Clustering Analysis Method of Ethnic Cultural Resources Based on Deep Neural Network Model

Mingjing Tang, Chao Sun(✉), and Li Liang

Yunnan Normal University, Kunming, China
1052962641@qq.com

Abstract. This paper proposes a method of clustering analysis of ethnic cultural resources based on deep neural network model. Firstly, the feature word extraction and vectorization of ethnic cultural resources texts are realized by doc2vec document vectorization tool. Then K-means clustering algorithm is used to cluster the ethnic cultural resources texts after vectorization, and the Elbow method is used to determine the best aggregation. So as to obtain the correlation between the texts of ethnic cultural resources, which is used for the collection, storage and intelligent service of massive ethnic cultural resources provides technical support. At the end of the paper, the ethnic cultural resources in the specific ethnic website are taken as an example to analyze the above methods.

Keywords: Deep neural network model · Ethnic cultural resources · Clustering analysis · Elbow method

1 Introduction

Ethnic culture is an indispensable part of the culture that has been deposited in China of five thousand years of history. However, with the continuous development of modernization and internationalization process of China, the protection and inheritance of ethnic culture, especially minority culture, faces a huge crisis. At present, the latest research on the protection of ethnic culture has the digital protection mentioned in the reference [1] and the reference [2]. For example, reference [3] have invented an algorithm that can effectively mine a large number of rock carving patterns; also the network and information protection mentioned in the reference [4] and the reference [5], such as the construction of various ethnic cultural websites. However, these methods all rely on the continuous collection of minority cultures. While in the face of massive ethnic cultural resources, they can't be quickly classification, collection and sharing, which is not conducive to the protection and inheritance of ethnic culture. Therefore, new theoretical guidance and tool support are urgently in need. The most ideal state of ethnic cultural resources management is to realize intelligent management of data resources [6], such as automatic collection, classification and sharing of cultural resources. Using distributed web crawling technology, natural language processing technology, text mining technology [7] to collect, parse, preprocess and other operations on the ethnic cultural resources text. And then the obtained text of ethnic cultural resources is analyzed by clustering so

X. Chen et al. (Eds.): ML4CS 2020, LNCS 12488, pp. 160–170, 2020.
https://doi.org/10.1007/978-3-030-62463-7_15

as to have a better understand on the deep semantics of ethnic cultural resources texts and obtain the association between ethnic cultural texts. All of these will help the automatic collection, identification and sharing of massive ethnic cultural resources, and provide technical support for the development of ethnic cultural resources in the direction of intelligent management.

The current text mining is mainly based on text feature extraction of deep learning. For example, the reference [8] introduces the related research of deep learning text extraction, and the reference [9] introduces the text mining technology that combines deep learning features. Text mining is widely used, For example, reference [10] analyzes the differences between Chinese and American science and technology policies Based on text mining and visual analysis; Reference [11] studies how the big data service of scientific and technological literature develops towards intelligent question answering, based on text mining technology; reference [12] research the automatic classification of product description based on text mining. Inspired by the reference [10] and the reference [12], this paper is based on the text data of ethnic cultural resources, for the purpose of to enhance the identifiability and comprehensibility of massive ethnic cultural resources and to facilitate the intelligent collection and sharing of massive ethnic cultural resources. This paper propos a clustering analysis method of ethnic cultural resources based on neural network probabilistic language model [13]. Firstly, the ethnic cultural resources are crawled through distributed web crawling technology. The ethnic cultural resources data are processed by natural language processing technology to preprocess and segment Chinese word. And based on document vectorization tool, the vectorization of ethnic cultural resources data is realized. Then, using appropriate clustering algorithm to cluster the vectorized ethnic cultural resources data, and using the Elbow Method to select the optimal cluster number; finally, analyze the clustering result to identify and discover the deep semantics and associations between the ethnic cultural resources Texts. Thus provide support for the mining of massive ethnic cultural resources. The specific workflow is shown in Fig. 1.

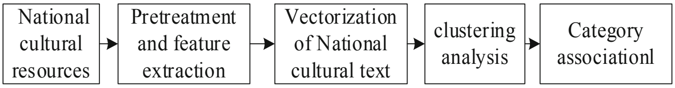
National cultural resources
Pretreatment and feature extraction
Vectorization of National cultural text
clustering analysis
Category associationl

Fig. 1. Ethnic cultural resources clustering process based on deep neural network.

