Design Engineering

ISSN: 0011-9342
Issue: 11 | Pages: 115 - 127

Application of Support Vector Machine in File Text Classification

Xiangzhen He[1], Fucheng Wan[1,*], Dengyun Zhu[1], Lei Zhang[1], Shengyin Zhu[1], Yuan Lin[2], Xuebin Yang[1], Yerong Hu[1], Yihao Zhang[1]

[1]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

[2]Lanzhou University of Finance and Economics Information Center, Lanzhou, Gansu 730000, China

*Corresponding Author:Fucheng Wan

Abstract:

As information technology is developing rapidly, numerous data is generated every day. Machine learning methods are applied more and more in text classification, information filtering, and information retrieval. Archives text data is a knowledge carrier that records historical files and gathers a lot of real and reliable information. It is of great significance to design suitable classification methods for support vector machines, to classify file data and to construct digital archives.

***Keywords**: Machine learning, Text classification, Support vector machines, Digital archives.*

I. INTRODUCTION

Nowadays, the electronic medium pro-duced has grown continuously, and traditional paper media have become less and less desirable. Internet information has penetrated into every aspect of people's daily life, such as Weibo, WeChat, email, and search engines. The massive amounts of data that each of these media have at one time or another have huge application value. But confront thesenumerous data resources, it is hard for people to fast and effectively search the real message they need. Therefore, how to manage these information reasonably and effectively is a problem faced and experienced in the current era of big data, and has also become an important research topic in the field of information processing.

There are many manifestations of information, such as speech, graphics, and text. However, at present, text is still the main form of expression of information. Both graphics and speech can be converted into textual forms for analysis. Text classification technology has widespread appliance as information filtering, information retrieval, search engine, digital library and other fields. Text Categorization or Text classification is an activity in which natural language text is marked by related categories in a predefined set. The popular explanation is the process of extracting the text of unstructured words into general tags. Select these generic tags from a set of predefined categories. Categorize your content and products that are relevant to you, making it easy for users to search and navigate on the website or application.

Since the text classification technology was put forward in the last century, it has been developed until now that both the rules-based and statistics-based text classification methods have been proposed by many computer scientists. The famous classification methods at this stage include support vector machine method, naive Bayes method, K nearest neighbor method, and decision tree method. In last several years, artificial intelligence has been widely mentioned, and the method of text classification based on deep learning has also been applied to a large extent.

Text classification technology began to appear in the Mid-20th century. In1957, H. P. Luhn mentioned the word frequency statistics classification [1]. VSM was proposed. In 1970, by Salton et al.VSM converts the features of the text into a vector form and then uses an appropriate amount of similarity to measuring text similarity. The problems of abstract mathematical are simple and easy to learn. In the case of a relatively late start in domestic text classification research, it is generally applied to Chinese text classification by using mature foreign classification techniques combined with Chinese text characteristics. In Chinese text,people display much interest in how to obtain and improve the algorithm of text classification in reasonable application and make better accuracy of the classification algorithm, lower algorithm classification time [2].

Two steps are required to complete the classification of text data. First, the categories of predefined data need to establish a classifier to represent,which is operation of classification stage. Construct a classifier after analyzing the classification algorithm and learning on training set. The data content of labels and training sample categories form the data set.

Then next is to establish a classifier,and evaluate the Pros and cons of this classifier through test samples. There are many classification methods, but research shows that SVM aremore effective than traditional classification way in classification performance, especially generalization ability [3].

Because of its good generalization ability and strong learning ability, the classification methods of support vector machines make the application of support vector machines very extensive [4]. They have made important achievements in the fields of text classification, face recognition, and hand writing recognition research results. This paper will start from the principle of support vector machine and verify its classification effect through the application on the corpus. This is of great significance for learning support vector machines.

II. TEXT CLASSIFICATION TECHNOLOGY

2.1 Text Classification Description

This process is mainly used to train a defined class or set of labels, train a certain method, construct a learning system according to the characteristics of text data, and use the learning system to train the training data as a classification system model. To achieve the effect of automatic classification. See in Fig 1.

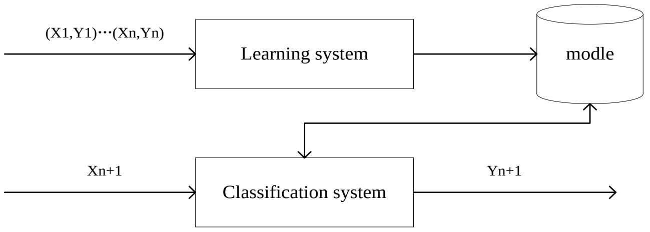

Fig 1: Classification problem description

In the above figure, Xn+1 is the input data. After the classification system, the output Yn+1 is the automatically classified category label.

