public opinion data is clarified by using data semantic association. This paper introduces the fuzzy set theory into the database and uses crawlers to obtain public opinion data and stores them in the database, to complete the data preprocessing through distributed text preprocessing, feature selection distributed computing, and text vectorization distributed computing. Taking the cloud computing platform as the core, the cross-dimension mining model of public opinion in online education data is constructed according to the dimension correlation analysis and preprocessing results. The simulation results show that the model has the advantages of wide range, fast speed, and high accuracy, and can provide data support for education reform.

Keywords Fuzzy association rules · Network education · Public opinion data · Cross-dimension · Data mining · Yaahp software

1 Introduction

As the rapid growth of information techniques utilized in different domains and applications [1–6], public opinion in online education refers to the tendentious individual attitude and subjective will expressed by netizens around the occurrence, development, and change of a certain educational phenomenon (including ideas, events, figures, policies, problems, etc.) in the virtual space of the Internet. With the further development of mobile

Internet and social network technologies, the expression forms of education public opinion are increasingly diversified, especially the public opinion data has shown multi-dimensional characteristics. In recent years, public policy formulation, implementation, and policy performance evaluation, including education policy have paid more and more attention, and the network has become the most concentrated display platform of national discourse. Besides, with the deepening of the comprehensive reform of education, the education authorities also hope to get feedback from the people on the adjustment of education policies and reform measures, to provide the basis for the decision-making of education reform. However, current works on education public opinion analysis focus on qualitative analysis of educational events, such as the use of questionnaires and other forms, resulting in limited data sources, and the analysis angle and contents are also limited to the analysis of statistical results [7]. With the development of big data and data mining technologies, we can more in-depth clarify the law of education public opinion dissemination, and mine the deep-seated views of educational events. Thus, valuable information can be extracted from massive, noisy, and fuzzy data [8].

✉ Jerry Chun-Wei Lin
jerrylin@iee.org

Li-xuan Li
lilixuan751029@sgu.edu.cn

Ying Huo
huoying@sgu.edu.cn

¹ School of Literature and Media, Shaoguan University, Shaoguan, China

² Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway

Therefore, taking the cross-dimension data of public opinion in online education as the starting point, this paper makes an in-depth study on such data.

Several studies have been proposed for handling fuzzy association-rule mining [5, 10–13]. First of all, through the analysis of fuzzy theory, association rule mining, and fuzzy association rule mining, the discussion of fuzzy association rules are realized. On this basis, the massive data is fuzzified to establish a fuzzy database, and the fuzzy data in the fuzzy database are classified and aggregated, and the continuous attributes are discretized to determine the frequent association rules of the fuzzy database, to realize the massive data mining. Liang et al. [14] proposed a big data mining analysis method based on the generation of fuzzy association rules. Based on the fuzzy classification structure, the fuzzy association rules are defined and the corresponding algorithm is given. The fuzzy language modifiers are introduced into the fuzzy association rules to obtain more comprehensible and linguistic knowledge. The improved fuzzy association rules are used to realize data mining. Zhang et al. [15] proposed a quantitative data mining method based on improved multi-level fuzzy association rules. The method uses high-frequency item sets, forming as a top-down mining process by deepening iteration, integrating the fuzzy-set theory, data mining algorithm, and multi-level classification technologies to find fuzzy association rules from transaction data sets. Thus, the hidden knowledge of quantitative value information stored in the multi-level transaction database to meet the user's customized information mining needs is then produced. Wang et al. [16] proposed a comprehensive association rule mining method for health examination data based on the extended FP-Growth method. This method extends the FP-Growth algorithm for mining positive and negative frequent patterns, namely the PNFP-Growth framework. The extended method also uses a pruning strategy to filter out misleading patterns by associating negative data items with positive data items. The designed model is applicable for public datasets and since the database is composed of real health examination information of thousands of people, which is easier for rule maintenance.

However, the data mining process is complex and the cross-dimension data cannot be identified when the above methods are used to mine public opinion information in online education. According to the detailed characteristics of public opinion data in online education and its cross-dimensional nature, this paper uses crawler and data distributed computing to complete the public opinion data preprocessing. With the support of the cloud computing platform, this paper uses fuzzy association rules to build an accurate and effective data mining model, to provide help for the reform and development of education in China.

2 Cross-dimension mining model of public opinion data in online education

2.1 Characteristics of public opinion in online education

In theory, network public opinion refers to the relevant situations of network public opinion, and also refers to the people's evaluation of the value of things and an expression of personal attitude, which is a unique phenomenon in the development of human society. The dissemination of social public opinion often exists in the streets and alleys, which is a traditional appeal method inherited for thousands of years. It is very difficult to collect social public opinions, which is inefficient and prone to bias. The network public opinion analytic is a new model in the social network. It expresses the public's "happiness, anger, sadness and joy" of social problems through the network media. With the help of network virtual space, the public opinion information in the network can be obtained through network technologies, which has high efficiency and strong fidelity.

The forms of network public opinion are very rich, such as voice, text, picture, video, and so on. Network public opinion data can be transmitted through the community forum, microblog, personal blog, instant chat software group, etc., which is convenient for interaction, easy to express, and rich in topics. Traditional social public opinion is obviously inferior in these aspects. Taking education as the main consideration, this paper summarizes several characteristics of the current public opinion in online education as follows:

- **The discourse space of public opinion in online education is flexible to access:** At the present stage, the role of invisible "gatekeeper" in the network media is gradually disappearing, and the openness and interactivity are stronger. Different interest groups related to education policies can put forward their own opinions on a wide range of educational issues and carry out interactive exchanges [16]. There is critical rhetoric in these public opinions, in which some of them belong to subjective criticism, some of them belong to conformity criticism, and some of them belong to cathartic speech. These three kinds of public opinion components have important practical significance for policy revision and policy performance evaluation.
- **Teacher talk is the leading force of public opinion on education network:** There are some differences in the knowledge base, information quantity, experience, and interest of Internet public opinion speakers on different educational issues and policy practices. The sample data shows that teachers' discourse has become a strong discourse in the education network public opinion, and teachers can often express their views from the perspective of their own experience. Compared with the immature

thinking of students and the disadvantage that educational experts cannot go deep into the front line of education, teachers are not only the executors of educational policies at the grass-roots level but also the observers of the implementation performance of educational policies [17]. They are also the groups most involved in educational policies. Also, they have a very clear understanding of education issues. In the forum, teachers often put forward their practical wisdom. Compared with the traditional media discourse and special topic interviews, the leading force of public opinion on education networks is teachers, followed by parents, and finally related interest observers, students, and education researchers.

2.2 Research on influencing factors of public opinion on education network

The influence factors of network public opinion are mainly three-factor theory and extended factor theory based on three factors. The three factors are the subject, object, and noumenon of public opinion. Based on three factors, the expansion of the theory of three factors is realized by increasing public opinion space. The main elements of network public opinion are netizens (subject), public affairs (object), netizens’ opinions on public affairs (ontology), increasing network platform, communication interaction, and influence. Using the analysis of ecology and system theory, it is proved that the public opinion on the education network is a complex ecosystem composed of public opinion subject, object, and environment. The factors of public opinion on education networks are extracted into six factors, such as subject, object, ontology, network platform, communication interaction, and influence; the first level (B layer) and second level (C layer) factors are determined and shown in Fig. 1.

To obtain the influence weight of the first and second level factors, the paper states a pairwise comparison scale “questionnaire on the distribution of educational network public opinion evolution factors” to the experts, to realize the

collection of public opinion data in online education, carry out the first weight assignment through the analytic hierarchy process, and issue the “rationality screening table for the last level index of the evolution influence factors of the public opinion on education network” to the ordinary netizens. Through the comprehensive evaluation method, the first assignment is verified and corrected, and the final weight value is obtained. The experts who assign the factors need to have certain research results or research background in this field. Through the investigation and interview of the authors and experts on network public opinion, 30 teachers, including doctors and teachers of Communication Science in the United States, experts with research experience in the field of communication and public opinion in China, and teachers with relevant public opinion research experience in Party schools are selected, involving 10 units. 30 questionnaires are distributed and collected; thus, the effective rate is 100%. Collecting the data of expert assignment, Yaahp software is used to carry out AHP and consistency calculation and get the weight of each factor, as shown in Fig. 2.