The organization structure of this paper is as follows: Sect. 2 crawls on ethnic cultural resources and performs data preprocessing and vectorization representation; Sect. 3 describes the clustering process of ethnic cultural resources data and analyzes the clustering results; Sect. 4 takes the ethnic cultural resources data in the specific website as an example to verify the method of this paper. Finally, Sect. 5 summarizes the full text and future work prospects.

2 Ethnic Cultural Resources Data Vectorization

At present, the main mode of dissemination of minority cultures is the various cultural websites of ethnic minorities. The content of ethnic cultural resources of such websites is relatively scattered and difficult to be discovered and utilized. Using distributed network crawler technology, natural language processing technology and data mining technology, we can collect, analyze and preprocess the ethnic cultural resources.

2.1 Feature Extraction

The ethnic cultural resources crawled from the webpage are stored in the format of the HTML document. In order to extract the useful ethnic cultural resources texts in the HTML webpage, the relevant document parsing library needs to be called to delete the head and other unrelated areas of the obtained ethnic cultural resources data text and conduct preprocessing operations such as unlabeling, so as to extract the text content of the webpage. Then use natural language processing tools to remove prepositions, adverbs and other meaningless words, retaining entity words such as verbs and nouns. The specific process is described as follows:

Algorithm 1. Feature word extraction algorithm for ethnic cultural resources

```
Input: ethnic cultural resources text
Output: feature word set $\varphi(s)$
Set $\varphi(s)=\emptyset$;
Scrapy (html);    // Climb the ethnic culture related pages;
BeautifulSoup(P);    // Extract node element text content;
for(p in P) {    // traverse for each text content;
    wordFilter(p);    //Remove stop words with less meanings such as prepositions, adjectives and adverbs, and retain vocabulary such as verbs and nouns;
    reductWord(p);    //Convert the vocabulary of various tenses into a general form, and make a part of speech reduction;
    add($\varphi(p)$);    //Add to feature set
}
end
```

First, Crawl the ethnic culture related webpage from the ethnic culture related website and save it, then extract the ethnic culture related articles in the node from the saved webpage text and save each article as a line, then traverse all the ethnic culture articles separately. Remove prepositions, adjectives, adverbs and other stop words in the article, retain verbs, nouns and other entity words. Finally, the tense of the vocabulary is transform into a general form and add to the feature set.

2.2 Ethnic Cultural Resources Text Vectorization

After completing the extraction of ethnic cultural resources, in order to measure the similarity between texts and then realize the cluster analysis of ethnic cultural resources texts [14]. It is necessary to vectorize each document information of ethnic cultural resources. Doc2vec is an unsupervised learning algorithm [15], which is used to predict a vector to represent different documents, it mainly adopts two models: Distributed Memory (DM) and Distributed Bag of Words (DBOW). The Distributed Memory Model works by predicting the current word based on its context information. Each paragraph is represented by a vector that represents a column vector in the paragraph matrix, each word is represented by a vector that represents a column vector in the word matrix, the paragraph vector and the word vector are averaged or connected to predict the next word in the context. The working principle of Distributed Bag of Words model is the same as that of Distributed Memory model, but predicts the context probability based on the target word, ignoring the input context word, Let the model predict a random word in the paragraph. Specifically, in each iteration of the random gradient descent, a window is sampled from the text, and then a word is randomly sampled from the window, and a classification task [16] is formed according to the paragraph vector. In this paper, the doc2Vec DBOW model is used to characterize the ethnic cultural resources characterized by feature words. The specific process is shown in Fig. 2.

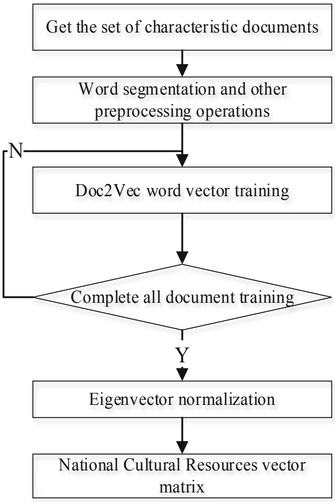
Get the set of characteristic documents
Word segmentation and other preprocessing operations
N
Doc2Vec word vector training
Complete all document training
Y
Eigenvector normalization
National Cultural Resources vector matrix

Fig. 2. Vectorization process of ethnic cultural resources.

As can be seen from Fig. 2, the process of ethnic cultural resources vectorization is as follows:

Firstly, the data crawling operation of the above section can obtain a feature document set representing the ethnic cultural resources, and perform data pre-processing operations such as deleting stop words and word segments.