F. Sebastiani describes the text classification task with the following mathematical model: The task of text classification can be understood as obtaining such a function Φ:D×C→{T,F},and,D={d1,d2,…,d|D|} is the document that needs to be classified, C={c1,c2,…,c|C|} represents the set of categories under the predefined classification system, The T value means that for (dj, ci), the document dj belongs to the class ci, and the F value means that for (dj, ci) the document dj does not belong to the class ci. To put it differently, the main purpose of text classification is to find a valuable function mapping, and accurately complete the function mapping from D×C to T/F value. This mapping process is essentially a so-called classifier.

The formal definition of text classification is as follows:

Let i = 1, M be the M documents in the document collection, j = 1, N be the predefined N categories of topics, you can give such a classification result matrix C = (cij); and an element cij in the matrix represents the relationship between the i-th document and the j-th category. That is to say, automatic text classification can be attributed to the process of determining the value of each element of the matrix C above; Use a Boolean value of 1 or 0. If the value of cij is 1, it means that document i belongs to the jth category. If the value is 0, then document i cannot be classified into category j, namely:

$$c_{ij} \begin{cases} 1, & Documentibelongstocategoryj \\ 0, & Documentidoesnotbelongtocategoryj \end{cases} \tag{1}$$

For single-category classification, that is, a document can only be classified into one category, we can add qualifications. For all elements in the jth row (j = 1, ..., N), it must meet:

$$\sum_{i=0}^{N} c_{ij} = 1 \tag{2}$$

In practical applications, according to different predefined categories, there are two types of classification systems: Two-class classifier and multi-class classifier. In terms of text labeling, text classification can be divided into single label and multi-label. The task of the text classification system is simply said: under the predefined classification system, the association between the text and the category is automatically determined according to the content relevance of the text.As far as mathematical be concerned, text classification is a function mapping process, which maps the text of unspecified categories to predefined categories. The mapping can be one-to-one mapping or one-to-many mapping, because usually a text can be associated with multiple categories at the same time.

2.2 The Process of Text Classification

From a macro perspective, this process can be divided into two stages: classifier training and classifier inspecting. In the process of designing the classifier, the preprocessing of text, the representation of the text, and the design of the classifier algorithm are designed. The final result output is the result of the automat-ic classification process. A simple process description is shown as Fig 2.

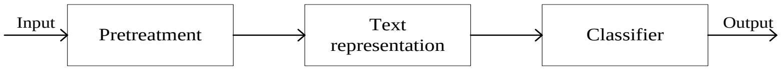

Fig 2: Process of text classification

When classifying Chinese texts, the preprocessing of texts consists mainly of text segmentation, stop words, and feature selection and feature extraction methods. The representation of text is mainly in the form of Vector Space Model (VSM), Probability Model, and Bool Model. Among them, VSM Model is a widely used model. The last is the design of the classifier. For different texts, its characteristics are inconsistent. Therefore, when designing the classifier for this feature, different classification algorithms will be selected. In the research of this paper, SVM is mainly chosen. The method of classifying the file text data is designed.

The SVMtext classification algorithm is mainly divided into four steps, Text feature extraction, text feature representation, normalization processing and text classification.

2.2.1 Text Feature Extraction

Currently, when extraction is processing, the feature selection process will be simplified by the feature independence assumption, reaching a compromise between calculation quality and calculation time. The general method is to select the best feature as the text feature subset by setting the threshold of feature according to the feature vector of the vocabulary in the text, and establish the feature model. (Be-fore feature extraction, segment words and stop words).

There are many methods for this feature extraction, and the most common method is to select features by word frequency. First calculate the weights by word frequency, sort them by weight from large to small, and then remove the useless words. These words are usually irrelevant to the topic. Any kind of articles may appear in large numbers. It is generally defined in the stop word list. After removing these words, a new sequence is sorted out, and then the top 8, 10 or more words with the highest weight can be selected according to actual needs to represent the core content of the text. In summary, the extraction steps of feature items can be summarized as:

(1) Word segmentation of all training documents, and the text is represented by the

dimensions of these words as vectors;

(2) Count all the words and their frequencies in the documents in each category, and then filter to remove stop words and single-word words;

(3) Count the total word frequency of words appearing in each category, and take some of the most frequent words as the characteristic word set of this category;

(4) Remove the words that appear in each category, and merge the feature word sets of all categories to form a total feature word set. The final feature word set is the feature set we use, and then use this set to filter the features in the test set.