2.3 Cross-dimensional correlation analysis

After defining the influencing factors of public opinion on education network, we can calculate the potential semantic association and sparse characteristics of public opinion data in online education, and obtain the dimensions between different education public opinion data, to realize the cross-dimension correlation analysis of data. In this paper, we need to calculate the correlation between the two kinds of potential data and shared data, to use these data as a bridge of knowledge transfer to realize the classification of test data. Because there is no co-occurrence relationship between training data and test data, similarity cannot be directly measured and analyzed [18]. However, if the two types of data are related, there may be a semantic correlation between them.

The Jensen-Shannon distance (JSD) between the distribution of shared data on the text and the distribution of potential data on text is used as the similarity measure between shared

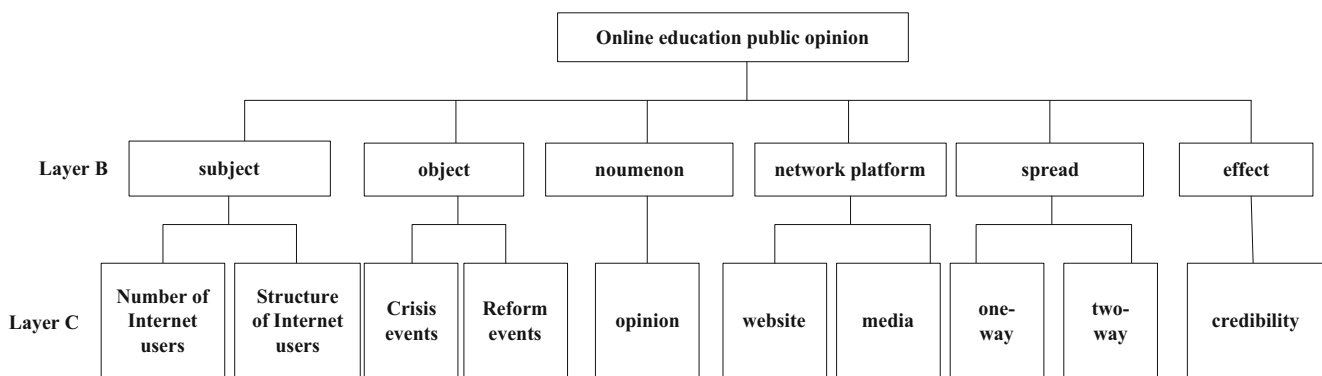


Fig. 1 Factors influencing public opinion on education network

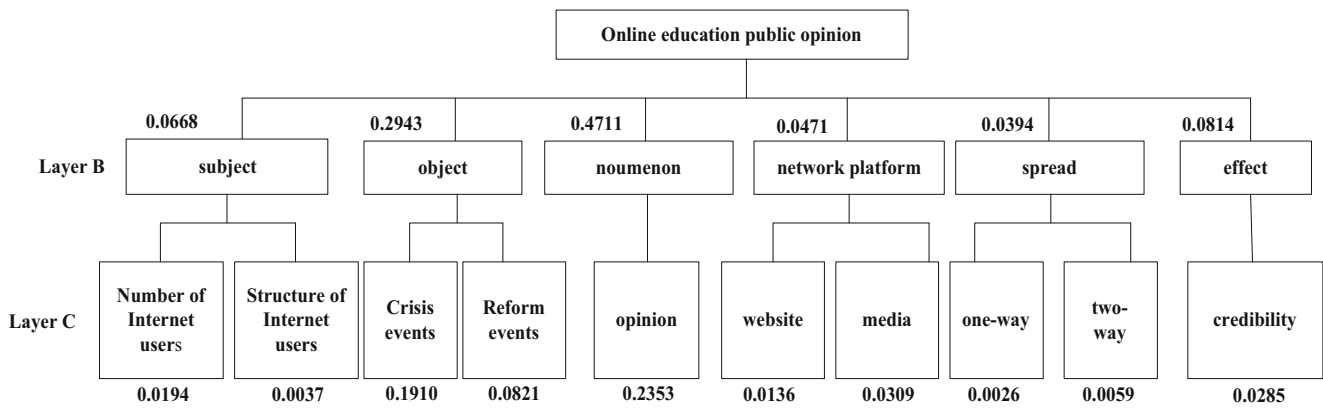


Fig. 2 Schematic diagram of factor weight

data and potential data. JSD is widely used to calculate the similarity between two probability distributions. Two discrete probability distributions of P and Q are defined as:

$$JSD(P||Q) = \frac{1}{2}(KL(P||M) + KL(Q||M)), \tag{1}$$

where M is the Kullback-Leibler distance between two text distributions. Taking Jensen-Shannon distance as the similarity measure, the similarity between the sharing data z_k and potential data z_i^t is described as follows:

$$S_{z_i^t, z_k} = \frac{1}{JSD(p(D|Z = z_k)||p(D|Z = z_i^t))} \tag{2}$$

After obtaining the correlation between the potential data and all the shared data, the relationship between the training data and the test data in the potential data is inferred according to the calculation results. This paper mainly uses the Pearson correlation coefficient (PCC) to judge the correlation. Pearson correlation coefficient is often used to weigh the linear correlation between two random variables [19], and the correlation between the training data z_i^s and test data z_j^t can be obtained, which is calculated as follows:

$$\rho(z_i^s, z_j^t) = \frac{\sum k(S_{z_i^s, z_k} - \bar{S}_{z_i^s})(S_{z_j^t, -\bar{S}_{z_j^t}})}{\sqrt{\sum k(S_{z_i^s, z_k} - \bar{S}_{z_i^s})^2} \sqrt{\sum k(S_{z_j^t, -\bar{S}_{z_j^t}})^2}}, \tag{3}$$

where, $\bar{S}_{z_i^s}, \bar{S}_{z_j^t}$ represents the average similarity between training data and test data and all shared data. According to the property of the Pearson correlation coefficient, if the value of ρ is positive, the training data and test data have different degrees of similarity with many shared data, while the calculation of ρ value is negative, which indicates that the training data and some shared data have high similarity, while the test data and these shared data have a low similarity. The text under the training data and test

data are represented as the feature vector of potential data, and the test data text d_j^t is represented as:

$$\phi(d_j^t) = [p(z_1|d_j^t) \dots p(z_k|d_j^t), p(z_1^t|d_j^t) \dots p(z_j^t|d_j^t)] \tag{4}$$

The data feature vector in formula (4) is mainly composed of two parts: one is the distribution of text on shared data, another is the distribution of text on test data. According to the measurement of the correlation between data, the training data set z_i^s with the highest correlation coefficient related to testing data z_j^t is used to replace z_j^t and transform the feature expression of test data in the potential data set. Through the transformation, the new characteristic of the data text d_j^t is tested and expressed as follows:

$$\varphi(\phi(d_j^t)) = [\phi(d_j^t)_{[1:k]}, \phi(d_j^s)_{[k+1:k+k^*]}] \tag{5}$$

Semantic concept expansion is to increase more semantic related concept sets. This process mainly involves semantic similarity calculation and semantic concept expansion between public opinion data concepts. After obtaining the anchor text c corresponding to each t in Wikipedia, many studies take the anchor text link or page content as extended conceptual semantics and background knowledge. The basis of these methods is character matching [20], and no more information related to anchor text semantics can be found. To expand more semantic related concepts and improve the accuracy of cross-dimensional data mining, we need to take an effective method to find more concepts similar to c semantics.

