Hindawi
Journal of Sensors
Volume 2022, Article ID 6492566, 13 pages
https://doi.org/10.1155/2022/6492566

Research Article

Text Data Processing and Classification Algorithm Based on Data Fusion and Granular Computing

Duo Ji[1] **and Peng Zhang**[2]

[1]*College of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang, 110854 Liaoning, China*
[2]*Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China*

Correspondence should be addressed to Duo Ji; jiduo@cipuc.edu.cn

Received 19 November 2021; Revised 29 December 2021; Accepted 3 January 2022; Published 24 January 2022

Academic Editor: Gengxin Sun

With the rapid expansion of network information and the emergence of a large number of electronic texts, how to organize and manage this massive information has become a major challenge. Automatic text categorization technology is to study how to let the machine classify unknown text through self-learning, thus solving the difficulties encountered in manual classification. Because granular computing can reduce the knowledge in solving complex problems, it is more convenient to summarize and acquire knowledge. It has become a hotspot in recent years, and it also provides new ideas for text classification research. The rough set model of granular computing can acquire knowledge by mining decision rules. The decision process is more transparent and easy to understand. It has been paid attention to and applied in text classification research. Based on the research of existing achievements, this paper makes a further study on the application of granular computing in text categorization. After analyzing the existing feature selection methods, the feature distribution is proposed based on the relationship between feature words and categories. By calculating the distribution distance between any two feature words, the feature words with similar distribution distances are aggregated, which effectively reduces the dimension of the feature space and also avoids the individual samples caused by the existing feature selection algorithm. A phenomenon that is discarded due to features. The experimental results show that the clustering method can obtain higher classification accuracy than other feature selection methods when using SVM as the classifier. SVM performs best, and the final text classification accuracy rate can reach 85.46%. According to the correlation principle of the rough set, feature selection is made for each information granularity, the selected feature is used as the condition attribute and the coordination matrix is constructed, and the most similar sample is heuristically searched to obtain the attribute reduction set.

1. Introduction

The rapid development of information technology, especially the development of the Internet, has brought people into the era of information exchange. The Internet provides a platform for people to exchange and share information and has become an indispensable part of modern life tools and work tools. In February 2019, the China Internet Network Center (CNNIC) released the "43th Statistical Report on Internet Development in China," showing that as of December 2018, the number of Internet users in China was 829 million, and the number of new Internet users was 56.53 million. The penetration rate reached 59.6%, an increase of 3.8% from the end of 2017. The number of mobile Internet users in China reached 871 million, and the number of mobile Internet users increased by 64.33 million. The proportion of Internet users using mobile phones increased from 97.5% at the end of 2017 to 98.6% at the end of 2018. Mobile Internet access has become one of the most commonly used Internet channels. With the continuous increase in the number of Internet users and the continuous growth of online information, people have encountered the problem of massive information such as retrieval and management brought about by information expansion.

How to effectively organize and manage this information has become an area facing the information science. With the continuous development of technology, text classification has gradually changed from a knowledge-based method to a method based on statistics and machine learning.

Most of the information on the Internet appears in the form of text or can be converted into text. Therefore, as a key technology for processing and organizing large amounts of text data, text categorization has become information filtering, information security, mail classification, information retrieval, and search. Basic technologies in the fields of engines, web forums, digital libraries, etc. and many research teams at home and abroad have conducted in-depth research on text classification algorithms. Tian [1] proposed that in text categorization, the performance of the classifier decreases as the feature dimension increases. The main purpose of feature selection is to remove irrelevant and redundant features in the function and to reduce the functional dimensions. Based on the word vector generated by Word2-Vec, the Word2Vec-SM algorithm is proposed to reduce the dimension of the feature. Bei [2] proposed an improved tf-idf-miow algorithm based on the traditional tf-idf algorithm and mutual information algorithm to meet the requirements of marine big data text classification. The results of automatic text classification experiments show that the tf-idf-miow recall rate in the oceanography field is 10.33% higher than the traditional tf-idf algorithm, and the f1 score is increased by 6.92%. Ni et al. [3] studied in detail the influence of parameters on classification accuracy when using support vector machine (SVM) and K-nearest neighbor (KNN) text automatic classification algorithm. The advantages and disadvantages of the two text classification algorithms are presented in the field of petrochemical processes. Chen [4] proposed a new classification model LDA- (Latent Dirichlet Allocation-) KNN (K-nearest neighbor). LDA is used to solve the problem of semantic similarity measurement in traditional text categorization. The sample space is modeled and selected by this model. Lianhong et al. [5] proposed a short text semantic extended representation method based on concept map. Firstly, the degree of association between the text feature words and the concepts in the concept map is calculated, and the concept with high degree of relevance is selected to form a conceptual dictionary of the current text. Then, the concept dictionary is added to the feature word set to obtain a semantic extended representation of the short text. Weiyin and Li [6] proposed a text classification model CNN-XGB based on convolutional neural network and XGBoost. Firstly, Word2Vec is used to represent the preprocessed data, followed by multi-scale convolution kernel convolutional neural network for data feature extraction. Finally, XGBoost is used to classify the features of deep extraction. Man et al. [7] fully validated the model by using multiple indicators to evaluate the model in the test data set. Compared with other models, the proposed model has better classification performance in the two-class and multiclassification tasks. Wang et al. [8] proposed a new feature word extraction algorithm based on chi-square statistics by extracting the feature words of text method and evaluated the text classification model through the improved new method. The experimental results show that the new method is significantly better than the traditional feature extraction methods in the evaluation results such as precision, recall, F1, and ROC_AUC. Yao et al. systematically study the web/text classification problem by combining sparse representation with random measurement. First, a very sparse data measurement matrix is used to map the original high-dimensional text feature space to a low-dimensional space without losing key information. Then, a general sparse representation method is proposed, which obtains the sparse solution by decoding the semantic correlation between the query text and the entire training sample. The authors conducted a large number of experiments using real-world data sets to check the proposed method, and the results showed the effectiveness of the proposed method [9]. Wang et al. [10] to improve the text classification effect and introduce the deep neural network isomorphic with BP neural network to initialize the initial weight of BP neural network. Experiment on multiple data sets it shows that this text method obviously improves the accuracy of text classification. Chaolei and Junhua [11] conducted experiments on the same data set. The results show that simulated annealing has stable global search performance and is an effective way to optimize SVM parameters. Chao and Junhua [12] show through experiments that compared with the traditional KNN algorithm, the improved algorithm has improved accuracy, recall rate, and F value. Compared with other classification algorithms, it has certain advantages. Junhong et al.'s [13] simulation experiments show that the proposed method can effectively solve the problem that the incompletely labeled text classifier can not effectively identify the boundary between the incompletely labeled text category and other categories under the current classification system, resulting in low data classification performance. Kai [14] introduced the process of text categorization and an overview of the three classifiers. Finally, the three classifiers were tested separately, and the experimental results were analyzed to find out the classification effect of the support vector machine classifier in the experimental environment, better than the other two classifiers. J. Ma and Y. Ma [15] showed that the method is superior to the commonly used long- and short-term memory models, multicategory logistic regression, and support vector machines in terms of accuracy and recall rate.

