Research and Implementation of Text Information Mining Technology Based on HMM

Xiaohui Wang [1, a], Wei Wang [2, b], Lu Li [3,c], Yang Zhang [2,d], Wei Yao [2,e], Lei Guo [2,f], Peng Sun [2,g]

[1]State Grid Henan Electrical Power Company, Zhengzhou 450000, China;

[2]State Grid Henan Electrical Power Company Electrical Power Research Institute, Zhengzhou 450052, China;

[3]State Grid Henan Province Power Company Maintenance Company, Zhengzhou 450052, China

[a]lydlwxh@sohu.com,[b]ww7839@163.com

Abstract

With the development of society, text information mining technology has been widely used in many fields.Hidden Markov (HMM)-based text information mining technology has achieved its role and value in many fields, but so far, there is little research on text information mining technology in the power field.This paper studies the application of text information mining in power system as a research background, and explores the research and implementation of text information mining technology in power system based on HMM.First of all, text information mining technology needs to preprocess the text, including word segmentation, word frequency statistics, remove stop words, feature selection and feature extraction, text representation and text classification.Then master the HMM algorithm to encode text information mining.Text preprocessing is the most important part of text information mining, and it also lays a foundation for the realization of text information mining.

Keywords

Hidden Markov; power system; text preprocessing; text mining.

1. Introduction

With the rapid development of information technology, the amount of data accumulated in the power system is increasing. Behind the huge textual information data, due to the diversification of data structure and storage methods, the data used is only a small part, resulting in a large amount of valuable data being wasted [1].In the wasted data, the data in the form of text accounts for a large proportion. How to obtain the required text information from more complex text data has received widespread attention in the power industry. Taking maintenance and maintenance as an example, grid companies have accumulated a large number of inspection and test records, inspection and elimination records, fault and defect description reports and event sequence records (SOE) [2].These logs and reports are mainly in the form of Chinese short texts with mixed numbers and alphabetic symbols. They contain a wealth of equipment historical running status information, maintenance effect information and reliability information. Prefecture-level cities with industrial and commercial development generate millions of pieces of information every year, reaching the TB level [3].The analysis and utilization of these data is beneficial to objectively evaluate and predict the healthy development of the operating status of power equipment.

At present, the use of text information in the power system mainly focuses on the statistics and analysis of unstructured data, and carries out in-depth analysis and mining of the implicit value laws. Faced with structured data and a large amount of unstructured data of short text, natural language processing (NLP) technology and data mining extraction technology [4] can be used to extract entities and relationships, and then extract useful information. In order to improve the lean level of power system operation management, and to continuously increase the amount of unstructured text data, it

is necessary to further explore the value of potential text information in unstructured data. In addition, the text information mining technology and the power industry are in a stage of rapid development, and the application of text information mining will become more and more extensive. This technology will inevitably play an increasingly important role in the field of power systems. Therefore, it is particularly important to carry out research on the application of power big data text information mining technology. Based on the above content, it is of great significance to develop a data mining system based on short text.

In the field of power, some scholars have applied data fusion in condition evaluation, fault diagnosis, reliability prediction and test optimization [5].But so far, the integration of structured and unstructured data in the power system has not yet touched, that is, the fusion between life-time data [6].One of the important reasons is the lack of research on structured and unstructured data, and there are no technical ways and solutions for obtaining power text information. The inability to build a detailed power corpus reaffirms the importance of text information mining in the power system industry.

With the development of named entity research, a statistical model-based identification method has been developed. The more representative methods include Miller's HMM-based method, Borthwick, Hai Leong Chieu and other methods based on maximum entropy. Fuchun Peng's method based on conditional random field And Kazama and other methods using the support vector machine for text recognition [7].In this paper, the HMM-based method is used to mine text information, extract important information from the power system text, and identify the text.