2.2.2 Text Feature Representation

Following is to compute the word weights:

$$w_{tk} = \frac{tf_{ik} * \log(N/n_k + 0.01)}{\sqrt{\sum_{i=k}[tf_{ik} * \log(N/n_k + 0.01)]^2}} \tag{3}$$

Where tf_{ik} represents the frequency of special diagnosis times tk appearing in document di, N represents the total training data, n_k represents the number of occurrences of tk in the training data set. As mentioned,words that appear frequently in a batch of files will be less discernible and weighted, and the lower the weight; while in a document, a word with higher frequency, the higher the degree of differentiation .The larger, the greater the weight.

2.2.3 Normalization Processing

Using algorithms to restrict to be processed data within the required specific range is standardization processing.

$$\frac{a - min}{max - min} = b \tag{4}$$

In the formula, ‘a’ represents keywords frequency , ‘min’ represents words in all texts that with minimum frequency , so ‘max’ represents words in all texts that with maximum frequency. This phase is standardization. When juxtapose word frequencies, large deviations are likely to occur. This is a way to make text classification more accurate.

2.2.4 Text Classification

After the above processing, a sample set that has been abstracted into vectorization is there, then calculate the similarity of this sample set according to the trained template file. If it is this category, use the other categories template to calculate until it is assigned to the relevant category. This is the way to use the svm model for text classification.

SVM-based system implementation (as shown), See in Fig 3.

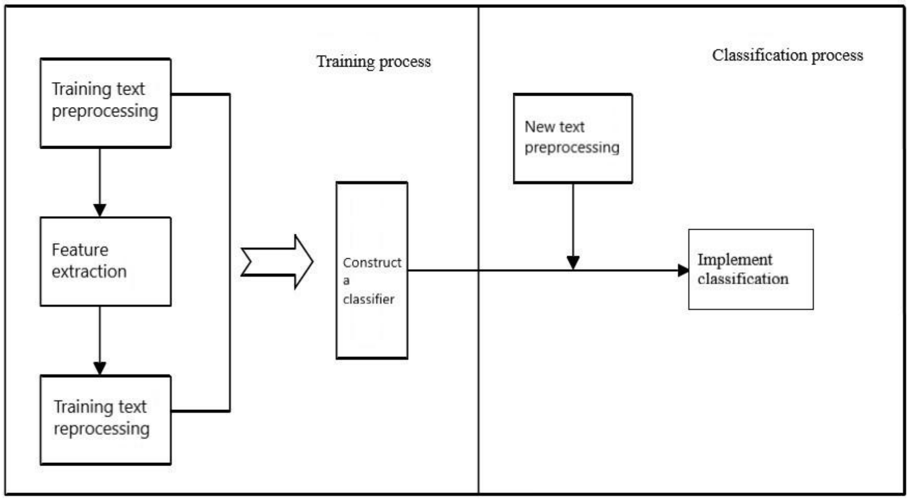

Fig 3: SVM-based system implementation

III. SUPPORT VECTOR MACHINE THEORY

Corinne Cortes and Vapnik proposed the SVM method in 1995, which is a method to resolve the binary classification problem. Finding the best high-dimensional classification hyperplane is the basic idea of SVM.

3.1 Support Vector Machine Thinking

The theory of SVM was proposed by Vapnik. Its main theoretical basis is the theory of statistical learning. Because of its good generalization performance and ease of use, SVM has extensive research in image, phonetics, video, and text.

The SVM is designed to seek an optimal hyperplane in linearly separable data. This optimal hyperplane can classify two different categories optimally. The core of the algorithm is about use mathematical thinking to solve this optimal hyperplane. Shown as Fig.4

In Fig.4,$w \cdot x - b = 0$ is a hyperplane that can classify two categories in the graph is optimized. The black and white points on the way represent two different categories.$w \cdot x - b = 0$and $w \cdot x - b = 0$ represents a straight line that can maximize separation of the two classification samples, $w \cdot x - b = 0$ and $w \cdot x - b = 0$ the distance between them is the separation distance. W represents the normal vector of the hyperplane, $w = (w_1; w_2; \dots; w_{d1})$ and b is the intercept between the hyperplane and the support vector. To facilitate the formula description, the formula is unified to represent:$w^T x - b = 0$ (1)