Although many research teams at home and abroad have their own research plans for text categorization methods, they all have some shortcomings: single method, low efficiency, and complicated calculation process. The granularity calculation method just has the advantages that these methods do not have. The calculation process is simple and clear, the operation is simple, the feature recognition efficiency is high, and subtle changes can be detected. Many research teams have seized the opportunity and made extensive calculations on the granularity.

The method of granular computing can reduce the dimension of knowledge when solving complex problems, which makes it easier to generalize and acquire knowledge. It has become a hotspot in recent years, and it also provides new ideas for the study of text classification. Xingguo et al.

[16] applied the granularity calculation to the vehicle identification. Aiming at the problem that the classification of fine-grained vehicle identification images has low recognition rate due to redundant features, a fine-grained vehicle identification algorithm based on singular value decomposition and central metric is proposed. The research shows that the method uses the Residual Network (ResNet) framework to test on the Cars-196 fine-grained model data set, and the accuracy rate can reach 93.02%, which is better than the current bilinear and attention model. Extended experiments prove that this method is equally applicable to other network frameworks. Haoru et al. [17] used the granular algorithm to screen out images with great influence on the recognition results to prevent overfitting; input the filtered images into the RPN network improved by soft-nms (Soft Nonmaximum Suppression) to obtain object-level image annotation. Dangwei et al. [18] applied the granularity algorithm to search for isolated regions. Aiming at the shortcomings of traditional particle swarm optimization algorithm for searching isolated regions and low search accuracy, a subgroup hierarchical coarse-grained particle swarm optimization algorithm was proposed. On the basis of the coarse-grained model, the subgroup is divided into several common subgroups, adaptive subgroups, and elite subgroups. Different subgroups adopt different evolution strategies in the evolution process. Jingrui and Dongyang [19] completed the multigranularity search of the initial layer data source through statistical expectation calculation, imported the initial layer probability calculation result into the multigranularity variable distribution calculation, and completed the multigranularity search of the middle layer data source. In order to ensure the effectiveness of the proposed method, the proposed method is compared with two traditional methods, and the efficiency is obviously improved and has high efficiency. The experimental results of Jinshuo et al. [20] and others show that the multithreading strategy based on CPU can achieve a 4x speedup ratio, and the parallel algorithm based on the unified computing device architecture (CUDA) can achieve a maximum speed up of 34 times and the proposed strategy. Based on the CUDA parallel strategy, it achieves a 30% performance improvement and can be used to quickly schedule computing resources in other areas of big data processing. Suzhi et al. [21] introduced the granularity idea to divide the initial data set into multiple subsets. Secondly, the improved similarity matrix was calculated for each subset combined with intraclass and interclass distance. Finally, the improved parallel AP aggregation was implemented based on MapReduce model. Experiments on real data sets show that the IOCAP algorithm has better adaptability on large data sets and can effectively improve the accuracy of the algorithm while maintaining the AP clustering effect. Ronghu and Yunjie [22] in order to improve coarse-grained parallel inheritance algorithm performance, shortening the solution time for the stereo warehouse path optimization problem, applying a single program multiple data stream (SPMD) parallel structure to the coarse-grained parallel genetic algorithm, and improving the algorithm. Yingjian et al. [23] use the granular algorithm to divide the circuit into multiple regions and use the logical fingerprint feature as the identifier of the region. By comparing the multivariant logical fingerprints of the partition in two dimensions of time and space, the hardware Trojan detection without gold chip is realized and diagnosed. Jin and Jianhua [24] designed a fine-grained remote attribute proof algorithm to solve the problem of large-scale remote identification of traditional attributes. For different remote proof requirements, the attribute was remotely proved and the terminal platform was more detailed. Yilin et al. [25] proposed the theory of the degree of weighted granularity superiority relationship pessimistic multigranularity rough set and weighted granularity dominant degree optimistic multigranular rough set. On this basis, a dynamic parallel updating algorithm based on the degree-weighted rough set approximation set of weighted granularity and dominant relationship is presented.

In order to solve the problem of complex data, difficult operation, cumbersome recognition process, and incomplete feature extraction in a text categorization method, this paper studies the text classification method based on granular algorithm. Based on the existing results, the application of granular computing in text categorization is further studied. The existing feature selection methods are analyzed. According to the relationship between feature words and categories, the feature distribution distance is proposed. The distribution distance is similar. Feature words are aggregated, which effectively reduces the dimension of the feature space and avoids the phenomenon that individual samples caused by the existing feature selection algorithm are discarded because they do not contain the selected features; the clustering method can be obtained when using SVM as the classifier has higher classification accuracy than other feature selection methods.

2. Method

2.1. Data Fusion

2.1.1. Introduction to Data Fusion. Data fusion in text data processing is to carry out multilevel comprehensive optimization and intelligent analysis through certain rules and finally complete the needs of users, that is, a process of obtaining more accurate description of perceptual object fusion information. In text data processing, the perceptual information obtained by data fusion technology is usually more persuasive than the data collected and analyzed by a node. The data fusion center fuses information from multiple sensors; it can also fuse information from multiple sensors and the observation facts of the human-machine interface (this fusion is usually a decision-level fusion). Extract the symptom information, under the action of the inference engine. Match the symptoms with the knowledge in the knowledge base, make fault diagnosis decisions, and provide them to users. Data fusion can ensure the accuracy of perceived data, reduce the network data traffic in processing, reduce the redundant data in the network, and play an important role in making reasonable decisions for applications.

2.1.2. Hierarchy of Data Fusion. According to the level of data abstraction in the fusion system, the fusion can be divided into three levels: data level fusion, decision level fusion, and feature level fusion.

(1) Data Level Fusion. The information processing of data level fusion is shown in Figure 1. Data level fusion is the fusion directly on the original data layer. The data is synthesized and analyzed before various sensors are preprocessed. Because the sensor detects the same feature data in the same environment, different types of feature data cannot be fused. The advantage of data level fusion is that it can maintain the complete amount of information of data without data preprocessing. However, it also has the disadvantages of large traffic, long analysis time, poor anti-interference ability, and poor real-time performance. In order to solve this problem, efforts should be made to develop a fusion algorithm model that has both robustness and accuracy. Focus on research on related processing, fusion processing, and system simulation algorithms and models, and conduct research on evaluation techniques and metrics for data fusion systems.

(2) Feature Level Fusion. Feature level fusion belongs to the middle level fusion. It first extracts the features of the original information from each sensor (the features can be the edge, direction and speed of the target) and then classifies, collects, and synthesizes the multisensor data according to the feature information. The information processing process is similar to the feature level fusion process. Because the extracted feature information is generally directly related to decision analysis, the fusion result can assist decision analysis to the greatest extent. Feature level fusion has low requirements for communication bandwidth; as long as the broadband reaches above 2 MHz, normal operation can be guaranteed. It realizes considerable information compression and is conducive to real-time processing, but its accuracy is reduced due to data loss. At present, many methods have been applied to feature level fusion. Common methods include image fusion, data compounding, information compounding, data compounding, and image compounding.

