An Optimized K-means Algorithm for Text Clustering

Jiani Zhao[1,2,*]

[1] *School of Information Management & Engineering, Shanghai University of Finance and Economics, Shanghai, China*

[2] *Air Transportation College, Shanghai University of Engineering Science, Shanghai, China*

Abstract

In the process of data mining, the two major problems confronted by K-means clustering analysis are the determination of the initial cluster center and the valuing of k. The traditional K-means algorithm has obvious subjectivity in the above-mentioned two aspects, which will directly affect the clustering effect. In this paper, an analysis method combining relational matrix and degree centrality is proposed to determine the initial center point and the k value of K-means algorithm. The improved K-means algorithm is applied to the clustering analysis of the Chinese entrepreneurial policy text collection, and the clustered topic effects are visually displayed through the word cloud graphs. This empirical analysis not only verifies its effectiveness and objectivity for the improved algorithm in processing large clusters of long text document clusters with random unknown number of categories and category topics, but also provides an approach for the objective classification of Chinese entrepreneurial policy text collections in the meanwhile.

Keywords: *K-means, clustering, text mining, relational matrix, degree centrality*

I. Introduction

The Internet has developed rapidly in recent years, and the amount of data of various types of text documents encountered in our life has also increased rapidly. We can use text or document clustering technology to organize large-scale text information, extract their important features and find valuable information, thus reducing the workload of manually sorting out considerable documents and improving the efficiency of document retrieval. The currently common text clustering analysis is mainly to use the relevant model algorithm to classify the text collections of the known category number, or regulate the number of categories artificially or objectively and then conduct clustering analysis. However, the document data set we will analyze is likely to be a messy document data set, such as a policy document data set in practice. The document subject is unknown, and the number of document subjects is unknown. If people subjectively specify the number of topics in the document data set, the subjective cognition of the researcher on the document data set will greatly affect the clustering effect. Therefore, it is difficult for traditional document clustering algorithms to obtain objective and accurate results in reality. In view of this, this paper makes related improvements to the two parts of clustering method and index parameters on the basis of traditional text clustering. The purpose of doing this is to conveniently and quickly perform clustering analysis on the number of unknown topics and the messy document data sets of unknown topics.

K-means algorithm is one of the most common algorithms in clustering analysis. It is a typical clustering method with no supervision, so there is no need to label by artificial experts, which greatly reduce the clustering cost. However, while having these merits, K-means algorithm has also some defects and shortcomings. For example, the traditional K-means algorithm has great randomness to the selection of the initial clustering center, and the K-means algorithm has a high degree of dependence on the selection of the initial clustering center. In addition, the choice of value is often based on the personal experience of the researcher, and the choice of value has a great influence on the final clustering result. When different values are selected for the same sample data set, the clustering results may be very different, and the final result will be greatly affected. Therefore, the selection of the

value ultimately determines the quality of the clustering result to a large extent.

Based on the above analysis, the research makes improvement to the above-mentioned problems on the basis of the traditional K-means algorithm. A new method of selecting the initial clustering center is proposed, and the method is combined with the Degree Centrality in the relationship matrix and social network analysis to confirm the initial clustering center. After that, the iterative algorithm is used to confirm the final clustering class number, so as to eliminate the errors brought by artificial estimation, thus promoting the clustering effect and correctness. In the end, we use the improved K-means algorithm to do clustering analysis for the 462 policy texts of the Chinese entrepreneurship policy text collections, and the phrase cloud pictures are used to inspect the clustering effect. The experimental result tell us that the improved K-means algorithm has achieved good results in processing long text document clustering with large groups of random unknown number of categories and category topics.

II. Literature Review

There are many existing data mining methods, including association rule analysis, cluster analysis, outlier analysis, classification and so on. Among them, clustering, a classification technique of unsupervised learning, is a process of dividing physical or abstract sets into similar object classes, so that objects in the same cluster have a high degree of similarity, while objects in different clusters have great differences, so as to obtain the potential classification information in the data. Currently, the main clustering algorithms are: Partition-based method, hierarchy-based method, density-based method, grid-based method and model-based method. As a classical clustering algorithm based on partition idea, K-means algorithm has the advantages of simple structure, fast convergence speed and strong local search ability. At present, it has been widely used in many fields such as statistics, marketing, customer classification and so on. Many algorithms are innovated and expanded around it.