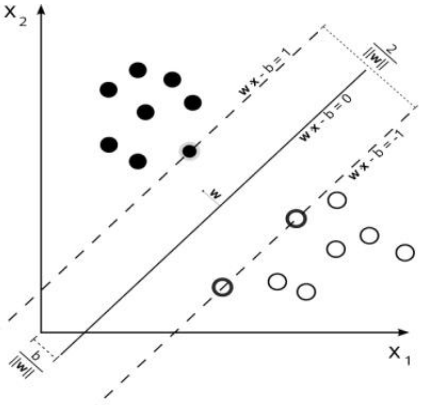

Fig 4: Support vector machine principle

As can be seen from the above figure, the distance between two support vectors can be expressed as $\frac{2}{||w||}$(2), and this algorithm is about to seek the maximum distance between these two categories, namely$\max_{w,b} \frac{2}{||w||}$(3). In order to facilitate the solution of this formula, the problem is converted to$\min_{w,b} \frac{1}{2}||w||^2$ (4). The final problem becomes that of seeking w, b and finding (4).

In the support vector, solving w and b directly is often a very difficult task, so the commonly used method is to introduce the Lagrangian factor and transform the problem into the situation shown in formula (5).

$$L(w,b,\alpha) = \frac{1}{2}\left||w|\right|^2 + \sum_{i=1}^{m}(1 - y_i(w^T x_i - b))(5)$$

In the above equation, the partial derivative of w and b is converted into its dtablual problem.

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j \; y_i y_j x_i{}^T x_j (6)$$

$$s.t. \sum_{i=1}^{m} \alpha_i \, y_i = 0 \; \alpha_i \geq 0, i = 1,2, \ldots m \qquad (7)$$

Finally, to solve α, you can find the value of w, b.

Apply the SVM algorithm to multiple dimensions, then transform the problem into a question that searching the largest classification interval to find the optimal hyperplane. Shown as Fig 5.

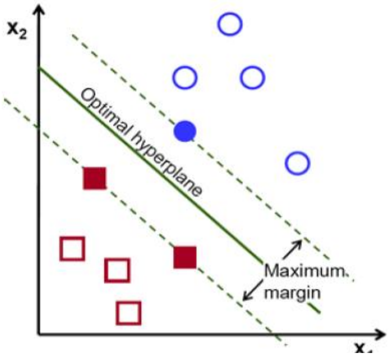

Fig 5: Multidimensional SVM thinking

3.2 Support Vector Machine linearly inseparable

SVM is designed to solve linear separable problems. The quantity of support vectors determines the algorithm complexity of the trained model. So SVM is not easy to produce overfiting; the support vector determines the model trained by svm. Even if all the non-support vector points in the training set are removed and the training process is repeated, the result is still the same model and if several support vectors acquired by a SVM training is relatively small, the model trained by the SVM is easier to be sublimated. However, data tends to be linearly inseparable. That is, the corresponding vector of the data set in space cannot be separated by a hyperplane. As shown below, see in Fig 6.

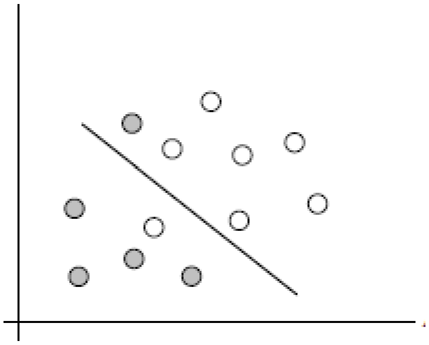

Fig 6: The corresponding vector of the data set

For the nonlinear classification problem in the input space, it can be transformed into a linear classification problem in a certain dimensional feature space by nonlinear transformation, and a linear support vector machine can be learned in a high-dimensional feature space. In the dual problem of linear support vector machine learning, the objective function and the classification decision function only involve the inner product between the instances, so there is no need to explicitly specify the nonlinear transformation, but the kernel function is used to replace the inner product. The inner product between two instances is the representation after the linear transformation of the kernel function.