(3) Decision Level Fusion. Decision level fusion is the highest level of data fusion. Firstly, after each sensor preprocesses the original data and makes a decision, it fuses their decision results to make the final decision results consistent as a whole. The information processing process of decision level fusion is shown in Figure 2. The advantage of decision level fusion is that it has good real-time and fault tolerance, less dependence on sensors, and less traffic. It can still work even when one or several sensors fail. However, because decision level fusion needs to preprocess the original data obtained by sensors to make their own decision results, the preprocessing cost is high.

Because different levels of fusion algorithms have different advantages, disadvantages, and scope of application. For choosing the fusion algorithm at which fusion level, it is necessary to comprehensively consider the sensor performance, computing power, communication bandwidth, detection parameters, and capital budget of each system. There is no universal structure that can be applied to all application backgrounds. In practical applications, different levels of fusion algorithms often appear in one system at the same time. The characteristics of three different levels of fusion methods are compared as follows.

2.1.3. Data Fusion Classification. In the existing data fusion applications, data fusion technology can be divided into different categories according to different standards. Data fusion technology is divided into data layer fusion, feature layer fusion, and decision layer fusion. See Figure 3 for details.

2.1.4. Common Data Fusion Methods. With the rapid development of information theory, artificial intelligence, target recognition and other fields, more and more data fusion algorithms also appear. At present, data fusion algorithms can be divided into three categories: methods based on physical model, methods based on parameter classification, and methods based on cognitive recognition model. There are many common data fusion methods in text data processing. The typical ones are neural network, fuzzy theory, D-S evidential reasoning, and principal component analysis. This paper chooses the method of parameter classification to solve the problem of massive information retrieval and management caused by information expansion.

We choose the data fusion algorithm of principal component analysis. Principal component analysis (PCA) is a simple and effective data compression algorithm, which is very consistent with the characteristics of text data processing. Principal component analysis is to project the perceptual data onto a new coordinate axis and calculate its eigenvector, so that the eigenvector corresponding to the largest eigenvalue becomes the first coordinate vector (called the first principal component); the eigenvector corresponding to the second largest eigenvalue becomes the second coordinate vector (called the second principal component), and so on. In this way, its main components are retained, which not only ensures the main characteristics of the data but also reduces the amount of data transmission. The data fusion algorithm based on principal component analysis divides the text data into multiple clusters. The cluster head will collect the information of its cluster members and then put the data into the observation matrix, which can be projected into a new space. The nodes of the cluster are evenly distributed in the sensing area. After clustering, the observation matrix in the cluster meets $a_{\text{value}} = 25$, $a_{\text{value}} = 50$, $a_{\text{value}} = 75$, respectively. The relationship with the standard reconstruction error and data fusion rate is shown in Table 1.

The table shows the following characteristics: (1) the standard reconstruction error rate decreases with the increase of data fusion rate. (2) When the data fusion rate in the control cluster is certain, the smaller the value of a_{value}, the smaller the standard reconstruction error rate of data, which also shows that the data similarity affects the reconstruction error rate of data.

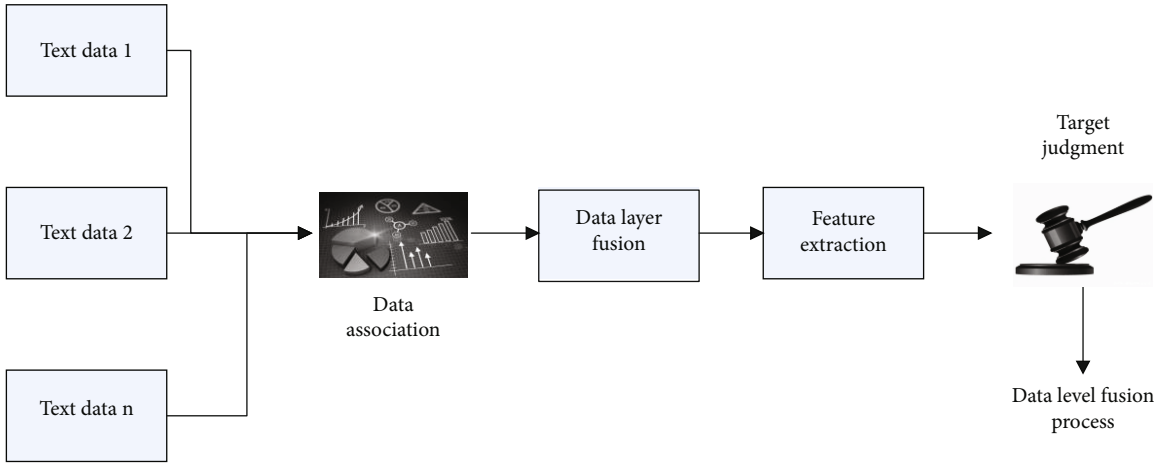

Figure 1: Data level fusion process.

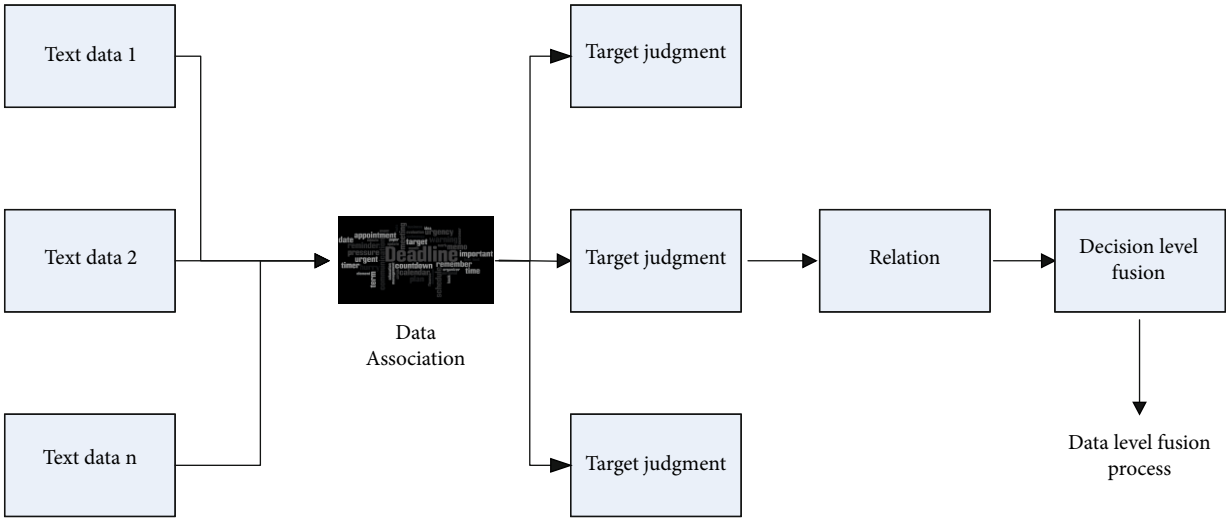

Figure 2: Decision level fusion process.