2. Text mining technology

This paper studies text mining based on HMM and related principles in the field of power. Nowadays, HMM-based text information mining technology is gradually mature, but there is little research in this field of power industry. This paper deals with the text according to the basic process of text information mining. The research scheme mainly includes text acquisition, word segmentation, feature selection and extraction, and text classification. The text mining process of this paper is shown in Figure 1:

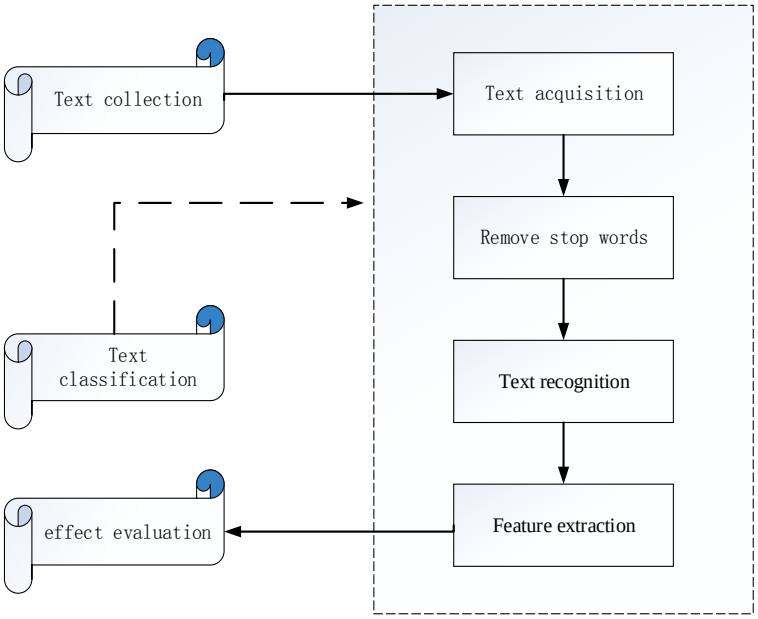

Figure 1 text mining process

2.1 Introduction to HMM Model

HMM is a statistical-based sequence analysis and learning model that was established by Baum et al. in the late 1960s and early 1970s [8].In recent years, it has been paid to the field of information technology, such as some improved models of speech recognition, natural language processing and text mining. There are corresponding applications in text representation, word segmentation, feature

selection and data mining involved in the classification process. The HMM information processing model is easy to set up, does not require large-scale dictionary sets and rule sets [9], and uses the probability distribution features of vocabulary for model building. It is easier to understand than artificial neural networks (Anmatic Neural Networks ANN) models. Partially reflects the semantic nature of the text.

Definition: A hidden Markov model is a triplet $\mu = (A, B, \pi)$. Including initialization probability vector, state transition matrix, confusion matrix. The number N of states in the model, the number M of different symbols that may be output from each state, the state transition probability matrix $A = \{a_{ij}\}$, among them, $a_{ij} = P(q_t = s_j | q_{t-1} = s_i), 1 \le i, j \le N$, $a_{ij} \ge 0$, $\sum_{j=1}^{N} a_{ij} = 1$,Observing the probability distribution matrix $B = \{b_j(k)\}$ of symbol v_k from state s_j ,among them, $\pi_i = P(q_1 = s_i), 1 \le i \le N$, $\pi_i \ge 0$, $\sum_{i=1}^{N} \pi_i = 1$.

In the HMM model, we do not know the specific state sequence of the model, only the probability function of the state, so the HMM is a double stochastic process [10].Among them, the state transition between models is concealed, and the observable random process is a random function of the hidden state transition process.