Then, based on the doc2vec document vector tool, the Distributed Bag of Words Model is constructed to train the document vector of ethnic cultural resource document set. In order to avoid the impact of content size in the text of ethnic cultural resources on the value of feature vectors, it is necessary to normalize the feature vectors of each document:

$$\bar{\Delta}x(i,j) = \Delta x(i,j)/\|\Delta x(i)\|2 \tag{1}$$

The ethnic cultural resources feature vector can be normalized to the [0, 1] by formula (1).

Finally, the characterization of the ethnic cultural resources vector matrix can be obtained. Assume that the entire ethnic cultural resources have n texts, and after vectorization, the following ethnic cultural resources vector matrix is obtained:

$$\delta i = [\delta i(1), \delta i(2), \cdots \delta i(n)] \tag{2}$$

According to Formula (2), $\delta i \in R^{n \times m}$, n is the number of texts of ethnic cultural resources, and m is the number of characteristic words of ethnic cultural resources.

3 Ethnic Cultural Resources Clustering

Clustering is one of the important research fields in data mining and pattern recognition and so on. It plays an important role in identifying the internal structure of data [17]. After the above ethnic cultural resources texts are vectorized, the feature vectors corresponding to each ethnic cultural text can be obtained. And then the similarity between the texts can be measured by the clustering algorithm to realize the association and differentiation of ethnic cultural resources. The clustering algorithm used in this paper is K-means clustering algorithm, and the Elbow Method is used to evaluate the clustering effect. The specific workflow is shown in Fig. 3.

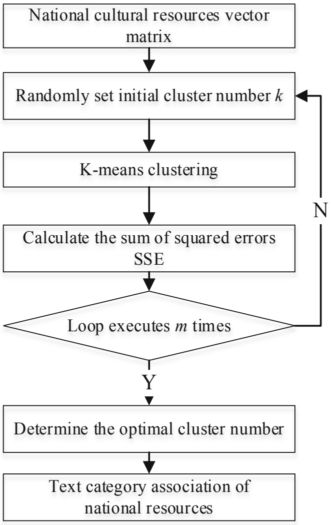
National cultural resources vector matrix
Randomly set initial cluster number k
K-means clustering
Calculate the sum of squared errors SSE
Loop executes m times
N
Y
Determine the optimal cluster number
Text category association of national resources

Fig. 3. Cluster analysis process of ethnic cultural resources text.

3.1 K-Means Clustering Algorithm

K-means clustering algorithm [18, 19] as a classical clustering analysis algorithm, which has the advantages of simple implementation, fast clustering and so on. The main principle of this algorithm is to divide samples into k clusters according to the distance between samples. Taking the ethnic cultural resources sample set as an example, the specific steps are as follows: First, *k* points are randomly selected from the vector set space as the initial cluster center. Then, the vector of ethnic culture and its nearest clustering center are classified into one category, and Then calculate the average of all vectors of each cluster to update the values of each cluster center. Finally, the above two steps are continuously iterated until the cluster center no longer changes to obtain the final clustering result [20]. The specific algorithm is described as follows:

Algorithm 2 . Ethnic cultural resources text vector clustering algorithm

```
Input: ethnic cultural resources text vector set, cluster number $k$
Output: $k$ clusters
1. Randomly select $k$ vectors in the text vector set of ethnic cultural resources as the initial clustering center of $k$ clusters;
2. For any vector, calculate its similarity with $k$ cluster centers by Euclidean distance method, and mark the category of vector as the cluster corresponding to the nearest cluster center;
3. Update the cluster centers corresponding to the $k$ clusters by calculating the average value of each cluster;
4. Iterate over the above two steps until the center point no longer changes.
5. end
```

3.2 Evaluation of Optimal Cluster Number Based on Elbow Method

Since the K-means clustering algorithm is an unsupervised learning task, the iterative method is used, and only the local optimal solution can be obtained. The selection of *k* is not easy to grasp, so the clustering effect needs to be evaluated. In this paper, the Elbow Method is used to evaluate the clustering effect of ethnic cultural resources [21].

Calculate the Sum of Squared Errors (SSE). The sum of squared errors of the text vector of ethnic cultural resources is calculated by formula (3):

$$SSE = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2 \tag{3}$$

Where, k represents the number of clusters, c_i is all elements in the i-th cluster, m_i represents the cluster center of the i-th cluster, and p represents each element in the cluster.