Non-negative matrix factorization (NMF) [21] is used to calculate the semantic similarity between public opinion data. Assuming that the initial decomposition matrix X is the concept document matrix of $m \times n$, m is the number of concept sets and n is the number of document sets, then the NMF algorithm is used to decompose two non-negative matrices W and H , where W is the data matrix of $m \times r$ and H is the

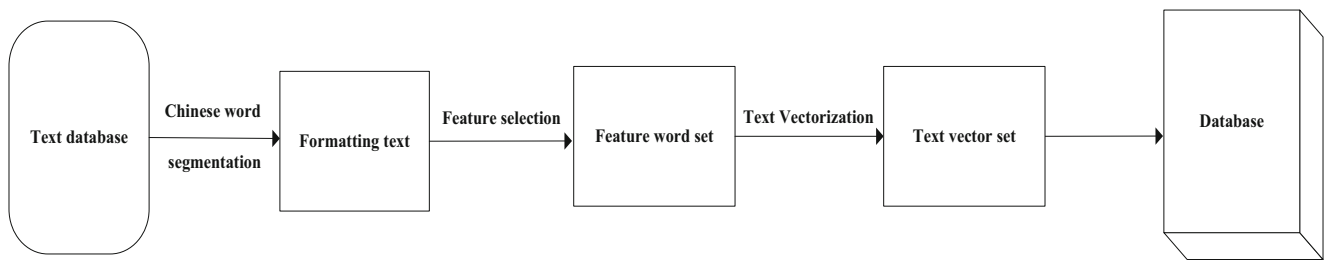


Fig. 3 Data preprocessing process

document matrix of $r \times n$, r is the column number of decomposition matrix W and the number of rows of H , representing the number of document set data. Different matrix decomposition methods have different application fields. In this paper, the NMF method is adopted. This method is based on the NMF method with sparsity constraints. In the iterative process of matrix decomposition, it can quickly find the nonnegative matrices W and H , to minimize the objective function. Finally, the cross-dimensional correlation analysis results of data are obtained, as shown in formula (6).

$$E(W, H) = \|X - WH\|^2 \tag{6}$$

Matrix decomposition is mainly divided into two stages. One is the projection process, that is, sparsity constraint is applied to the matrix generated in the process of decomposition iteration to find the optimal projection vector of each row or column of the matrix under the condition of invariant constraint; the other is the decomposition process, in each iteration process, the optimal solution obtained in the first stage is decomposed and iterated according to the nonnegative constraint and sparsity constraint until the stop condition is satisfied. In this way, the complete process of cross-dimension correlation analysis of public opinion data in online education is completed, which can provide convenience for the smooth development of subsequent mining work.

3 Construction of mining model based on fuzzy association rules

3.1 Analysis of fuzzy association rules

Association rule mining is one of the widely used research methods in the field of data mining. According to the types of variables to be processed by the association rules mining algorithm, they are divided into Boolean and numerical association rules. The data that the former needs to deal with are discrete, which can represent the relationship between different variables, while the latter is to process the numerical field, discretize, segment, or directly process the data. According to the abstract level of the data to be processed, it can be divided into single-layer and multi-layer association rules [22]. For the former, all variables are at the same level, without considering that the data, in reality, may have different levels; for the latter, multi-layer data is fully considered, that is, the rules contain multiple hierarchical attribute variables. If association rule mining is to deal with the data dimension, it can be divided into single dimension association rule and multi-dimension association rule. For the former, the data is one-dimensional, while for the latter, the data to be processed is multidimensional, that is, single-dimensional association rules are used to deal with the relationship between single attributes, and multidimensional association rules are used to deal with some relationships between attributes [23].

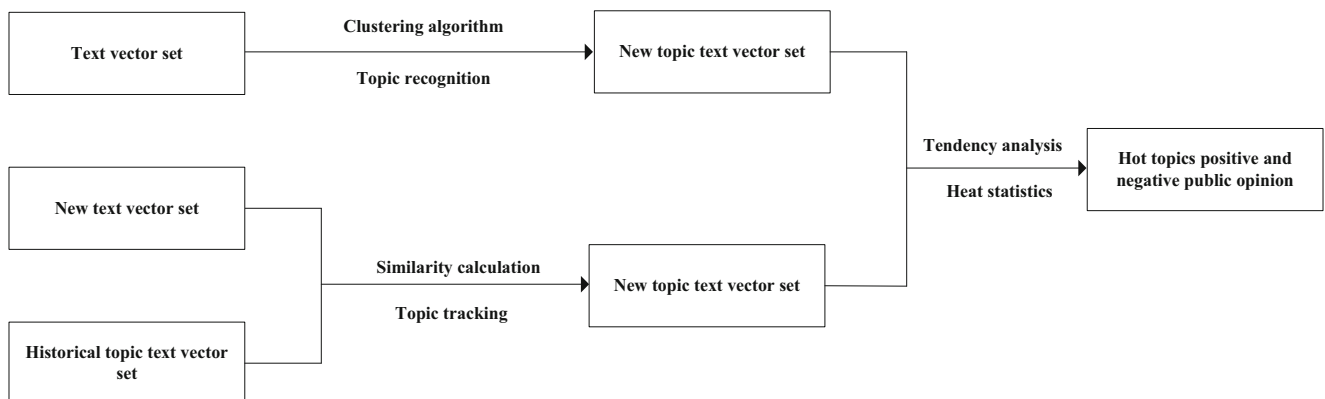


Fig. 4 Flowchart of public opinion analysis

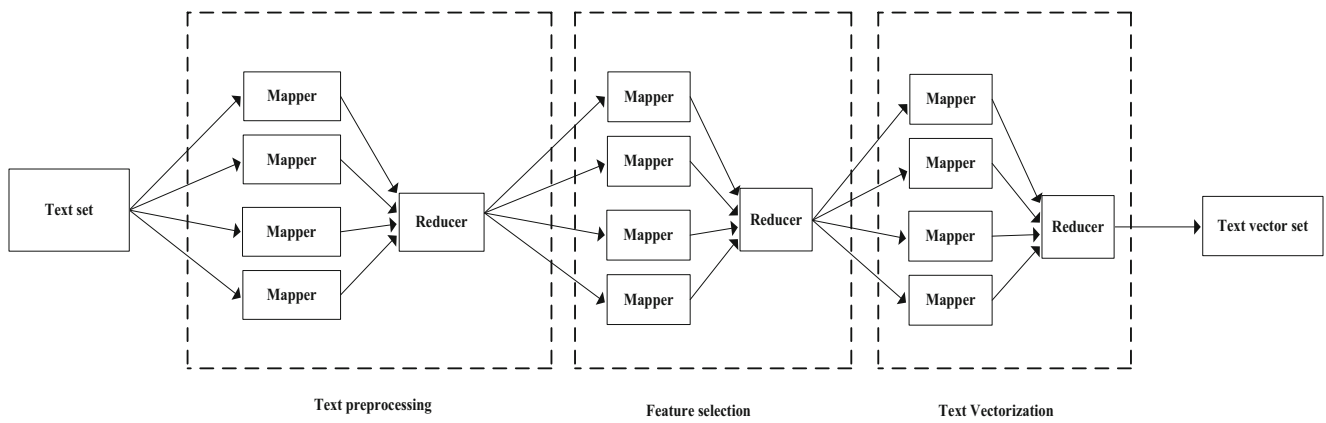


Fig. 5 Preprocessing job of sequential MapReduce text

According to the problem of data certainty in association rules, it is divided into deterministic association rules and fuzzy association rules. Due to the diversity and complexity of things in real life, it is difficult to use accurate concepts to describe many things. For example, a person’s height and body shape are fuzzy concepts, which are difficult to express with precise numbers. Therefore, fuzzy association rules appear, which makes fuzzy association rules more practical in some cases.

Fuzzy-set theory is based on the general set theory and extends the value range from [0, 1]. If there are elements ω and set B in the universe, then in the general set theory, the relationship between ω and B is as follows:

$$L(B, \omega, D) = \begin{cases} 1, \omega \in D \\ 0, \omega \notin D \end{cases} \quad (7)$$

In the theory of fuzzy sets, $f_x(\omega)$ is set to denote the membership degree of element ω to set B , and its value range is [0, 1], then the relationship between ω and B is as follows:

$$L(B, \omega, D) = f_x(\omega), f_x(\omega) \in [0, 1] \quad (8)$$

Fuzzy set divides the data set B into K classes, sets a cluster center in each class, and sets U to represent the fuzzy

membership matrix. Then the objective function of fuzzy theory can be expressed as follows:

$$J_s(U, k_1, \dots, k_K) = \sum_{i=1}^K \sum_{j=1}^m (u_{ij})^q d_{ij}^2, \quad (9)$$

where d_{ij} is the clustering between the cluster center of the i -th cluster interval and the j -th data, q is the incremental parameter of massive data, and q is the membership degree of the j -th data belonging to the i -th cluster center and it needs to satisfy:

$$\sum_{i=1}^K u_{ij} = 1 \quad (10)$$

According to the established objective function, it is necessary to ensure the minimum value of the objective function and complete the construction of fuzzy sets.