The basic principle of the HMM model is shown in Figure 2:

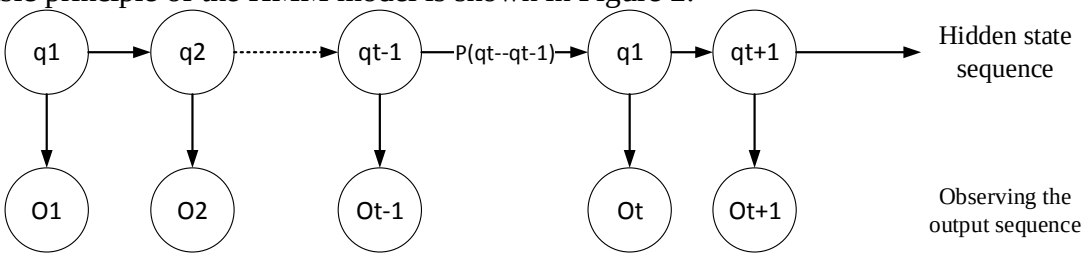

Figure 2 HMM diagram

The description of the HMM model can be used to solve three basic problems [11].The first is to evaluate problem: the probability (evaluation) of finding an observation sequence for a given HMM model; The second is the decoding problem: searching for the hidden state sequence (decoding) that is most likely to generate an observation sequence. The third problem is learning problem: generating a HMM (learning) for a given sequence of observations.

Evaluation:For such problems, we describe a hidden Markov model (that is, a collection of triads) and an observation sequence for different systems. We want to know which HMM is most likely to produce this given sequence of observations. A forward algorithm can be used to calculate the probability of an observed sequence after a given hidden Markov model, and thus the most appropriate hidden Markov model.

Decoding:The decoding problem is that a given sequence of observations searches for its most likely sequence of hidden states, that is, the search generates a hidden state sequence of the output sequence. For decoding problems, the Viterbi algorithm (Viterbi) can be used to determine the search for known observation sequences and the most likely hidden state sequences under the HMM.

Second, another broad application of the Viterbi algorithm is the part-of-speech tagging in NLP.In the process of part-of-speech tagging, each word in the sentence is an observation state, and the part of speech is a hidden state.For each word in each sentence, by searching for its most likely hidden state, we can find the most likely part of speech token state for each word in a given context.

Learning:The learning problem is to generate a hidden Markov model from the observed sequence. This is also the hardest one of the HMM-related issues. Based on an observation sequence (from a known set) and a set of hidden states associated with it, a most suitable hidden Markov model is estimated, i.e., the most appropriate triplet (p_i, A, B) for the known sequence description is determined.

Therefore, the problems usually solved by the hidden Markov model include the following three:

For a system that is most likely to match the observed sequence -the evaluation, solved using a forward algorithm;

For a sequence of observations that have been generated, determine the most likely sequence of hidden states - decoding, solved using the Viterbi algorithm;

For the generated observation sequence, determine the most likely model parameters - learning, using forward-backward algorithms.

2.2 Text acquisition

Text preprocessing begins with getting text, and text acquisition can use a breadth-first and depth-first algorithm to write crawlers to grab useful information. There are two kinds of reptiles, which are divided into vertical reptiles and general reptiles. Generally, general reptiles are used. The vertical reptiles mainly obtain the texts of related topics at the relevant sites, and the general reptiles have no restrictions on this. There are many open source crawler systems on the web (such as Python's Scrapy).Secondly, text can be obtained by means of file reading. In this paper, we use the specified text to mine text information.

2.3 Remove stop words

Remove stop words, the stop words mentioned here refer to words that are not needed in the text information mining process. These symbols and the expression of words in the text do not have any effect, but they exist in a large amount in the text to be processed, so they need to be removed. Secondly, many other words can be removed for some different texts, such as adjectives and so on, as well as other part of speech.

2.4 Text segmentation

The first pre-processing that we need to do after getting the text and removing it is to need to segment the words we want. The current participles are all based on statistical participles, most of which come from established corpora, and the participles recombine the required sequence of words according to certain rules [12].English letters and words are separated by spaces because they are separated by spaces, but natural language does not have a space character, so the word segmentation requires a special algorithm to solve.