The Relationship Between K Value and SSE Value (Elbow Shape). As shown in Fig. 4, in order to obtain the best clustering effect, the k-means clustering algorithm needs to be repeated n times and the error sum of squares needs to be calculated each time, and obtain the correspondence between K values and SSE values. Draw the relationship between K value and SSE value (elbow shape), the K value point corresponding to the elbow portion ($k = 4$) in the relationship graph is selected as the optimal cluster number.

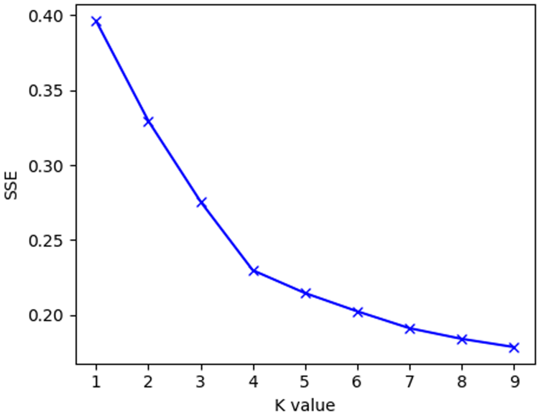

Fig. 4. K value and SSE value.

4 Experimental Analysis

In the experimental part of this paper, five virtual machines will be created by using virtualization technology in the OpenStack environment, one of which is the Master node and the other four are the slave nodes. Install the crawler module Scrapy on each slave node to implement the main function of resource crawling; The Redis database is installed in the master node to realize URL queue management and maintenance of multiple distributed crawlers. Scrapy is a screen grabbing and web grabbing framework developed by python, which is used to grab web sites and extract structured data from pages [22].

4.1 Experimental Data and Steps

The experimental data in this paper were obtained by crawling articles related to ethnic culture from the Ethnic Affairs commission of the People's Republic of china (http://www.seac.gov.cn) and Ethnic network (http://www.minzu56.net), and a total of 6830 ethnic cultural resources web pages have been obtained. The specific experimental process includes: ethnic cultural resources data crawling, ethnic cultural resources text preprocessing, ethnic cultural resources text vectorization, ethnic cultural resources text clustering, identifying and discovering the deep semantic and relationship between the texts of ethnic cultural resources according to the clustering results.

The programming language used in the experimental part of this article is the Python language, and all experimental steps are completed using a Python-based development library. For example: The crawler module Scrapy is used to realize the crawling of national cultural resources; Use Python's natural language development library NLTK to complete the ethnic cultural resources text preprocessing. In the vectorization stage of ethnic cultural resources text, the doc2vec development interface is called by Python's gensim library to construct a text vector training model, model training and feature extraction for ethnic cultural resources texts, and automatic mapping to *k* resource category clusters through unsupervised learning. In the text processing stage, a total of

6,830 ethnic cultural resources webpage texts were generated, which resulted in 6830 feature vectors in the text vectorization stage. The K-means clustering algorithm was implemented by Python's sklearn library to clustering the ethnic cultural resources text vector matrix.

4.2 Analysis of Experimental Results

The relationship between the sum of the squared errors (SSE) and the cluster number k is shown in Fig. 5. For the more accurate evaluation, we also draw the change of the Silhouette Coefficient [23] as a reference. We can see from the figure that when $k = 7$, the SSE polyline chart shows a larger inflection point (elbow). At the same time, the Silhouette Coefficient also reaches the maximum at this point. Therefore, $k = 7$ is the choice of the optimal cluster number k value. As shown in Fig. 6, 6830 texts in the ethnic cultural resources data are mapped to 7 ethnic culture types, of which type 7 contains the most text, accounting for 1782 texts; type 1 contains the least text, accounting for 423 texts. So far, the discovery and identification of the relationship between the texts of ethnic cultural resources has been completed.

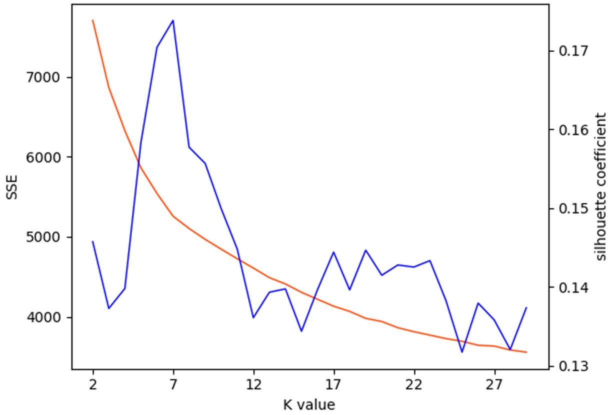

Fig. 5. The relationship between SSE and Silhouette Coefficient and K.