For this type of problem, the solution given by SVM is: find a certain function. In the calculation of the classification algorithm, these functions map the data that is indivisible in the low-dimension to the high-dimensional, forming a linearly separable case. So for the linearly inseparable case, turn the problem into:

$$f(x) = w^T \emptyset(x) - b = \sum_{i=1}^{m} \alpha_i y_i \, k(x_i, x_j) - b\mathrm{m} \tag{8}$$

In the above formula, $\mathrm{k}(\mathrm{x}_\mathrm{i}, \mathrm{x}_\mathrm{j})$ is the given kernel function.

So, in the case of linear inseparability, the main task is to design the kernel function. Researchers have proposed many methods for kernel functions. The following two are used in the design of classifiers.

(1) Linear kernel: This kernel function is mainly used in the case of linear separability. It is characterized by a small number of parameters and a fast processing speed. For general linearly separable data, the classification effect reaches an ideal state.

$$K(x_i, y_i) = < x_i, y_i > \tag{9}$$

(2) Radial Basis Function (RBF): In support vector machines, the radial basis kernel function is the most commonly used method. Its expression is:

$$K(x_i, y_i) = exp(-\gamma ||x_i - y_i||)^2 \tag{10}$$

IV. EXPERIMENTAL DESIGN

4.1 Experimental Data

The experimental data is derived from archive text data. It crawled through the data of Gansu Provincial Department of Bureau Administration in the form of reptiles. Through the form of data cleaning, the file text data was structured and expressed in a form that conforms to the characteristics of the file text.

The dissertation separately classifies the data of five different categories of archives and texts, including the departments of the Public Security Department, the Department of Health, the Commission for Discipline Inspection, the Department of Transportation, and the Office of People and Society.

4.2 Design of Stop Word List

Because the unique features of the archive text data itself are different from other ordinary texts, the traditional stop word list is not adopted in the text preprocessing stage. Combined with the nature of the archive text itself, this article will design a proprietary suspension. This table, part of which is shown in TABLE I:

TABLEI. Stop word list

1	about
2	notice
3	is
4	.

5	,
6	file
7	title
8	time
9	work
10	report
11	opinion
12	and
13	regulations
14	decide
15	speech
16	government
17	office

4.3 Evaluation Standard

As for the criteria for judging the file text data classification results, the paper uses traditional methods to calculate the classification accuracy p, recall rate R and F1 values, and compares the time efficiency when classifiers are designed using different kernel functions.

$$P = \frac{correctclassification}{actualclassification} \times 100\% \qquad (11)$$

$$P = \frac{correctclassification}{Thenumberofclassifica \quad tionsdue} \times 100\% \qquad (12)$$

$$F_1 = \frac{2PR}{p+R} \times 100\% \qquad (13)$$

4.4 Experimental Results

The file text data is classified and processed using linear kernel functions and support vector machines with radial basis kernel functions. The results are shown in TABLE II and TABLE III.

TABLE II. Support vector machine (Linear)

	P	R	F1	Time(s)
Public Security Department	90.3%	93.5%	91.4%	333
Health Department	82.5%	86.5%	88.7%	346
Commission for Discipline Inspection	88.5%	86.8%	84.5%	330
Transportation Department	91.6%	90.8%	91.2%	298
People Club Department	90.8%	92.5%	91.6%	305

TABLE III. Support vector machine (RBF)

	P	R	F1	Time(s)

Public Security Department	86.5%	88.9%	88.7%	380
Health Department	87.6%	83.9%	85.7%	389
Commission for Discipline Inspection	90.3%	86.5%	88.2%	368
Transportation Department	93.2%	92.5%	89.2%	353
People Club Department	92.5%	94.3%	93.4%	309

The P, R, and F1 values of the results obtained from the two experiments were compared to obtain Fig 7, Fig 8, and Fig 9.