Association rule mining is mainly composed of antecedent X and consequent Y represented in the form as $X \rightarrow Y$ which has two parameters: support and confidence. If X occurs, Y will occur with confidence. Association rules show the implicit relationship between data, and association rule mining is to discover the implicit relationship between data [24]. Association rule mining can be divided into generating frequent itemsets and generating rules. Firstly, the massive

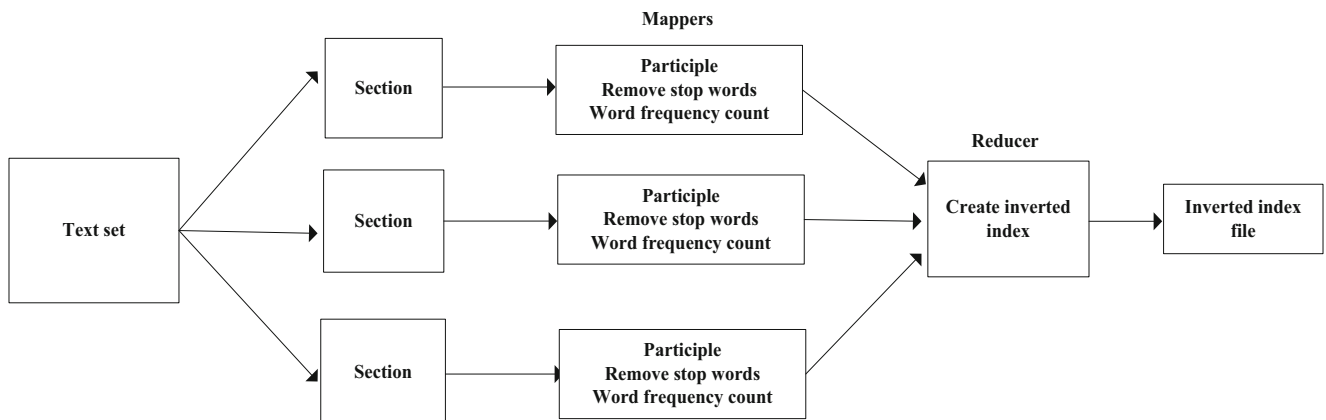


Fig. 6 Schematic diagram of text distributed preprocessing

Fig. 7 Flowchart of feature selection in Map stage

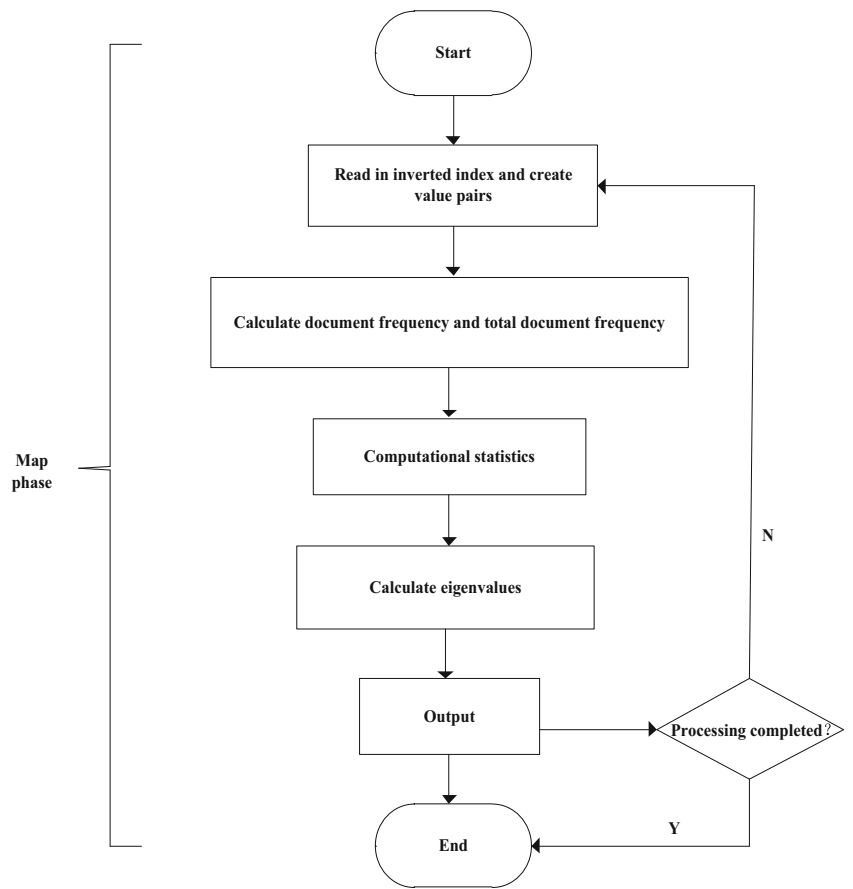


Fig. 8 Flowchart of feature selection in Reduce stage

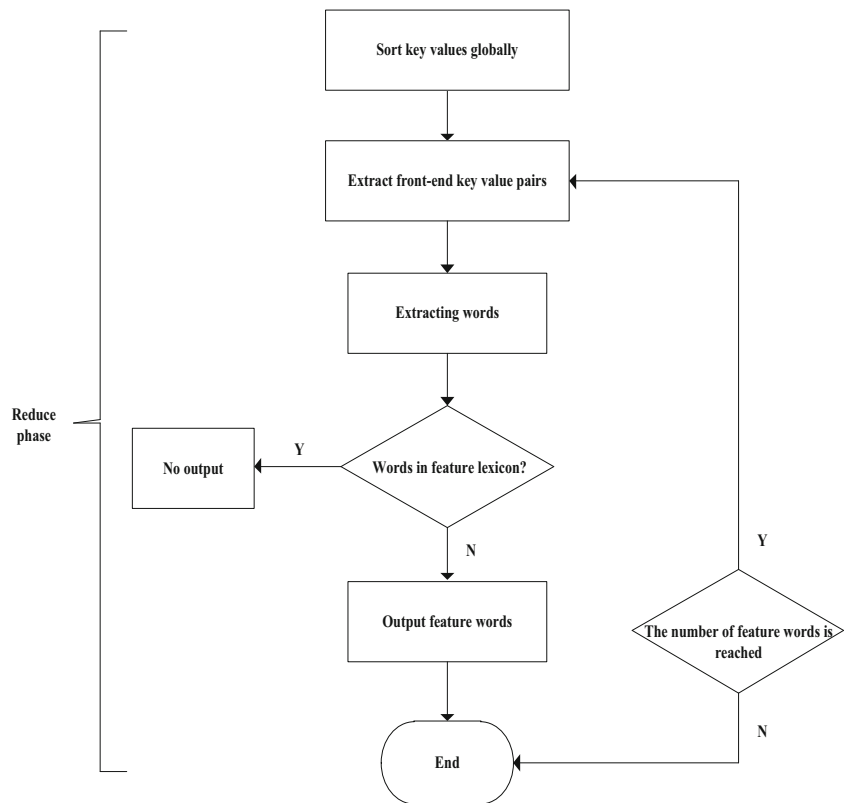
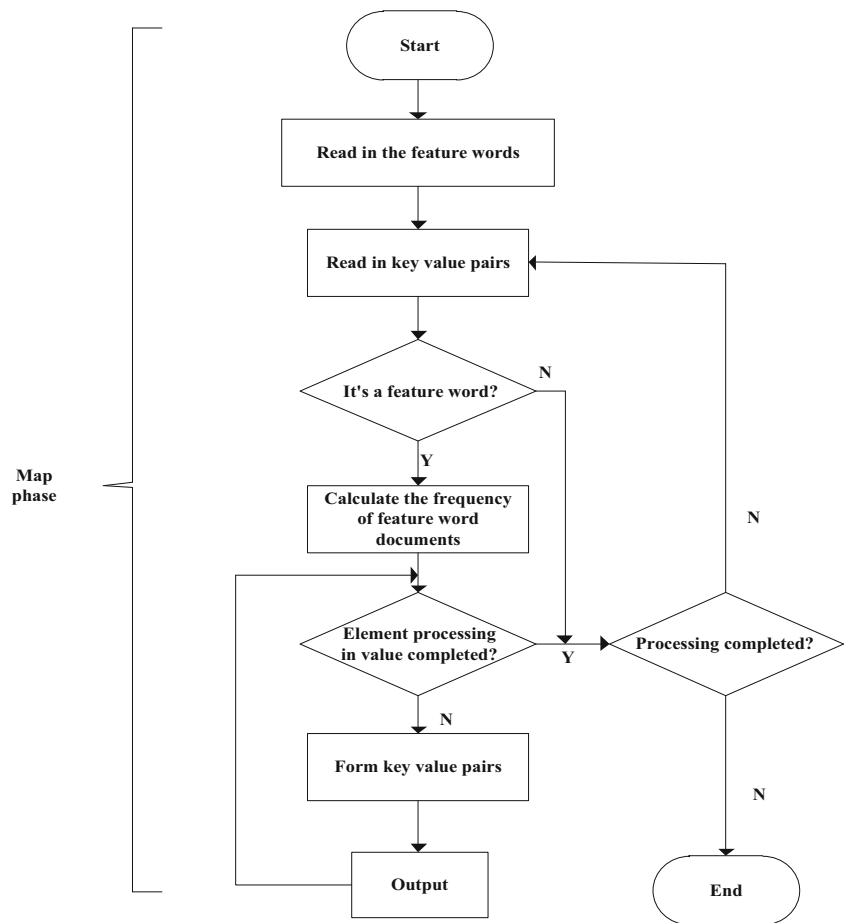