For the implementation of word segmentation, we apply the second problem in the hidden Markov model [13], which is the problem of decoding. We will use the Viterbi algorithm to implement word segmentation. The Viterbi algorithm is a very special and widely used dynamic programming algorithm. It was originally proposed for the shortest path problem of the directed graph of the fence network (Lattice), which is the optimal path [14].As long as the problem that needs to be described in applying the implicit Markov model can be decoded by the Viterbi algorithm, including the participle algorithm we need to apply - the Viterbi algorithm. It also shows that the hidden Markov model is widely used. It also shows that the hidden Markov model is widely used. The following is a detailed introduction to how the Viterbi algorithm is applied to the implementation of word segmentation, and how to use dynamic programming to implement word segmentation.

First, define a Viterbi variable $\delta_t(i)$,The Viterbi variable is the maximum probability that the HMM will reach state s_i along a certain path at time t and output the observation sequence $O_1O_2\cdots O_t$:

$$\delta_t(i)=\max_{q_1q_2,\cdots,q_{t-1}} P(q_1,q_2,\cdots,q_t=s_i,O_1O_2,\cdots,O_t|\mu) \tag{1}$$

Similar to the forward variable, $\delta_t(i)$ has the following recursive relationship:

$$\delta_{t+1}(i)=\max_j\left[\delta_t(j)\bullet a_{ji}\right]\bullet b_i(O_{t+1}) \tag{2}$$

This recursive relationship allows us to apply dynamic programming search techniques. In order to record at which time the HMM reaches the state s_i by which of the most probable paths, the Viterbi

algorithm sets another variable $\psi_t(i)$ for the path memory, and let $\psi_t(i)$ record the previous one of the states s_i on the path status (at time t-1).

2.5 Feature Extraction

Feature extraction and feature selection are the most important part of the text information mining process, because it directly determines the category direction of the text to be mined. That is to lay a certain foundation for the classification of text. The objectives of feature selection are as follows:

1. Improve the accuracy of the forecast;

2. Construct faster, less expensive forecasting model;

3. Be able to have a better understanding and explanation of the model.

For text classification, the method of feature extraction is crucial, and it is generally the most common use of statistic (CHI) in power system text categorization [15].

The basic idea of CHI [16] is to determine the correctness of the existence of the theory by observing the deviation between the actual value and the theoretical value. The specific operation steps are to assume that the two variables are indeed independent of each other, and then observe the degree of deviation between the actual value and the theoretical value. If the deviation is small enough, then we think that the error is a natural sample error, because the measurement is not accurate enough or accidental, which proves that the two are indeed independent, then the original assumption is correct. However, if the deviation is so large that such an error must not be caused by accident or by measurement inaccuracy, we believe that the two are actually related, that is, to negate the original hypothesis and accept the opposite hypothesis.：

The traditional CHI statistical method assumes that the non-independent relationship between feature item *w* and category *c* is similar to the χ^2 distribution with one-dimensional degrees of freedom, and the CHI statistic for *w* for *c* can be calculated as:

$$\chi^2(w,c)=\frac{N(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \tag{3}$$

A represents the frequency of the document containing the feature item *w* and belonging to category *c*.

B represents the frequency of the document containing the feature item *w* but not belonging to category *c*.

C represents the frequency of the document belonging to category *c* but not containing feature item *w*,

D represents the frequency of the document that neither belongs to *c* nor contains feature item *w*.

N represents the total number of documents in the corpus.

The CHI statistical method is used to measure the degree of correlation between the feature item *w* and the category *c*. When the feature item *w* and the category *c* are independent of each other,$\chi^2(w\ c)$ 0.The stronger the correlation between the feature item *w* and the category *c*, the larger the value of $\chi^2(w\ c)$, and the more the identification information related to the category *c* contained in the feature item *w* at this time [17].