2.2. Feature Selection and Feature Extraction

2.2.1. Document Frequency (DF). The document frequency is determined according to the number of documents containing feature items, and the feature whose document frequency is higher than a certain threshold is selected as the feature item. The calculation formula is as follows:

$$\mathrm{DF}(f_k, c_i) = p(f_k \mid c_i). \tag{1}$$

2.2.2. Information Gain. The information gain indicates the average amount of information of the document class when the document contains a certain feature value. The calculation formula is as follows:

$$\begin{aligned}\mathrm{IG}(f_k) = \sum_{i=1}^{d}\left(p(f_k \mid c_i)\log\left(\frac{p(f_k \mid c_i)}{p(c_i)p(f_k)}\right)\right) \\ + p(\bar{f}_k \mid c_i)\log\left(\frac{p(\bar{f}_k \mid c_i)}{p(c_i)p(\bar{f}_k)}\right).\end{aligned} \tag{2}$$

2.2.3. Expected Cross Entropy (ECE). It is expected that the cross entropy is similar to the information gain, but it only

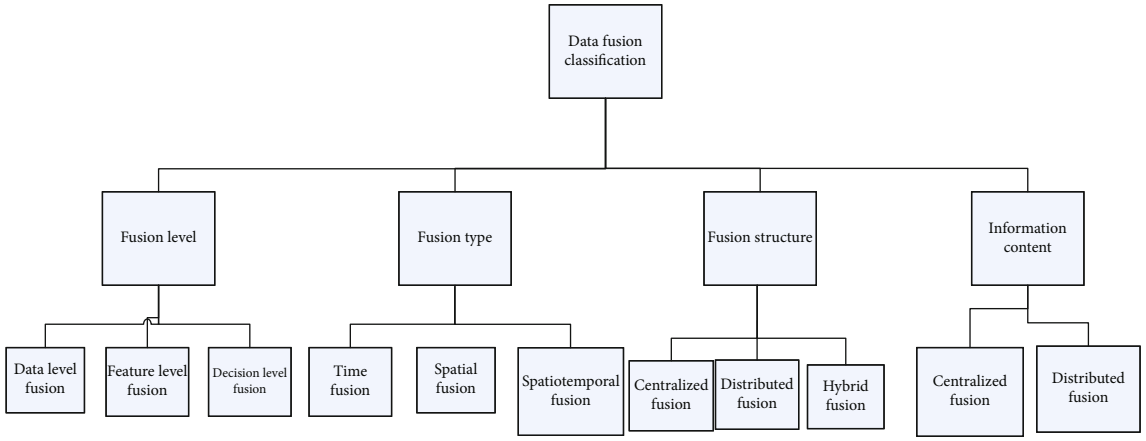

Figure 3: Classification of data fusion.

Table 1: Characteristics of principal component analysis.

Data fusion rate	0	15	30	45	60	75	90
$a_{\text{value}} = 25$	0.1	0.075	0.05	0.03	0.02	0	0
$a_{\text{value}} = 50$	0.16	0.13	0.08	0.05	0.035	0.02	0
$a_{\text{value}} = 75$	0.18	0.16	0.12	0.055	0.04	0.025	0

considers the occurrence of features in the text, and the formula is as follows:

$$\text{ECE}(f_k) = \sum_{i=1}^{|d|} p(c_i \mid f_k) \log \frac{p(c_i \mid f_k)}{p(c_i)}. \tag{3}$$

2.2.4. Mutual Information. Mutual information represents the correlation between text features and text classes. The formula is as follows:

$$\text{MI}(f_k, c_i) = \log\left(\frac{p(f_k \mid c_i)}{p(c_i)p(f_k)}\right). \tag{4}$$

2.2.5. CHI Statistic. The CHI statistic, also known as x2 statistic, assumes that the x2 distribution with the first degree of freedom between the feature and the category, the x2 statistic, is calculated as follows:

$$\text{CHI}(f_k, c_i) = \frac{N \times \left[p(f_k \mid c_i) \times p(\bar{f}_k \mid \bar{c}_i) - p(f_k \mid \bar{c}_i) \times p(\bar{f}_k \mid c_i)\right]^2}{p(f_k) \times p(\bar{f}_k) \times p(c_i) \times p(\bar{c}_i)}, \tag{5}$$

where A represents the number of texts containing the feature f_k and belongs to the category c_i, B represents the number of texts containing the feature f_k but not belonging to the category c_i, C represents the number of texts that do not contain the feature f_k but belongs to the category c_i, and D represents the number of texts containing feature f_k and does not belong to the category c_i.

2.3. Support Vector Machine (SVM). The most important point in SVM is the choice of kernel functions. The performance of different kernel functions is different. In the era of big data, a lot of data is linear and inseparable. In order to make performance better at this time, we must choose the most suitable kernel function. The main role is to map the linearly inseparable numbers in the input space into a high-dimensional space so that the feature data is separable. But to construct a kernel function V, you must know the mapping of input space to feature space. To know this kind of mapping, you should understand the distribution of the data set, but in many cases, you do not know the specific distribution of the processed data set, so it is difficult to choose a kernel function that conforms to the input space. You can choose the following common kernel function to replace your own kernel function:

2.3.1. Linear Kernel Function.

$$\kappa(x, x_i) = x \cdot x_i. \tag{6}$$

The linear kernel function is used to solve the problem of linear separability. From the above formula, we can know that the dimension of the feature space to the input space is the same. The parameters are small and fast. It is suitable for linearly separable numbers. When you first choose a linear kernel function, if the effect is not ideal, then switch to another kernel function.

2.3.2. Polynomial Kernel Function.

$$\kappa(x, x_i) = \left((x \cdot x_i) + 1\right)^d. \tag{7}$$

The polynomial kernel function can realize the mapping of low-dimensional feature data to high-dimensional data, but there is an obvious disadvantage that there are many parameters. When the order of the polynomial is high, the element value of the kernel matrix approaches infinity or infinity, and the calculation is performed. The complexity is too big to calculate.

2.3.3. Gaussian (RBF) Kernel Function.

$$\kappa(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{\delta^2}\right). \quad (8)$$

The Gaussian kernel function is a highly localized function that maps a sample of data into a high-dimensional space. The advantage of this kernel function is that its performance is better regardless of the number of samples, large or small, relative to the polynomial kernel. There are fewer function parameters. Therefore, in most cases, when you do not know which kernel function to use, you can choose the Gaussian kernel function first.