Due to the high dimension of text vector matrix of ethnic cultural resources, it is not convenient for data visualization analysis [24]. In this paper, Principal Component Analysis (PCA) is adopted to decrease the dimension of the clustering results of the text of ethnic cultural resources, and calls Python's painting library matplotlib, Visualize the effect of clustering. The clustering effect is shown in Fig. 7. It is better to divide the ethnic cultural resources into seven different types.

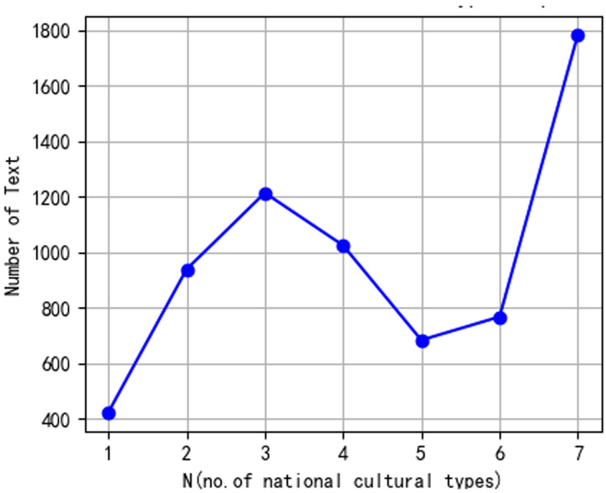

Fig. 6. Number of text and cultural type.

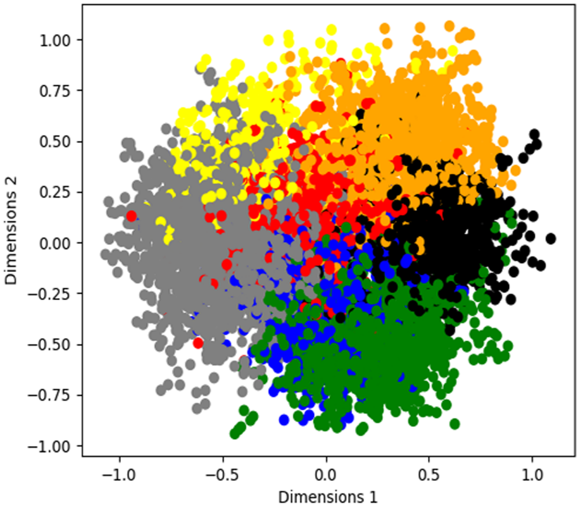

Fig. 7. Clustering effect of ethnic cultural resources text.

5 Conclusion

In this paper, a clustering analysis method of ethnic cultural resources based on deep neural network model is proposed, which adopts the unsupervised text vector representation method (doc2vec) based on deep neural network, using cosine distance formula to calculate the topic similarity between vectors effectively reduces the dimension of vector space and improves the efficiency of training, which is conducive to the mining of massive ethnic cultural resources texts. At the same time, the K-means clustering algorithm is used for text clustering, and the clustering effect evaluation method combining

Elbow Method and Silhouette Coefficient is adopted to select the optimal clustering number k value, it solves the problem that the number of clusters k in K-means algorithm is difficult to be determined. Finally, through experimental analysis, the method has high accuracy and efficiency, and effectively divides ethnic cultural resources into different categories. And realizes the differences and associations between ethnic cultural resources, revealing the deep semantics of ethnic cultural resources, and providing support for the massive ethnic cultural resources mining and intelligent services. In the next step, we can apply the clustering analysis method of ethnic culture based on the deep neural network model to the construction of the intelligent management system of ethnic culture, and realize the automatic collection, classification and sharing of ethnic culture resources.

Acknowledgment. This work was supported by the National Natural Science Foundation of China (Grant No. 61662085), Natural Science Foundation of the Department of Education of Yunnan Province of China (Grant No. 2017ZZX073), and Program for innovative research team (in Science and Technology) in University of Yunnan Province.