2.6 Improvement of feature extraction

Since the traditional CHI statistical method is a normalization process, only the number of occurrences of a feature item in all documents is calculated, and the number of times the feature item appears in a document is not calculated. Some feature items may appear less frequently in all documents, and appear more frequently in a document. The feature item may be processed according to the low frequency word, but the contribution rate of the feature item in this document is relatively high, which affects the accuracy of the classification. Therefore, we introduce parameter α

(frequency) to adjust, in order to solve the problem that CHI method is not reliable for low-frequency word statistics.

Multiply the parameter α in the formula (3), which effectively solves the low-frequency word problem and improves the accuracy of text classification.

In addition to dealing with low-frequency words, this paper also deals with negative correlations that do not contribute much to the classification. The correlation between the keyword characteristics in the feature selection process and the categories in the power system is nothing more than two types, one is positive correlation and the other is negative correlation. For the traditional CHI statistical method, in the above formula: If AD-BC>0, it means that the feature item is positively related to the relevant category of the power system. That is, the feature item appears that the description text may belong to a certain category of the power system. At this time, the higher the value of the CHI statistic, the more likely the text contains the feature item when it belongs to one of the categories. Conversely, if AD-BC < 0, the feature item is negatively correlated with the power system related category. That is, the feature item appears that the description text may not belong to a certain category of the power system. At this time, the higher the value of the CHI statistic, the more likely the text does not belong to a category when the feature item is included. Thus, a feature word appears less frequently in a certain type of document, but appears more frequently in the rest of the document. If calculated according to formula (3), the score obtained in this case is relatively large, but the contribution rate of the text classification is not large.

After a rigorous analysis of the above CHI statistical methods, we have modified the CHI statistical method of the above analysis, and discard the situation in which the feature items are negatively correlated with the categories. In the case of the previous CHI statistical method, we will improve the $x^2(w,c)=0$ in the case of $AD\text{-}BC \le 0$.The CHI statistic for a feature item is generally used as the average or maximum value of the CHI statistic for that category for all categories. Based on the modified CHI statistical method, the CHI value of the feature item w is specified as the sum of the CHI statistic of the feature item for all categories. That is to say, when calculating the CHI value of the feature word, we only need to consider the case where the feature item is positively correlated with the category. For other cases, the summation statistics are no longer considered.

The CHI feature extraction step is as follows:

(1) The total number of texts in the statistical sample set (N);

(2) Statistics on the frequency of occurrence of each word in the positive text (A), the frequency of occurrence of each word in the negative text (B), the frequency in which the positive text does not appear), and the negative text does not appear in frequency;

(3) Calculate the chi-square value of each word;

(4) Sort the results of each word according to the formula of the chi-square test from large to small, and then select the first n words as features, then n is the feature dimension.

3. Case analysis

3.1 Study design

This paper takes the transformer fault text data as the research object and mines the power system data. In this paper, a total of 35,289 pieces of data were collected and preprocessed by HMM. The data is extracted by CHI and the improved CHI (CHI_2), and the extracted data is classified into the support vector machine (SVM) and the k-nearest neighbor (KNN) classifier for classification.

The classification criteria need to be defined before text classification. The classification of fault texts by the power system is generally divided into “emergency”, “important” and “general”. The classification effect is represented by the accuracy rate, the recall rate and the F value. This paper also classifies the transformer fault text according to the classification standard.

3.2 Comparative analysis of examples

There are 2800 fault data after preprocessing by HMM. These data are extracted by CHI and CHI_2 respectively, and the extracted data are respectively classified into SVM and KNN classifier for classification.