2.3.4. Sigmoid Kernel Function.

$$\kappa(x, x_i) = \tanh(\eta < x, x_i > + \theta). \quad (9)$$

If sigmoid is a kernel function, the support vector machine is equivalent to a multilayer neural network. In the actual machine learning modeling, if you have a certain prior knowledge of the data in advance, you should choose a kernel function suitable for the data distribution. If you cannot know, you should use the cross-validation method to try different kernel functions. The smaller the effect, the better the kernel function with the smallest error. Of course, when using SVM modeling, multiple kernel functions can also be combined to form a mixed kernel function. Of course, according to the experience of predecessors, the number of selected features and the size of the sample must follow certain rules. Figure 4 is the classification principle of the sample linear separable support vector machine in two dimensions.

The machine finds a decision boundary and separates the positive and negative categories. Then, the machine will find the distance from all sample points to this decision boundary and find the closest points to this decision boundary. Among them, the larger the distance, the better the decision boundary.

2.4. Rough Set. The rough set was proposed by Z. Pawlak, a professor at the Warsaw University of Technology in Poland. As a mathematical theory for dealing with uncertain, incomplete data and inaccurate problems, rough sets have been widely used in artificial intelligence, pattern recognition, data mining, and machine learning and knowledge discovery. It examines knowledge from a new perspective and uses knowledge as a classification ability. The size of classification ability is determined by the granularity of knowledge. The uncertainty of knowledge is caused by the large granularity of the composition domain knowledge, and this granularity of knowledge will represent the classification by the division of equivalence classes of equivalence relations. The following are the advantages of rough sets: the mathematical foundation is mature; no prior knowledge is required; the operation is simple; and the theories for dealing with other uncertain problems are strongly complementary. The following is the disadvantage of rough set: rough set can only solve discrete data.

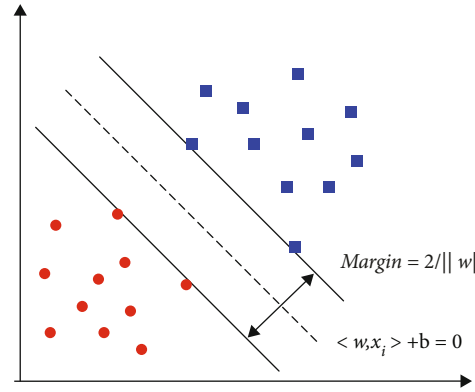

Figure 4: Optimal classification surface in the case of two-dimensional linear separability.

A rough set uses an information system as a description object. An information system is a collection of objects described by a set of multivalued attributes. Each object and its attributes have a value as its descriptive symbol. The information system can be represented by an information table, the rows of the information table corresponding to the research object, the columns corresponding to the attributes of the object, each row containing the descriptors representing the attributes of the corresponding object feature items, and the category information of the corresponding objects. Rough sets can also represent classification rules in decision information tables and attribute reduction based on the importance and dependencies of attributes to generate decision rules for each class. The test data set is used to calculate the confidence and gain of the candidate rule to verify the extracted candidate rule as the final classification rule. Before establishing candidate rules, the decision table is divided into two groups in a random manner: a% of the data is regarded as the training data group; 1-a% of the data is regarded as the test data group. Figure 5 is a rough set flow chart.

Firstly, the demand data is processed. Based on the similarity definition in the algorithm definition, the equivalent item is calculated for each demand item; based on the rough set, the similarity threshold is calculated, and the initial equivalence class is modified; and the validity index of the cluster is calculated. Judge the quality of the clustering results, and obtain the clustering results that meet the needs of customers through repeated calculations.

2.5. Definition of Text Classification Method Based on Granularity Calculation. The particle size calculation mainly includes three parts: particles, grain layer, and grain structure. Among them, the particle is the most basic element that constitutes the particle size calculation model and is the primitive of the particle size calculation model. The grain layer is the overall composition of all the particles obtained according to the granulation criterion of a practical demand and is an abstract description of the problem space. The grain structure is the relationship structure formed by the interconnections between the grain layers. The complexity of the grain structure determines the complexity of the

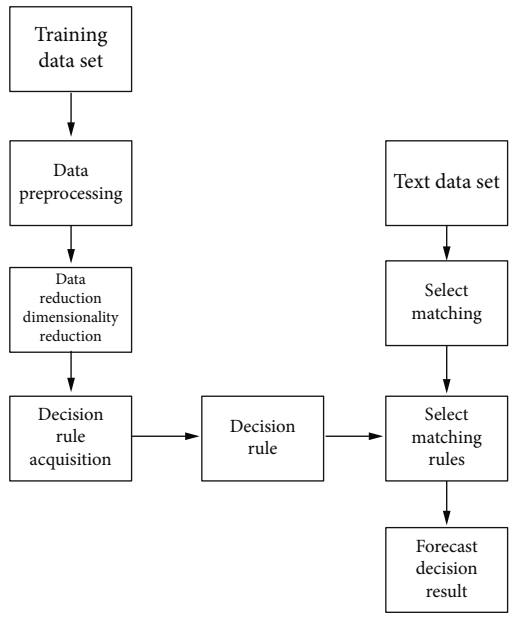

Figure 5: Rough set flow chart.

problem solving. Granularity calculations can be solved in two ways, namely, granulation and particle calculation.

Definition 1. Set a corpus D, where D contains m texts d; that is, after each word of D is segmented, multiple feature words w are obtained, and then each feature word w is trained by Gensim library. The corresponding feature word vector w is obtained, and the dimension is k-dimensional. Set the word vector set obtained by the entire corpus $W = \{w_1, w_2, \cdots, w_n\}$. Based on the perspective of granular computing, this set of word vectors is called the word vector space.

Definition 2. For a word vector space, the feature word similarity $W = \{w_1, w_2, \cdots, w_n\}$ based on the space is defined as

$$R_W^{\lambda} = \left\{ \left(\mathbf{w}_i, \mathbf{w}_j \right) \in W \times W \mid S\left(w_i, \mathbf{w}_j \right) \geq \lambda \right\}, \quad (10)$$

where $S(w_i, w_j)$ represents the similarity between the word vector w_i and w_j; this paper uses the Euclidean distance to measure the similarity; λ is a threshold, which satisfies $0 < \lambda \leq 1$. It can be seen that the definition of the similarity of the characteristic words is a special binary relationship. Responsiveness and symmetry are satisfied, but the transfer characterization is not necessarily satisfied, so it can induce an overlay on the word vector space W.

It can be seen that the feature word similarity segmentation divides the whole word vector space into one feature word class, which is equivalent to granulating the entire word space. Each granulated word class maintains a high similarity inside, and the similarity the definition is embodied by the threshold λ, so the lambda value has a significant influence on the final granulation result.

3. Experiment

3.1. Data Source. The experiment uses the TanCropusV1.0 Chinese corpus, firstly classifying the corpus into 12 text granularity sets by class. Then, the stop words and 1-gram words of 12 text granularity sets (5504 total) were removed, and the characteristics of 12 text granularity sets were evaluated by DF, GSS, ECE, and CHI. In a descending order, select the top 20 features as the condition attributes for each granularity set and calculate the purity of each granular set.