Fig. 9 Task flowchart of text distributed vectorization Map



database to be mined is selected, to realize the generation of frequent itemsets through parameter settings. On this basis, the association rules are generated, and the association rules are stored in the association rules database to complete the association rules mining.

To improve the effectiveness of data mining, based on general association rules, the quantitative data in a massive database M is transformed into Boolean data, and the Boolean database is obtained. The fuzzy set theory is introduced into the database, and the Boolean database is transformed into a fuzzy database M_f . Then the association rules in the fuzzy database are found by the Apriori-based classification algorithm [25]. Therefore, fuzzy association rules can be expressed as $X \leftrightarrow Y$ an implication. i_j is set as the element in the massive database set, the value range is Δ_j , and Δ_j is associated with the fuzzy set of massive data.

For any itemset X , for the i -th data in the fuzzy database, the support degree X is as follows:

$$D_{sup\tau_i}(X) = T(X_{1_i}, X_{2_i}, \dots, X_{\tau_i}), \tag{11}$$

where X_{j_i} denotes the fuzzy value of X_j in the i -th data; T is the generalized triangular norm of content data. Then the support degree of association rules can be expressed as follows:

$$D_{sup\tau_i}(X \leftrightarrow Y) = \frac{\sum_{i=1}^n D_{sup\tau_i}(X \cup Y)}{|D_f|} \tag{12}$$

The implication degree of association rules in the whole fuzzy database is as follows:

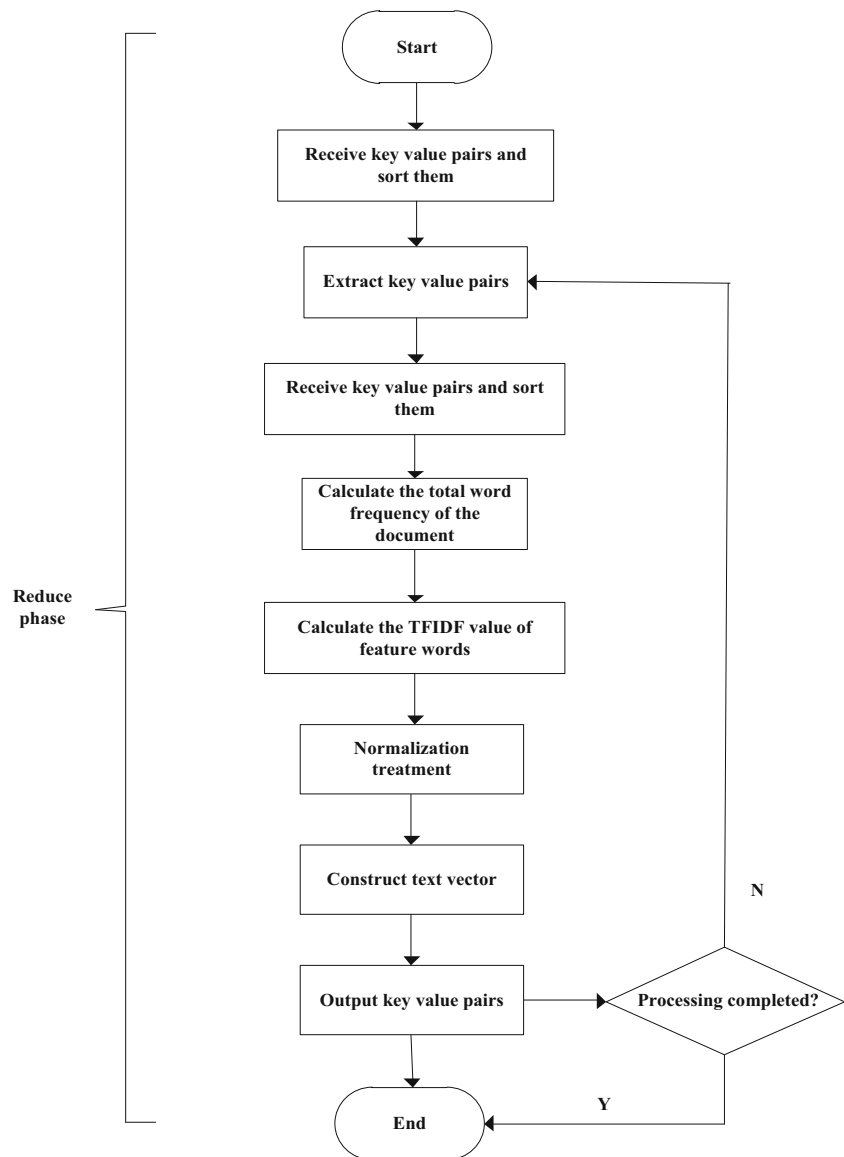
$$D_{im\tau_i}(X \leftrightarrow Y) = \frac{\sum_{i=1}^n D_{im\tau_i}(X \leftrightarrow Y)}{|D_f|} \tag{13}$$

Through the above discussion, the realization process of fuzzy association rules mining is clarified, which provides favorable conditions for the final realization of cross-dimension mining of public opinion data in online education.

3.2 Public opinion data acquisition and preprocessing

At present, the crawler is an important way to obtain Internet information actively. It is mainly divided into general web crawlers, theme web crawlers, and deep web crawlers. The general web crawler crawls according to a set of preset network addresses and parses other links in the page after obtaining the page for a follow-up visit; the theme web

Fig. 10 Task flowchart of text distributed vectorization Reduce



crawler mainly avoids a large number of irrelevant pages to save time and space, and the way to avoid irrelevant address is to analyze and judge the address; through relevant methods, the deep web crawler analyzes and classifies the addresses collected by crawling, and then accesses them [26].

The crawler is used to obtain public opinion data and store it in the database, which contains a variety of information. However, not all the information is related to the subsequent analysis and processing, such as the pictures and public opinion words on the page are not related. The public opinion information obtained by the above methods cannot be directly used for classification and analysis. The original public opinion information must be transformed into a form that can be understood and processed by the computer [27, 28]. In this paper, a text processing method based on statistical model segmentation is used to calculate the maximum probability of the sentence under the hypothesis of a specific word

segmentation result, and the probability value is taken as an important basis for the judgment standard of word segmentation.

The main task of public opinion information preprocessing is to de-duplicate and de-noising the collected web pages, get the formatted information, and construct the text vector. This process is the preparation stage of public opinion data analysis. The workflow is shown in Fig. 3.

Public opinion analysis mainly uses the HDFS distributed file system of the Hadoop framework for data storage and uses the Map/Reduce programming model to realize data processing, which can improve the parallel computing ability. The process mainly completes public opinion topic identification, public opinion topic tracking, and public opinion topic evaluation. The workflow is shown in Fig. 4.