First, compare the feature extraction of CHI and CHI_2 methods in KNN classifier. The results are shown in Table 1:

Table 1 Comparison of CHI and CHI_2 method feature extraction in KNN

	Accuracy	Recall rate	F value
CHI	0.68	0.72	0.74
CHI_2	0.68	0.75	0.81

It can be seen from Table 1 that the improved accuracy, recall rate and F value of the classification evaluation index of the feature extraction method CHI_2 are significantly higher than the CHI method. It shows that the feature extraction effect is obviously improved after the improvement of CHI. Next, the feature extracted data is put into the KNN classifier, and the result is shown in Figure 3:

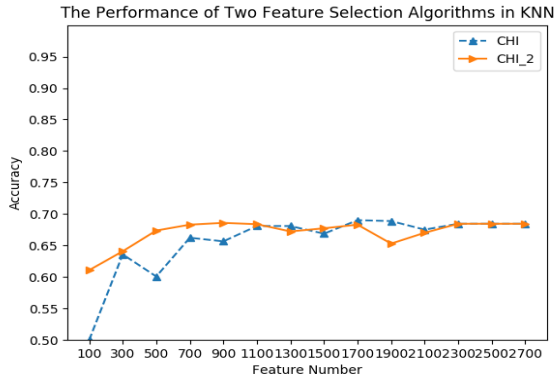

Figure 3 Classification accuracy of two feature selection methods on KNN

Figure 3 depicts the accuracy of the CHI and CHI_2 methods for classifying transformer fault text on the KNN classifier. It can be seen from the classification experiments conducted by selecting different number of feature items that when the feature items are less than 1100, the accuracy of the CHI method and the CHI_2 method is basically increasing, and the accuracy of the CHI method fluctuates greatly. Overall, CHI_2 is more stable than CHI. When the data set is larger than 1100, the accuracy of the two feature extraction methods tends to be stable, and the accuracy rate is about 0.7.

Then, compare the feature extraction of CHI and CHI_2 methods in SVM classifier. The results are shown in Table 2:

Table 2 CHI and CHI_2 method feature extraction in SVM comparison

	Accuracy	Recall rate	F value
CHI	0.79	0.81	0.89
CHI_2	0.79	0.72	0.81

It can be seen from Table 2 that the classification evaluation index (accuracy rate, recall rate and *F* value) of CHI_2 is significantly higher than that of CHI method, and the three evaluation index values of SVM are higher than KNN. It shows that after the improvement of CHI, the feature extraction effect is obviously improved, and the classification effect of SVM is better than KNN.Next, the feature extracted data is put into the SVM classifier, and the classification result is shown in Figure 4:

Figure 4 depicts the accuracy of the CHI and CHI_2 methods for classifying transformer fault text on the SVM classifier. It can be seen from Fig. 2 that the accuracy of the three feature selection methods in the SVM classification model is always increasing when 100 to 1300 feature items are selected.

When the selected feature items are more than 1300, the accuracy of the two methods is higher and tends to be flat, but the overall trend of CHI_2 accuracy is always higher than the CHI method. When the SVM is classified by the CHI method and the CHI_2 method, the final accuracy rate is 0.75 or more.

In summary, the improved feature extraction effect of CHI_2 is more reliable than the CHI method, and the classification effect of SVM classifier is higher than that of KNN classifier.

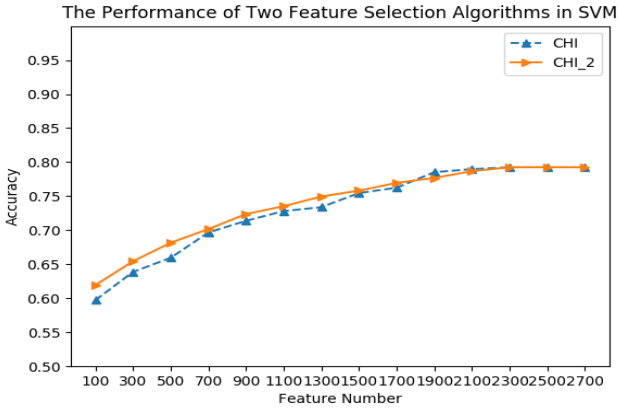

Figure 4 Classification accuracy of two feature selection methods on SVM

4. Conclusion

This paper is based on the HMM method for information mining of transformer fault texts, which has far-reaching significance for the research of information mining technology in text processing.