Select five from any of the 12 text granularity sets (e.g., G1, G4, G7, G8, and G11) to experiment; first, divide the five text granularity sets into test sets and training sets in a ratio close to 3 : 2, and select the characteristics of the training set separately, select the top 20 features of each CHI evaluation in the granularity set, and then "compress" the text set; that is, remove the text with the same result after feature selection, and count the number of texts and missing text after feature selection.

3.2. Experimental Platform

3.2.1. Skip-Gram Model Experimental Environment Configuration. There are many practical and convenient libraries. This article uses the Word2Vec in the Gensim library to complete the training process of word embedding. From the data preprocessing of the text to the completion of the word embedding training, the environment of the whole experimental process is shown in Table 2.

In this experiment, in addition to the user's comment data, I also added Wikipedia data as a library to train the word embedding of each word. The reason for this is that the corpus has a certain scale, and the effect of training is more. Well, it can fully reflect the correct position of these words in the vector space, so that a higher degree of word embedding can be obtained.

3.2.2. Skip-Gram Model Parameter Configuration. When using the Skip-Gram model to train the word embedding of each word, it mainly involves setting two parameters: first is the size c of the training window, and second is the length of the word vector, that is, the k of the word embedding. In theory, the larger the window c, the better the completeness of the model, but if c is too large, it will lead to many irrelevant words being trained. Therefore, in the course of the experiment, the random selection window is selected, and the window size is generally less than or equal to 10. For example, after determining the size of c, for a word, the distance $R = \text{rand}(1, c)$ is selected as the result of selecting the R words before and after the word as the final generated prediction result. Because of the corpus involved in this article, each text contains only a few to dozens of words, and the text is relatively short, so the size c of the window is set to 8.

3.3. Classification Algorithm Calculation

3.3.1. Information Granularity Rule Acquisition. Input: information system $S = (U, A, V, f)$

Output: rule set RUL

Table 2: Data preprocessing and extraction feature experimental environment.

Lab environment	Environmental configuration
Operating system	Centos6.5
CPU	Intel Core I5-650 3.20 GHz
RAM	8GB
Programming language	Python3.6
Word segmentation tool	ICTCLAS2016
Training tool	Word2Vec

(Step 1) The first step is to divide the domain according to the decision attribute value into different information granularities $G = \{G_{d_1}, G_{d_2}, \cdots, G_{d_s}\}$.

(Step 2) Determine if impurities are included between the particle sizes.

(Step 3) The third step is to perform attribute reduction on each information granularity in G, set the rule preamble length including the impurity attribute, and obtain the intragranular rule RUL.

(Step 4) Combine the rules between the granularities $\mathrm{RUL} = \bigcup_{G_F \in G} \mathrm{RUL}_{G_G}$.

(Step 5) Output RUL

3.3.2. Information Granularity Attribute Reduction. Input: information granularity G_i

Output: reduction attribute set REUDi

(Step 1) Calculate the coordination matrix H_i of G_i.

(Step 2) The second step is to find the most similar samples x_i, x_j in H_i, determine the elements $H_{i,j}$, delete other elements in sub H_i containing subscript i (or j), and initialize the reduction attribute subset R_i and the sample division subset X_i.

$$\begin{aligned} R_i &\longleftarrow H_{i,j}, \\ X_i &\longleftarrow x_i, x_j. \end{aligned} \tag{11}$$

(Step 3) The third step is to search for the sample x_k which is the most similar to H_{ij} and simultaneously update the reduction attribute subset R_i and the sample division subset X_i and delete the other elements in the H_i subscript containing k.

$$\begin{aligned} R_i &\longleftarrow H_{ij} \cap x_k, \\ X_i &\longleftarrow X_i \cup x_k. \end{aligned} \tag{12}$$

Table 3: The purity of the granularity of the text information corresponding to the four feature extraction methods.

Granular set number	DF	GSS	ECE	CHI
G1	35%	65%	35%	100%
G2	80%	95%	80%	100%
G3	60%	90%	60%	95%
G4	30%	70%	30%	95%
G5	40%	50%	40%	85%
G6	70%	80%	60%	100%
G7	55%	75%	55%	100%
G8	60%	95%	60%	100%
G9	60%	85%	60%	90%
G10	30%	80%	30%	95%
G11	20%	30%	20%	70%
G12	45%	65%	45%	100%
Mean	48.75%	73.33%	47.92%	94.17%

Table 4: Word2Vec.

Tunable parameter	Value
Number of iterations	20
Model selection	Skip-Gram
Method selection	HS
Context window	8
Sample value	Le-4
Lowest frequency	5

(Step 4) Repeat step 3 until the length of R_i is less than or equal to the specified threshold and add R_i to REUDi, and if $G_i = X_i$, terminate the search.

(Step 5) Output the reduced attribute set REUDi.

4. Results

4.1. Feature Extraction Purity Analysis. The experiment uses the TanCropusV1.0 Chinese corpus, firstly classifying the corpus into 12 text granularity sets by class. Then, the stop words and 1-gram words of 12 text granularity sets (5504 total) were removed, and the characteristics of 12 text granularity sets were evaluated by DF, GSS, ECE, and CHI. In a descending order, select the top 20 features as the conditional attributes of each granularity set, and calculate the purity of each granularity set. The results are shown in Table 3.

As can be seen from Table 2, the text granularity set obtained by DF and ECE for feature selection has a lower average value of 48.75% and 47.92%, respectively, and when ECE is used, the purity of a single text size set is up to 80%. The minimum is 20%. The text granularity set obtained by CHI feature selection has the highest average value, reaching 94.17%, wherein the single text granularity set has the highest purity of 100%, the lowest is 85%, and the six text granularity sets have a purity of 100%. It can be seen that among

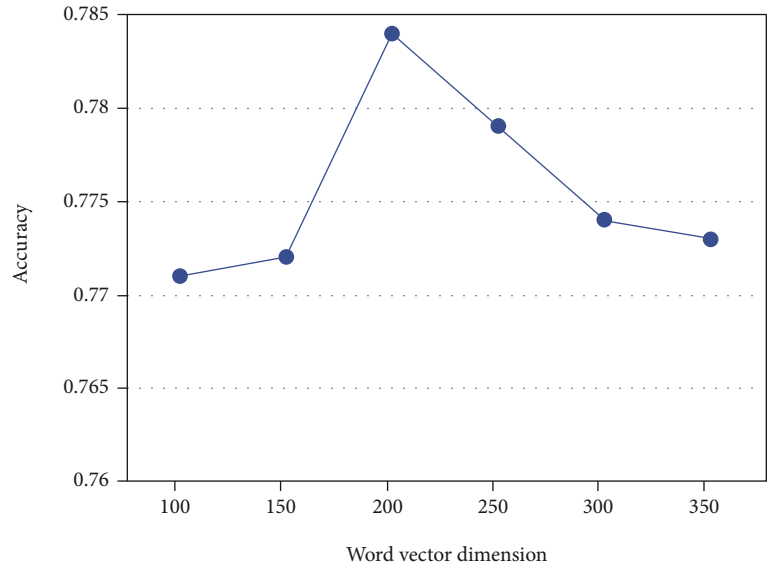

Figure 6: Accuracy rate results.

the four feature selection methods, the 12 feature granularity sets have the same feature words in the top 20 feature words obtained by DF and ECE, and the same top 20 feature words obtained by CHI are the same. There are fewer feature words, so CHI is more suitable as a feature selection method for text granularity sets.