Data preprocessing is divided into three steps: distributed text preprocessing, feature selection distributed computing,

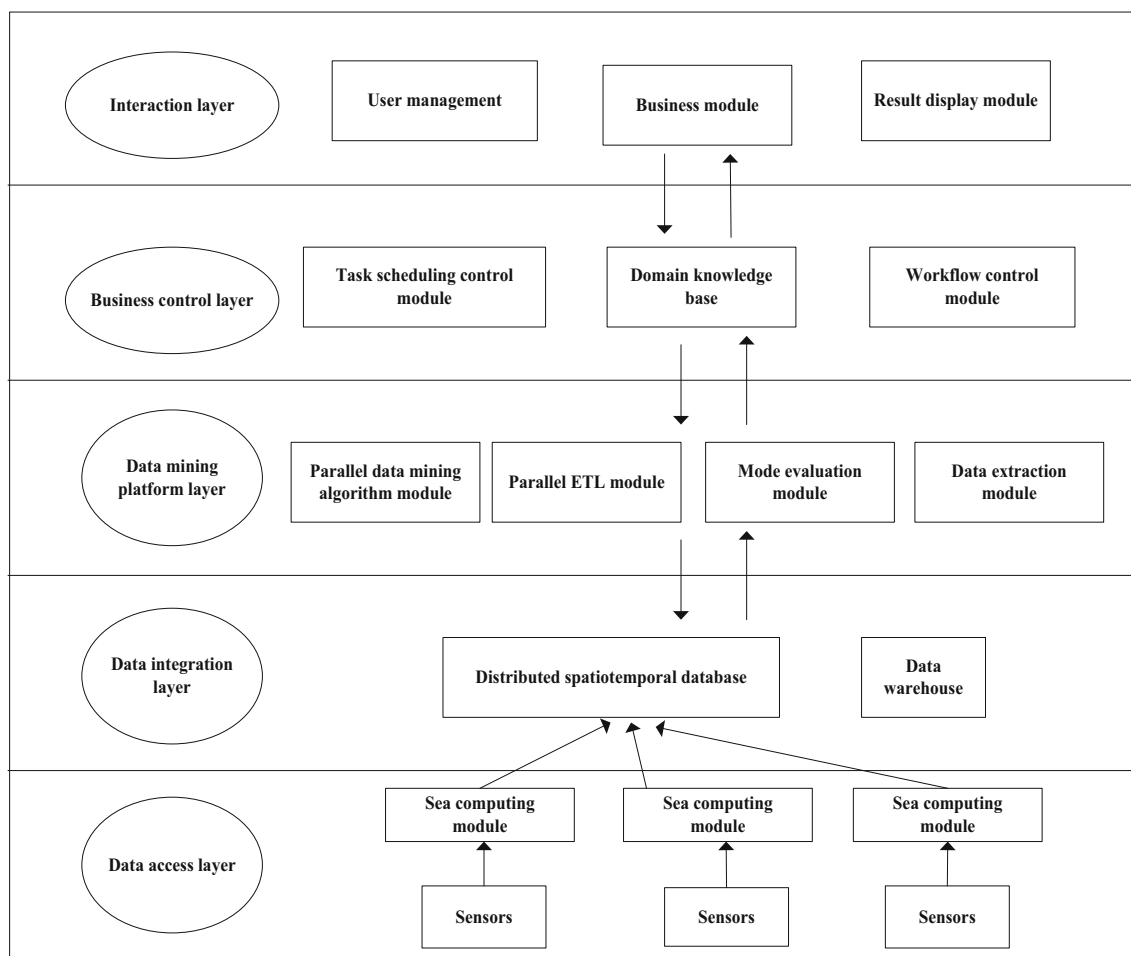


Fig. 11 Schematic diagram of model architecture

and text vectorization distributed computing. To store a large amount of text data on the Hadoop platform, HDFS distributed file is used and the MapReduce parallel model is used to optimize the text preprocessing process to realize distributed parallel preprocessing of large-scale text data. The key to this

process is how to reasonably decompose the whole job into multiple MapReduce tasks, and effectively decompose MapReduce calculation into two functions of Map and Reduce. In the process of numerical calculation, the degree of job parallelism can be maximized. Besides, information preprocessing is divided into three steps: text preprocessing, feature selection, and text vectorization. These three steps can be completed by MapReduce parallel computing model to complete distributed parallel computing. At the same time, the preprocessing process can also use the MapReduce parallel framework of a sequential combination to concatenate the whole process. The output content after the operation of MapReduce in the previous step is regarded as the input content of the next MapReduce process. The data to be processed can be converted into a text vector set after three steps of MapReduce. The implementation process is shown in Fig. 5.

3.2.1 Distributed text preprocessing

The distributed text preprocessing process includes four parts: Chinese word segmentation, stop word removal, word frequency statistics, inverted index generation. Using the

Table 1 Parameter values of the model in framework Weka

Parameter name	Parameter value
C	100,000
Class_weight	None
dual	FALSE
fit_intercept	TRUE
intercept_scaling	1
max_iter	100
multi_class	ovr
n_jobs	1
penalty	12
random_state	None
solver	liblinear
verbose	0

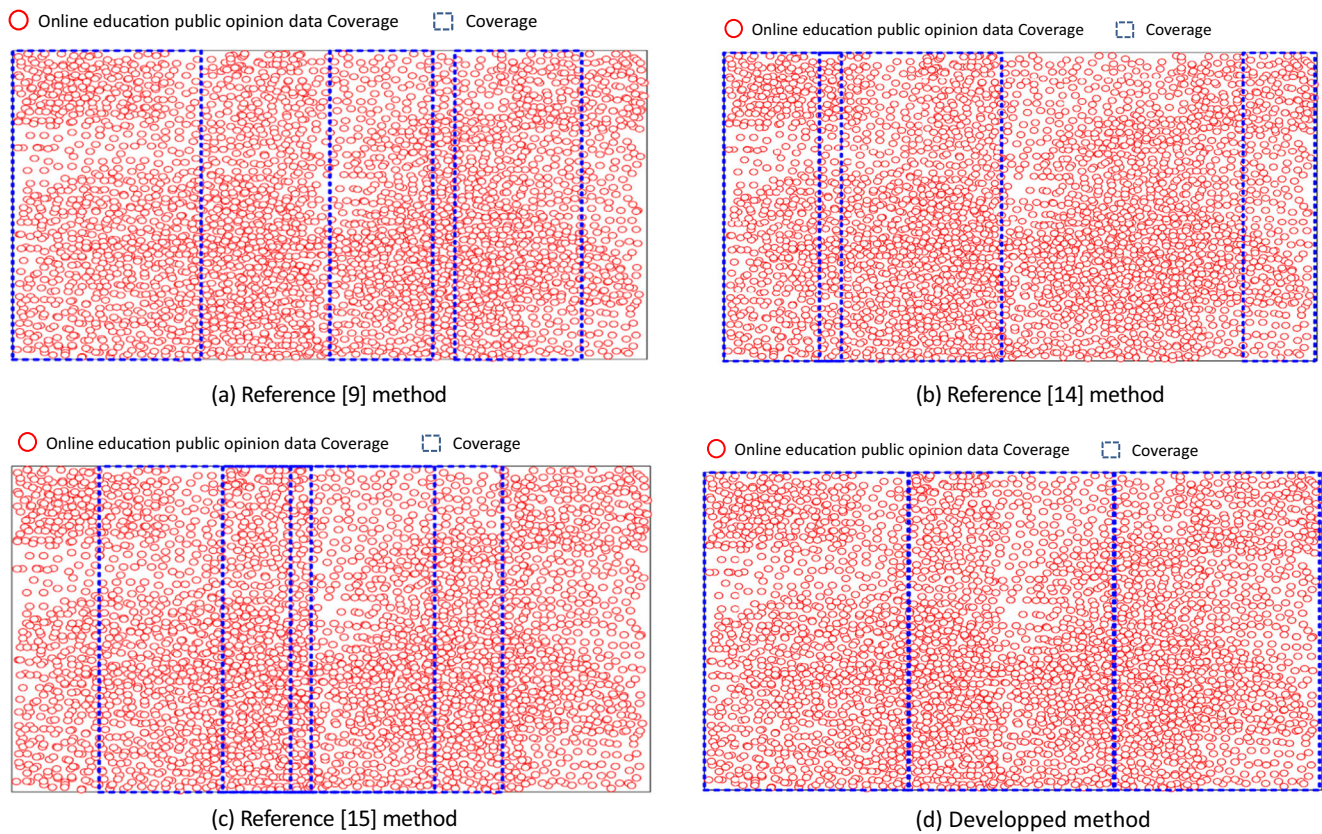


Fig. 12 Result comparisons (a) Reference [9] method (b) Reference [14] method (c) Reference [15] method (d) Developed algorithm

MapReduce framework to complete the whole task, in the process of constructing an inverted index, we need to calculate and count the word frequency of each word. We need to transfer a lot of fixed size message data which is divided into small pieces between the nodes set as DataNode. If it is set in the Map stage, the amount of data to be transmitted will be greatly increased in the process of system operation. Therefore, in this method, the process of constructing inverted index is arranged to be executed in the reduce stage, and the three steps of word segmentation, removal of stop words, and word frequency statistics, which require a relatively small amount of data transmission, are arranged in the map stage, to achieve the maximum extent of the whole preprocessing process and execution. The distributed preprocessing process is shown in Fig. 6.