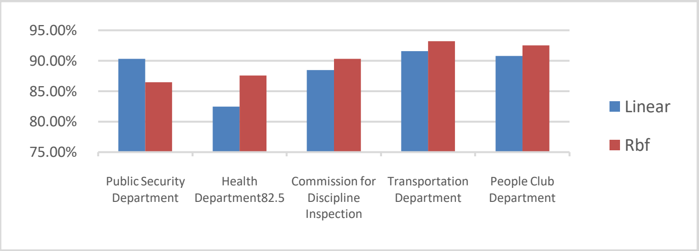

Fig 7: P value

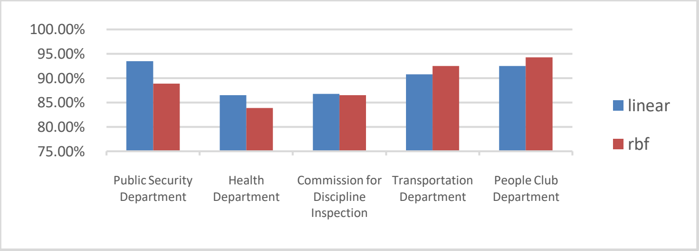

Fig 8: R value

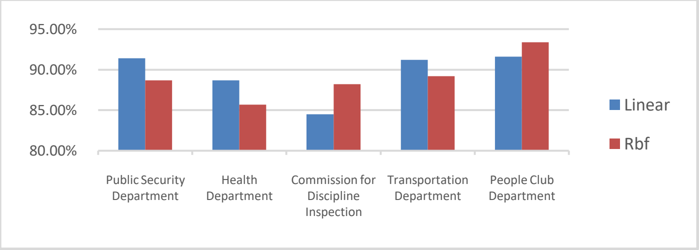

Fig 9: F1 value

Finally, in the classification process of the two algorithms using the kernel function, a comparison is made between the two time rates, and the results shown in Fig 10 are obtained.

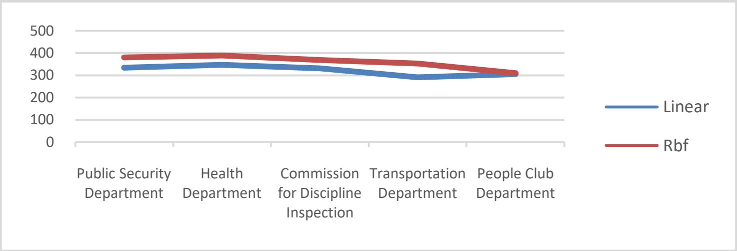

Fig 10: Comparison of two kernel function rates

V. EXPERIMENTAL SUMMARY

In the process of classifying archive text data, support vector machine algorithms are designed using different kernel functions. The resulting classification results are generally similar. The accuracy, recall, and F1 value of text data classification are approximated result. For the time efficiency of the two kernel methods, the linear kernel function is faster than the radial basis method. This is precisely because the choice of parameters in the linear kernel function is less than that of the radial basis kernel function. This leads to a faster processing of linear kernel functions.

VI. ACKNOWLEDGEMENTS

This work was supported by Northwest MinZu University the central university of basic scientific research business funds (No.31920190059).

REFERENCES

[1] Luhn H. P. (1958). Luhn, h.p.: the automatic creation of literature abstracts. ibm journal ofresearch and development 2(2), 157-165. Ibm Journal of Research & Development, 2(2): 159-165.

[2] Joachims T (1997) Text categorization with support vector machines. Fakultäten.

[3] Mccallum A, Nigam K (1998) A comparison of event models for naive bayes text classification.

[4] Chen Y (2015) Convolutional Neural Network for Sentence Classification.

[5] Johnson R, Zhang T (2014) Effective use of word order for text categorization with convolutional neural networks. Eprint Arxiv.

[6] Omara I, Wu XH, Zhang HZ, et al. (2017) Learning pairwise SVM on deep features for ear recognition. Proceedings of the 2017IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS).

[7] Zhu H, Liu XX, Lu RX, et al. (2017) Efficient and privacy preserving online medical prediagnosis framework using nonlinear SVM. IEEE Journal of Biomedical and Health Informatics.

[8] Christina M. Neubig, Liesbet Vranken, Jutta Roosen et al. (2020) Action-related information trumps system information: influencing consumers' intention to reduce food waste. Journal of Cleaner Production 261.

[9] Ye BL, Xue LY, Fang YL et al. (2020) Quantum coherence and quantum Fisher information in the XXZ system. Physica E: Low-dimensional Systems and Nanostructures 115.

[10] Hsieh CH, Shen C, Chen CC, et al. (2020) Knowledge-based system for resolving design clashes in

building information models Automation in Construction 110.

[11] Li DF, Hu Y, Lan MM (2020) IoT device location information storage system based on blockchain Future Generation Computer Systems 109.

[12] Chung KF, Cheung R, Li V, et al. (2015) Resource management using information system in Chinese medicine service. European Journal of Integrative Medicine 7.

[13] Yang L, Xu WH, Zhang XY, et al. (2020) Multi-granulation method for information fusion in multi-source decision information system. International Journal of Approximate Reasoning 122.

[14] "Natural Language Processing and ChineseComputing", Springer Science and Business Media LLC. 2018.