References

1. Xiao, W., Lu, Y.: Digital protection analysis of national traditional culture based on big data. Inf. Commun. **2019**(05), 177–178 (2019)
2. Liu, X.C., Song, W.: Research on digital protection of national traditional culture under the condition of big data. J. Cent. Univ. Natl. (Nat. Sci. Ed.) **25**(03), 44–49 (2016)
3. Zhu, Q., Wang, X., Keogh, E., Lee, S.-H.: An efficient and effective similarity measure to enable data mining of petroglyphs. Data Min. Knowl. Disc. **23**(1), 91–127 (2011). https://doi.org/10.1007/s10618-010-0200-z
4. Mei, H.: Research on the information service mode and document information resource construction of national libraries under the network environment. Commun. Res. **1**(08), 176 (2017)
5. Kummer, T.-F., Leimeister, J.M., Bick, M.: On the importance of national culture for the design of information systems. Bus. Inf. Syst. Eng. **4**(6), 317–330 (2012). https://doi.org/10.1007/s12599-012-0236-2
6. Manjunath, T.N., Ravindra, S.H., Umesh, I.M., Ravikumar, G.K.: Realistic analysis of data warehousing and data mining application in education domain. Int. J. Mach. Learn. Comput. **2**(04), 419–422 (2012)
7. Sun, Y.: A text mining approach to analyze public media science coverage and public interest in science. Int. J. Mach. Learn. Comput. **4**(06), 496–500 (2014)
8. Liang, H., Sun, X., Sun, Y., Gao, Y.: Text feature extraction based on deep learning: a review. EURASIP J. Wirel. Commun. Network. **2017**(1), 211 (2017). https://doi.org/10.1186/s13638-017-0993-1
9. Chen, X.: The feature extraction of the text based on the deep learning. In: Advanced Science and Industry Research Center. Proceedings of the 2014 International Conference on Network Security and Communication Engineering (NSCE 2014). Advanced Science and Industry Research Center: Science and Engineering Research Center, 2014, no. 5 (2014)
10. Wu, Y., Yuan, Y., Gong, Z.D.: A comparative study of Sino-US science and technology policies under the background of artificial intelligence——based on text mining and visualization analysis. J. China Acad. Electron. Sci. **14**(08), 891–896 (2019)

11. Wen, Y.K., Wen, H., Qiao, X.D.: Let knowledge generate wisdom——study on text mining and question answering technology based on artificial intelligence. J. Inf. **38**(07), 722–730 (2019)
12. Lee, H., Yoon, Y.: Engineering doc2vec for automatic classification of product descriptions on O2O applications. Electron. Commer. Res. **18**(3), 433–456 (2017). https://doi.org/10.1007/s10660-017-9268-5
13. Xiong, F.L., Deng, Y.H., Tang, X.W.: The core architecture of word2vec and its application. J. Nanjing Normal Univ. (Eng. Technol. Ed.) **15**(01), 43–48 (2015)
14. Rajhans, M., Pradeep, K.: Clustering web logs using similarity upper approximation with different similarity measures. Int. J. Mach. Learn. Comput. **2**(03), 219–221 (2012)
15. Wang, C.C.: A model of chinese sentiment analysis with more general applicability. In: Information Engineering Research Institute (USA), Asia Pacific Human-Computer Interaction Research Center (Hong Kong). Proceedings of 2018 4th ICMSMA International Conference on Advances in Intelligent Information Technologies (ICAIIT 2018). Information Engineering Research Institute (USA), Asia Pacific Human-Computer Interaction Research Center (Hong Kong): Intelligent Information Technology Application Society (2018)
16. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents (2014)
17. Mohamed, N.E., Monzer, M.Q.: Analysis of some algorithms for clustering data objects. Int. J. Mach. Learn. Comput. **4**(01), 99–105 (2014)
18. Sun, J.G., Liu, J., Zhao, L.Y.: Research on clustering algorithms. J. Softw. **2008**(01), 48–61 (2008)
19. Liu, P., Teng, J.Y., Ding, E.J., et al.: Spark-based large-scale text k-means parallel clustering algorithm. Chin. J. Inf. Sci. **2017**(04), 150–158 (2017)
20. Yogi, W.R., Devi, F.: The comparative study on clustering method using hospital facility data in Jakarta District and surrounding areas. Int. J. Mach. Learn. Comput. **9**(06), 749–755 (2019)
21. Kanungo, T., Mount, D.M., Netanyahu, N.S., et al.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)
22. Ryan, M.: Python Network Data Collection. People's Posts and Telecommunications Press, Beijing (2016). 2016.3
23. Sarunya, K.: A novel outlier detection applied to an adaptive k-means. Int. J. Mach. Learn. Comput. **9**(05), 569–574 (2019)
24. Supaporn, B., Thuttaphol, I., Nittaya, K., Kittisak, K.: Text-independent speaker identification using deep learning model of convolution neural network. Int. J. Mach. Learn. Comput. **9**(02), 143–148 (2019)