Table 2 Experimental results of cross-dimension data mining of public opinion

Method	Accuracy (%)	Recall (%)	F-value
Designed algorithm	88	86	0.86
[9]	75	79	0.71
[14]	73	78	0.68
[15]	77	80	0.76

3.2.2 Distributed computing for feature selection

Feature selection is to reduce the dimension of an inverted index file after text preprocessing to get the feature word set of text set. Although the traditional text feature selection algorithm is widely used, its calculation process is relatively complex and time-consuming, which cannot meet the requirements of processing efficiency in the network data environment. Using the MapReduce parallel computing model, the whole calculation process is decomposed into two calculation processes: Map and Reduce, and distributed parallel design and processing are adopted for the calculation program. The content of the text is cross-dimensional, and the process of I/O and feature calculation is carried out in order. In the process of

Table 3 Cross-dimension mining results of public opinion data in online education of the proposed model

Databases	Accuracy (%)	Recall (%)	F-value
Mechanics	90	89	0.89
Biology	89	88	0.85
Software	88	88	0.86
Electronics	89	91	0.86
Water conservancy	90	90	0.88

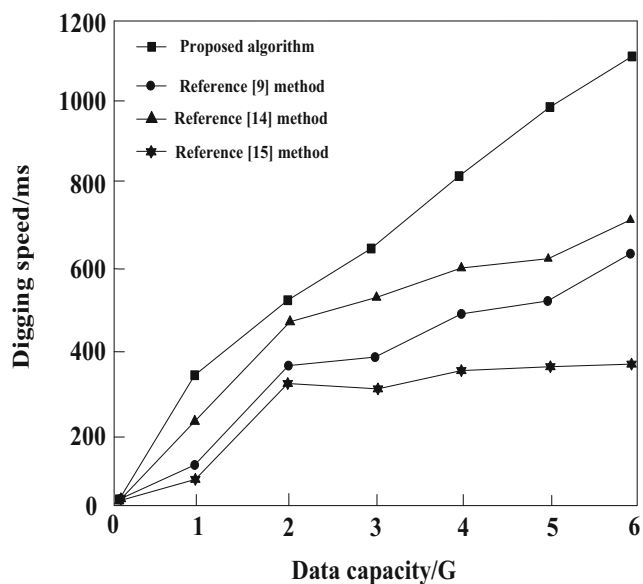


Fig. 13 Comparisons of execution efficiency

processing, the inverted index file output from the previous stage can be divided into separate small data blocks, and then stored on different nodes in the cluster.

Map function mainly realizes parallel reading of file blocks and calculation of eigenvalues. Through the good automatic sorting function of MapReduce, the obtained eigenvalues are arranged in a certain order. Finally, a Reduce function is designed to extract the specified number of feature words to complete the whole selection process. The two calculation processes of Map and Reduce are respectively shown in Figs. 7 and 8.

3.2.3 Distributed computing of text Vectorization

The method of TF-IDF weight calculation is used in text vectorization. To improve the calculation efficiency through

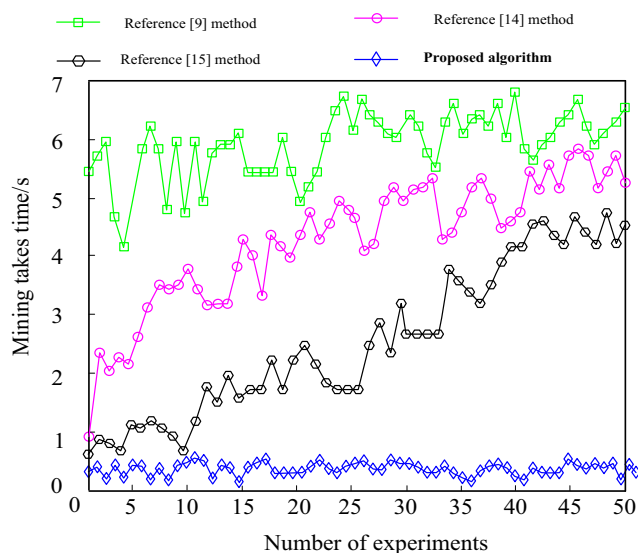


Fig. 14 Mining time-consuming comparison results

distributed computing, the TF-IDF weight calculation method is based on MapReduce optimization. In the map stage, each Mapper reads the text block in the index file by the bar as a unit, calculates the frequency of a feature word in the document through each record, and outputs the result in the form of standard key-value pairs. Then the MapReduce parallel framework compares the keys in the key-value pairs according to the set method, sorts the keys locally according to the comparison results, and finally sends the sorted results to the Reducer for further execution. In this process, according to the different document numbers, the TF-IDF values which represent the characteristics of feature words are processed in a standard dimensionless way, and the processed values are taken as the main factors of the text vector to complete the whole vectorization of the text file. Figures 9 and 10 show the detailed calculation process of Map and Reduce in-text distributed vectorization.

3.3 Implementation of data mining

In recent years, with the rapid development of the big data network industry, more and more application terminals are connected to the big data network. The development of cloud computing provides new solutions and solutions for network data mining [29, 30]. Based on fuzzy association rules, taking the cloud computing platform as the core, this paper constructs a cross-dimension mining model of public opinion data in online education according to the results of dimension association analysis and preprocessing, to lay a solid foundation for public opinion analysis in online education.

The cross-dimension mining model of public opinion data in online education based on cloud computing is divided into five layers: data access layer, data integration layer, data mining platform layer, business control layer, and interaction layer, as shown in Fig. 11.

4 Simulation experiments

To verify the validity of the cross-dimension mining model of public opinion data in online education based on fuzzy association rules, an experimental test is conducted. The experiments take the engineering education field as the research background, including five fields of Engineering Education such as machinery, biology, software, electronics, and water conservancy. Three data mining models for fuzzy association rule mining [9, 14, 15] are then compared with the developed approach to test the mining ability of four compared methods for public opinion in online education. The experimental network consists of 16,522 concepts and 94,231 relationships. After preprocessing the basic text, it maps to the network to form the knowledge concept model and finally transforms it into the required results. In the experiments, the machine learning

framework Weka¹ is used for data validation. The default parameters of the classifier in the framework Weka are used in the experiment, as shown in Table 1.

The public opinion data of five fields are divided into corresponding test set and training set, and eight experiments are conducted. Each time, 80% of public opinion data is selected as a training set and 20% public opinion data as a test set. Firstly, the cross-dimension mining range of public opinion data in online education of the four methods is compared, and the results are shown in Fig. 12.

From Fig. 12, it is obvious to see that the developed algorithm has a wider mining range and can obtain more complete cross-dimension mining results of public opinion data in online education. Besides, the experimental evaluation such as accuracy, recall rate, and F-value are then compared. The accuracy focuses on the ability to check the accuracy of the evaluation method; the recall rate focuses on the overall ability of the evaluation method; F-value combines the accuracy and recall rate to reflect the overall index of the method. The results are then described in Table 2.

As can be seen from Table 2, the accuracy, recall rate, and F-value of the proposed method are better than those of the three approaches [9, 14, 15]. The proposed method can improve the accuracy of mining results with the help of the relationship between knowledge and concepts in the network. The results also show that the designed algorithm can map the linear inseparable data in the original space to the semantic network to a certain extent, which enhances the linear separability and greatly improves the mining accuracy. The cross-latitude mining results of the data mining model based on fuzzy association rules for the above six kinds of public opinion data in online education are further tested, and the results are shown in Table 3.