4.2. Influence of Vector Dimensions and Thresholds on Experimental Results. This experiment mainly considers the influence of the word vector dimension on the experimental results. In this experiment, Wikipedia is used as the training set, and the Word2Vec tool is used to train the word vector. The setting of each parameter is shown in Table 4. The word vector dimension is a multiple of 5 from 100 to 350. Experiments are performed using Word2Vec+SVM to determine the influence of the word vector dimension on the experiment. The experimental results are shown in Figure 6. Table 5 is a comparison table of experimental results.

From the above comparative experimental results, table analysis is as follows:

(1) In the traditional machine learning model experiment, linear SVM performs best, and the final text classification accuracy can reach 85.46%. The reason is analyzed: the objective function of the support vector machine model is to minimize the structural risk, which greatly reduces the model's requirements for data volume and data distribution, so the performance is the best when the number of samples is small. Compared with the traditional machine learning algorithm, this method is more excellent in performance, even higher than the best performing SVM, which is nearly 8% higher in classification accuracy

(2) In the three traditional convolutional neural network model experiments, the best performance is the CNN-non-static model. Moreover, the model is nearly 3% higher in final classification accuracy than the CNN-rand model. The reason is as follows: on the one hand, the artificial random initialization feature representation can not be very abstract data input distribution; on the other hand, the Word2Vec trained word vector is used as the CNN input feature in advance, and in iterative training, the input feature is keep up to date

Table 5: Comparison of experimental results.

Model	Accuracy (%)
Word embedding+linear SVM	85.46
Word embedding+LR	85.20
Word embedding+random forest	82.61
CNN-rand	88.52
CNN-static	90.36
CNN-non-static	91.79
Text method	93.25

4.3. Classification Accuracy Analysis. After text classification for all 1918 test texts using the improved rough set text classification technique, the classification results of Tables 3 and 4 were obtained according to the evaluation methods given. The comparison shows the following:

(1) For small-scale test sets, the correct number of recalls and precision of texts have increased significantly

(2) There are significant reductions in the number of false recalls and nonrecalls in all test sets, such as environmental, educational, and economic; the wrong recall rate for computers and transportation is even reduced to zero

Table 6: Test results analysis table.

	G1	G4	G7	G8	G11
Number of test texts	207	50	263	193	178
Correct match number	174	50	231	182	168
Error match number	32	0	32	11	9
Number of lost text	1	0	0	0	1
Recall rate	84.45%	100%	87.83%	94.30%	94.92%
Precision rate	86.14%	84.75%	89.53%	100%	89.36%

(3) The recall and F1 values of all test sets are increased. The improved text classification technology reduces the false recall rate of computer and traffic to zero. Therefore, the recall rate of these two types reaches 100%

(4) The overall recall rate and precision rate of the overall ten text categories are improved to varying degrees

According to Bayesian theory, this is equivalent to the introduction of prior knowledge, which guides the model to converge to the optimal solution along a better direction during the training process. The CNN model is characterized in that it does not require manual feature selection in advance, which greatly reduces manpower consumption, and the input features are continuously updated during the training process, which indicates that the CNN text classification process is a combination of feature selection and training.

4.4. Text Test Results. During the test, the test set is divided into five groups according to the category, and then the five groups are matched with the rules in the rule base, and the classification results are statistically analyzed. The results are shown in Table 6.

It can be seen from Table 6 that the total number of texts participating in the test is 891, of which 807 are correctly classified and 84 are classified incorrectly. Since the test text set does not contain the feature words in the rule base, 2 articles are lost (unable to be judged). The macro average accuracy of the five categories is 89.96%, the macroaverage recall rate is Macro.r. 92.30%, and the macro average F1 is 91.11%. The microaverage accuracy Micro_p of the five categories is equal to the microaverage recall rate of Micro_r of 90.55% and the microaverage F1 of 90.55%.

5. Discussions

According to the degree of similarity of the conditional attributes between the particle sizes, the concept of particle size purity is proposed. It is proved by experiments that the top 20 features obtained by CHI are evaluated as the conditional attributes of each information granularity. Different from the traditional method of attribute reduction by constructing discernible matrix, this chapter constructs the attribute matrix by deconstructing the decision matrix. According to the search method proposed in this chapter, the five categories are trained to obtain attribute reduction sets, and 34 rules with rule precedence greater than or equal to 2 are obtained. Experimental results show that these classification rules have a high classification ability.

There are significant reductions in the number of false recalls and nonrecalls in all test sets, such as environmental, educational, and economic; the wrong recall rate for computers and transportation was even reduced to zero; for the overall ten text categories, the recall rate and precision rate have been improved to different extents; for small-scale test sets, the correct number of recalls and precision of texts have increased significantly.

Linear SVM performs best, with a final text classification accuracy of 85.46%. Compared with the traditional machine learning algorithm, this method is more excellent in performance, even higher than the best performing SVM, which is nearly 8% higher in classification accuracy. In the three traditional convolutional neural network model experiments, the best performance is the CNN-non-static model. Moreover, the model is nearly 3% higher in final classification accuracy than the CNN-rand model.

Through three sets of comparative experiments, it can be known that compared with the traditional machine learning classification algorithm, the proposed method achieves better results in classification effect; using word embedding to initialize text features, compared with artificial randomization, text feature initialization is more excellent in classification effect; the method of this paper also achieves better classification accuracy than the traditional best convolutional network model. Finally, this chapter gives a detailed conclusion analysis for the experimental parameter settings and experimental results.

6. Conclusion

(1) In terms of feature dimension reduction, unlike the existing methods of selecting features by evaluation function, this paper proposes a new feature clustering method, which aggregates different feature words by calculating the distribution distance between features. To reduce the feature dimension, this can prevent part of the sample caused by the feature evaluation function from being discarded because it does not contain the selected feature. The experiment also proves that the clustering method can obtain higher classification accuracy through SVM test

(2) Using the Skip-Gram neural network language model to train the word embedding of each word, that is, the word vector of the feature word, construct the word vector space based on the word vector of all feature words, and then construct the relevant granulation relationship to the word vector space. As a result of granulation and granulation, each feature word in the word vector space has a feature word class, also called feature word granule, and the feature words inside each feature word class maintain a high degree of similarity, so the feature words in the feature word class are selected to expand, and the purpose of text expansion is achieved

(3) In terms of classification algorithm, this paper combines the relevant theory of granular computing with text classification. Firstly, the test text set is granulated, which reduces the complexity of attribute reduction in rough set. Second, for a single information granularity, by constructing a synergistic matrix and heuristic search attribute reduction set. The experiment extracted 34 rules from 1811 training samples. These rules were used to test 891 unknown samples. The average macroaccuracy was 89.96%, and the microaverage accuracy was 90.55%

(4) The experimental results show that the clustering method can obtain higher classification accuracy than other feature selection methods when using SVM as the classifier. According to the correlation principle of the rough set, feature selection is made for each information granularity, the selected feature is used as the condition attribute and the coordination matrix is constructed, and the most similar sample is heuristically searched to obtain the attribute reduction set

The experimental results show that in the three traditional convolutional neural network model experiments, the best performance is the CNN-non-static model. Moreover, the final classification accuracy of this model is nearly 3% higher than that of the CNN rand model.