This paper improves the CHI's shortcomings to obtain the CHI_2 method, and puts the CHI_2 processed text into the classifier. The effect is better than the CHI method, which provides an effective method for feature extraction.

This paper compares the classification effects of KNN and SVM two classifiers, and obtains the SVM classification accuracy rate higher than KNN, which provides an important research idea for text classification to select the classifier direction.

Acknowledgements

This paper is supported by the Science and Technology Project of the State Grid Henan Electrical Power Company Electrical Power Research Institute - "Research on Transformer Operation Feature Extraction and Deep Fault Diagnosis Method".

References

[1] QU Chaoyang,Chen Shuai,Yang Fan,Zhu Li.A Method for Power Big Data Preprocessing Attribute Reduction Based on Cloud Computing Technology[J].Automation of Electric Power Systems,2014,38(08):67-71.

[2] China Electrical Engineering Society Informationization Special Committee. White Paper on China Power Big Data Development [M]. Beijing: China Electric Power Press, 2013.

[3] Luo Xueli,Xu Shuzhen,Wang Sen,Yang Li,Duan Jiajie.Study on Unstructured Data Retrieval of Power Enterprises[J].Computer and Digital Engineering,2014,42(04):729-733.

[4] Geng Xuefeng,Zhou Guodong.Research on Deep Learning for Natural Language Processing[J].Acta Automatica Sinica,2016,42(10):1445-1465.

[5] RUN C, WALTZ D, ANDERSON R N, et al.Machine learning for the New York City power grid[J].IEEE Trans on Pattern Analysis and Machine Intelligence, 2012.

[6] Zhang Dongxia,Miao Xin,Liu Liping,Zhang Yan,Liu Xue.Study on the development of big data technology in smart grid[J].Proceedings of the CSEE,2015,35(01):2-12.

[7] Huang Zhengwei,Tang Fangyan.Study on Garbage Text Recognition Based on SVM

Classification Model[J].Mathematics in Practice and Theory,2016,46(07):144-153.
[8] Li Xinde, Pan Jindong, DEZERT Jean. A Sequence Aircraft Target Recognition Algorithm Based on DSmT and HMM[J].Acta Automatica Sinica, 2014, 40(12): 2862-2876.
[9] Yu Tao,Zou Jianhua.Study on the Method of Gait Recognition Based on Bayes Rule Combined with HMM[J].Chinese Journal of Computers,2012,35(02):2386-2396.
[10] Qian Zhiyong,Zhou Jianzhong,Tong Guoping,Su Xinning.Study on Automatic Word Segmentation of Chu Ci Based on HMM[J].Library and Information Service, 2014, 58(04): 105-110.
[11] He Min, Peng Yuqian, Liu Hongli, Hu Jiusong. Robust User Behavior Recognition Based on Improved Hidden Markov Model[J].Journal of Hunan University(Natural Science), 2018,45(02):127-132.
[12] Liu Yonghua,Li Aiping,Duan Liguo,Geng Peng,Wang Hongxiang.Subjective text recognition featuring subjective clues[J].Computer Engineering and Design,2015,36(09):2572-2577.
[13] Yang Jian,Wang Haihang.Text classification algorithm based on hidden Markov model[J].Computer Applications,2010.
[14] Hu Zhewei,Guo Ruipeng,Lan Haibo,Liu Haitao,Sang Tiansong,Zhao Feng.An isolated island model for distributed power distribution networks based on directed graphs[J].Automation of Electric Power Systems,2015,39(14):97-104
[15] Li Jie,Li Huan.Study on feature extraction and sentiment classification of short text review products based on deep learning[J].Information Studies & Practice,2018,41(02):143-148.
[16] Xiong Zhongyang,Zhang Pengzhao,Zhang Yufang.Study on text classification feature selection method based on χ~2 statistics[J].Computer Applications,2015.
[17] Fan Cunjia,Wang Yousheng,Wang Yuting.An Improved CHI Text Feature Selection Method[J].Computer and Modernization,2016(11):7-11