It can be seen from Table 3 that the developed algorithm in this paper has excellent results in five fields. The main reason is that the domain knowledge of machinery and water conservancy is relatively mature and high-quality concepts; the relationships can be obtained through relevant news and encyclopedia knowledge, and there is less information intersection between machinery and water conservancy and other categories. In reality, there is more interdisciplinary cooperation in biology, software, and electronics, and the concepts contained in their information are not only limited to their fields but also involve other fields. Because this method can effectively extract cross-dimensional data semantics, it also has resulted in these three fields.

To verify the efficiency of the proposed method, the performance in terms of execution time is carried out on the Hadoop platform. The public opinion in online education documents is randomly selected from the SogouT corpus for experiments, and the public opinion documents are segmented. The size of the public opinion mining test database is 1 to 6GB. The results are shown in Fig. 13.

It can be seen from Fig. 13 that the execution efficiency of the proposed method is significantly higher than that of the three methods. When the data capacity is set to small, the speed difference of each method is not obvious. When the classification content is larger than 3GB, the mining time difference begins to increase significantly. When the data capacity reaches 6GB, the mining ability of this algorithm is optimal. This is because the proposed method reduces the moving and accessing time of classified data through HDFS distributed storage and MapReduce parallel processing, and improves the efficiency of data mining significantly. Based on the above experiments, the mining time of the four methods is compared, and the results are shown in Fig. 14.

The analysis of Fig. 14 shows that the mining time of the design model is always less than 0.5 s, which is far lower than other methods. It also indicates that the developed algorithm can quickly realize cross-dimension mining of public opinion data in online education.

5 Conclusion

To clarify the content of public opinion in online education and carry out education reform, a cross-dimension mining model of public opinion data in online education based on fuzzy association rules is designed. Based on the characteristics of public opinion and the influence factors of public opinion in online education, this paper analyzes the correlation between cross-dimension data of public opinion, collects data by using a crawler, and preprocesses the public opinion data; in the cloud computing environment, it uses fuzzy association rules to create data mining model to achieve efficient and accurate cross-dimension mining of public opinion data. Compared with other data mining methods, the proposed method is more innovative and practical and is more applicable to real cases. However, there are few communication variables considered in the study of public opinion influence factors, thus it is necessary to take this problem as the breakthrough point in future research to further improve the public opinion mining model of online education, which provides important support for education development.

Acknowledgements The paper was supported by 13th five year plan discipline co construction project of philosophy and Social Sciences in Guangdong Province with No.GD18XXW07.

Funding Open access funding provided by Western Norway University Of Applied Sciences.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were

¹ <https://www.cs.waikato.ac.nz/ml/weka/>

made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Liu S, Liu D, Srivastava G, Połap D, Woźniak M (2020) Overview and methods of correlation filter algorithms in object tracking. *Complex & Intelligent Systems*. <https://doi.org/10.1007/s40747-020-00161-4>
- Luo L, Yu L, Chen H, Liu Q, Wang X, Xu J, Heng P-A (2020) Deep mining external imperfect data for chest X-ray disease screening. *IEEE Trans Med Imaging* 39(11):3583–3594
- Lin JC-W, Yang L, Fournier-Viger P, Hong T-P (2019) Mining of skyline patterns by considering both frequent and utility constraints. *Eng Appl Artif Intell* 77:229–238
- Lin JC-W, Shao Y, Djenouri Y, Yun U (2021) ASRNN: a recurrent neural network with the attention model for sequence labeling. *Knowl-Based Syst* 212:106548
- Lin JC-W, Srivastava G, Zhang Y, Djenouri Y, Aloqaily M (2021) Privacy preserving multiobjective sanitization model in 6G IoT environments. *IEEE Internet Things J* 8(7):5340–5349
- Duan Y, Li J, Srivastava G, Yeh J-H (2020) Data storage security for the internet of things. *J Supercomput* 76(5):8529–8547
- Liu S, Wang S, Liu X, Lin C-T, Lv Z (2021) Fuzzy detection aided real-time and robust visual tracking under complex environments. *IEEE Trans Fuzzy Syst* 29(1):90–102
- Liu S, Liu X, Wang S, Muhammad K (2021) Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT assisted complex environment. *Neural Comput Applic* 33(4):1055–1065
- Mao X (2018) Research on massive data mining method based on fuzzy association rules. *Microelectronics & Computer* 35(02):89–93
- Lin C-W, Hong T-P, Lu W-H (2010) An efficient tree-based fuzzy data mining approach. *Int J Fuzzy Syst* 12(2):150–157
- Lin C-W, Hong T-P, Lu W-H (2010) Linguistic data mining with fuzzy FP-trees. *Expert Syst Appl* 37(6):4560–4567
- Lin C-W, Hong T-P (2013) A survey of fuzzy web mining. *WIREs Data Mining and Knowledge Discovery* 3(3):190–199
- Lin JC-W, Li T, Fournier-Viger P, Hong T-P (2015) A fast algorithm for mining fuzzy frequent itemsets. *J Intell Fuzzy Syst* 29(6):2373–2379
- Liang F, Zhao L (2019) Big data mining analysis algorithm based on fuzzy association rules generation. *Comput Appl Softw* 36(10):285–292
- Zhang D, Zhang Y (2019) Quantitative data mining algorithm based on improved multilevel fuzzy association rules. *Comput Appl Res* 36(12):3619–3622
- Wang B, Chen D, Shi B, Zhang J, Duan Y, Chen J, Hu R (2017) Comprehensive association rules mining of health examination data with an extended FP-growth method. *Mobile Netw Appl* 22(2):267–274
- Dou C, Zheng Y, Dong Y, Zhang Z, Ma K (2018) Hybrid model for renewable energy and loads prediction based on data mining and variational mode decomposition. *IET Gener Transm Distrib* 12(11):2642–2649
- Baumann TS (2018) Appraisal of geodynamic inversion results: a data mining approach. *Geophys J Int* 207(2):667–679
- Zhu X, Wu G, Coulon F, Wu L, Chen D (2018) Correlating asphaltene dimerization with its molecular structure by potential of mean force calculation and data mining. *Energy Fuel* 32(5):5779–5788
- Wu X, Zhang X (2019) An efficient pixel clustering-based method for mining spatial sequential patterns from serial remote sensing images. *Comput Geosci* 124(5):128–139
- Yu W (2019) Discovering frequent movement paths from taxi trajectory data using spatially embedded networks and association rules. *IEEE Trans Intell Transp Syst* 20(3):855–866
- Kupka J, Rusnok P (2019) Fuzzy four-fold tables: their properties and use in fuzzy association analysis. *Int J Approx Reason* 108(5):89–106
- Park SH, Synn J, Kwon OH, Sung Y (2018) Apriori-based text mining method for the advancement of the transportation management plan in expressway work zones. *J Supercomput* 74(3):1283–1298
- Zhou G, Zhou G-Y (2018) Modeling and mining domain shared knowledge for sentiment analysis. *ACM Trans Inf Syst* 36(2):1–18
- Segatori A, Bechini A, Ducange P, Marcelloni F (2018) A distributed fuzzy associative classifier for big data. *IEEE Trans Cybern* 48(9):2656–2669
- Wu D-Y, Zhang X-Y, Zhou X-L (2018) Mining and correlation analysis of association rules between properties and therapeutic efficacy of Chinese Materia medica based on strategy pattern. *Chin Med J* 131(22):2755–2757
- Eide A (2018) Substituting model-based indicators in harvest control rules by observations using fuzzy logic methodology. *ICES J Mar Sci* 75(3):977–987
- Liu S, Guo C, Al-Turjman F, Muhammad K, de Albuquerque VHC (2020) Reliability of response region: A novel mechanism in visual tracking by edge computing for IIoT environments. *Mech Syst Signal Process* 138:106537
- Letrache K, El Beggar O, Ramdani M (2019) OLAP cube partitioning based on association rules method. *Appl Intell* 49(2):420–434
- Liu S, Pan Z, Cheng X (2017) A novel fast fractal image compression method based on distance clustering in high dimensional sphere surface. *Fractals* 25(4):1740004 2017

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.