Data Availability

This article does not cover data research. No data were used to support this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Liaoning Collaboration Innovation Center for CSLE, National Natural Science Foundation of China (Grant No. 61832004), and Projects of International Cooperation and Exchanges NSFC (Grant No. 62061136006).

References

[1] W. Tian, "A method of feature selection based on Word2Vec in text categorization," in *2018 37th Chinese Control Conference (CCC)*, pp. 9452–9455, Wuhan, China, 2018.

[2] L. I. U. Bei, "Oceanographic big data text categorization algorithm based on improved mutual information," in *Proceedings of 2017 2nd International Conference on Artificial Intelligence: Techniques and Applications (AITA 2017)*, p. 5, Shenzhen, Guangdong, China, 2017.

[3] J. Ni, G. Gao, and P. Chen, "Chinese text auto-categorization on petro-chemical industrial processes," *Cybernetics and Information Technologies*, vol. 16, no. 6, pp. 69–82, 2016.

[4] W. Chen, "Research on text categorization model based on LDA-KNN," in *Proceedings of 2017 IEEE 2nd advanced information technology, electronic and automation control conference (IAEAC 2017). IEEE Beijing Section, global union academy of science and technology, Chongqing global union academy of science and technology, Chongqing geeks education technology co., ltd: IEEE BEIJING SECTION (multinational Institute of Electrical and Electronics Engineers Beijing Branch)*, p. 8, Chongqing, China, 2017.

[5] D. Lianhong, S. Bin, and Z. Hongwei, "A short text classification method based on knowledge graph extension," *Information Engineering*, vol. 4, no. 5, pp. 38–46, 2018.

[6] G. Weiyin and W. Li, "Text classification based on convolutional neural network and XGBoost," *Communication Technology*, vol. 51, no. 10, pp. 2337–2342, 2018.

[7] Z. Man, X. Zhanguo, L. Bing, and Z. Yong, "A character-level text classification method for full convolutional neural networks," *Computer Engineering and Applications*, vol. 56, no. 5, p. 7, 2020.

[8] L. Wang, Z. Tao, C. Cai, A. Zhu, and Q. Luo, "Study on text classification of TF-IDF based on chi-square statistics improvement," *Electronic World*, vol. 6, pp. 24-25+28, 2019.

[9] L. Yao, Q. Z. Sheng, X. Wang, S. Wang, X. Li, and S. Wang, "Collaborative text categorization via exploiting sparse coefficients," *World Wide Web*, vol. 21, no. 2, pp. 373–394, 2018.

[10] T. Wang, Z. Long, L. Huaiquan, L. Li, and C. Siqi, "Text classification method based on key word learning," *Journal of Shandong Normal University(Natural Science)*, vol. 34, no. 1, pp. 54–60, 2019.

[11] G. Chaolei and C. Junhua, "Study on Chinese text classification based on SA-SVM[J].," *Computer applications and Software*, vol. 36, no. 3, pp. 277–281, 2019.

[12] H. Chao and C. Junhua, "Chinese text classification based on improved K nearest neighbor algorithm," *Journal of Shanghai Normal University(Natural Science)*, vol. 48, no. 1, pp. 96–101, 2019.

[13] D. Junhong, L. Xiaoyu, and M. Dejun, "A text classification training method based on incomplete labeling," *Microprocessor*, vol. 40, no. 1, pp. 20–24, 2019.

[14] L. Kai, "Study on Chinese text classification method," *Computer Knowledge and Technology*, vol. 15, no. 4, pp. 242–244, 2019.

[15] J. Ma and Y. Ma, "Semantic-driven classification of judicial document learning methods," *Computer Applications*, pp. 1–6, 2019.

[16] J. Xingguo, J. Wan, C. Xiaodong, L. Haiou, and C. Yi, "A fine-grained vehicle identification algorithm based on singular

value decomposition and central metric," *Journal of Xidian University*, pp. 1–6, 2019.

[17] Z. Haoru, Z. Yong, and L. Guozhu, "Study on fine-grained image classification algorithm based on RPN and B-CNN," *Computer Applications and Software*, vol. 36, no. 3, pp. 210–213+264, 2019.

[18] L. Dangwei, Z. Yongzhe, L. Kewen, and C. Zhenwen, "The coarse-grained particle swarm optimization algorithm for subgroup stratification," *Computer Engineering and Design*, vol. 40, no. 2, pp. 389–393, 2019.

[19] L. Jingrui and P. Dongyang, "Research on data source multigranularity search algorithm based on probability and statistics," *Computer Products &Circulation*, vol. 2, pp. 89–152, 2019.

[20] L. Jinshuo, L. Yangmei, J. Zhuangyi, J. Deng, Q. Haigang, and J. Pan, "A large-scale data fine-grained parallel optimization method based on PMVS algorithm," *Journal of Wuhan University(Information Science Edition)*, vol. 44, no. 4, pp. 608–616, 2019.

[21] Z. Suzhi, Y. Wei, C. Xiaoni, and L. Penghui, "A coarse-grained parallel AP clustering algorithm based on intra-class and inter-class distance," *Journal of Central China Normal University(Natural Science)*, vol. 52, no. 6, pp. 781–787+797, 2018.

[22] C. Ronghu and H. Yunjie, "Application of SPMD-based coarse-grained parallel genetic algorithm in path optimization of solid warehouse," *Software Guide*, vol. 17, no. 12, pp. 108–112, 2018.

[23] Y. Yingjian, L. Min, and Q. Yiyang, "Design and implementation of a hardware Trojan detection algorithm for coarse-grain reconfigurable arrays," *Journal of Electronics & Information Technology*, pp. 1–8, 2019.

[24] Q. Jin and L. Jianhua, "A fine-grained attribute analysis algorithm for terminal trusted remote proof," *Computer and Digital Engineering*, vol. 46, no. 10, pp. 1970–1973, 2018.

[25] Z. H. A. O. Yilin, J. I. A. N. G. Lin, M. I. Yunlong, and L. I. Jinhai, "A dynamic parallel updating algorithm for degree-grained rough set approximation sets based on weighted granularity and dominance relation," *Computer Science*, vol. 45, no. 10, pp. 11–20, 2018.