However, there are two main defects in traditional K-means clustering analysis: (a) In the K-means algorithm, the selection for initial center has a great influence on clustering results; (b) In the K-means algorithm, it needs to be given in advance k. Therefore, many scholars have studied the above problems and put forward a series of K-means optimization methods. Shi Z [1] selected the data as far away as possible as the initial clustering center, which solves the problem that the clustering results depend too much on the selection of the initial clustering center. Ghoul A E and Sahbi H [2] selected the clustering center through the median of data samples, which can weaken the influence of outliers on clustering results to a certain extent, thus solving the problem that traditional K-means algorithm is easy to fall into local optimum. However, the time cost of the algorithm is too large, which is not suitable for large-scale data clustering. Chen G C et al. [3] showed that the density sensitive similarity measure was used to calculate the object density to generate the initial cluster center, so as to optimize the stability of clustering results. Erzhou Zhua and Ruhui Ma b [4] proposed an optimization algorithm AFS-KM based on artificial fish swarm, in which attributes were weighted by information gain to calculate the distance between entities. Xie X et al. [5] found the farthest data object as the initial cluster center, and then splits the cluster, iterating repeatedly until finding a specified number of initial cluster centers. Tunali V et al. [6] considered maximum minimum distance algorithm and AP algorithm together to find the better number of clusters. Behera C K and Bhaskari D L [7] optimized K-means algorithm by using big data technology and combining map and reduce framework of Hadoop platform. Jiang X P et al. [8] put forward a new NDK-means method, which determined the effective density radius by standard deviation, and then selected representative sample points in high density area, and used these sample points as the initial clustering center. Zhou S B et al. [9] proposed a K-means CAN algorithm based on the sensitivity of results and local optimization, which used the intersection of sub-clusters of different clustering results to establish a weighted connected graph, and merged sub-clusters according to the connectivity of each node in the graph. Yu H T et al. [10] combined K-means algorithm with ant colony algorithm to improve the clustering quality. Chen G P et al. [11] used density method and the distance between objects to determine the initial clustering center, and selected the farthest k points in high density area as the initial clustering center. Yu S et al. [12] solved the nonconvex problem by using the kernel coefficient and

determined the number of clusters. Liu Y et al. [13] used the maximum-minimum algorithm for distance to select maximum distance to the point of high density from the selected center point as the initial centre of current.

III. Research Method

3.1. Traditional K-Means algorithm

According to the number of clustering K, K-means algorithm divides the existing data set into k clusters, and adopts the iterative updating method. In the first round, the object set is divided into k initial clusters according to the k initial center points randomly selected, and then the class to which each object belongs is iteratively re-divided according to the center of each cluster. The average value of each cluster will be used as the center point of the next iteration until the center point no longer changes, that is, the final clustering result is produced.

Input: a data set containing data objects and the number of clusters.

Output: clusters that meet the convergence of clustering criteria function.

Step 1: randomly select k data objects from data set D as initial clustering centers$kC_1, C_2, \dots \dots C_k$;
Step 2: Calculate the distance between each data object and k cluster centers, $d(x_i, c_j), i = 1,2, \dots, n; j = 1,2, \dots, k$, where $d(x_i, c_j)$ is commonly used, such as Euclidean distance or Manhattan distance;
Step 3: According to the calculated distance, divide each data object into the nearest cluster. That is, if $d(x_i, c_j) = min\{d(x_i, c_j), j = 1,2, \dots, k\}, x_i \in Y_j, Y_j$ represents the cluster with the cluster center c_j;
Step 4: Recalculate k clustering centers of the new cluster, $c_j = \frac{1}{n}\sum_{x_i \in Y_j,} x_i, j = 1,2, \dots, k$;
Step 5: Continue to perform steps 2-4 until the clustering criterion function($E = \sum_{i=1}^{n}\sum_{j=1}^{k}\|x_i - c_j\|^2$) converges or the clustering center point(c_j) no longer changes.

The criterion function E is the sum of the distances between each data point and its cluster center c_j. As the number of clustering continues to increase, the E value will also dynamically change. Under ideal conditions, the E value will gradually shrink and become smaller, the mutual distance between objects in the cluster will become smaller, and the mutual distance between clusters will gradually become larger. Therefore, the algorithm seeks a good clustering scheme by constantly seeking smaller and stable E values. When E gradually shrinks to a minimum, it will produce better clustering results.

3.2. The introduction of degree centrality

In social network analysis, degree centrality is a very important and direct method to analyze the importance of nodes in the whole network. The higher the centrality of a node, the more other nodes it is connected to, and the higher the importance of the node in the network.

In an undirected graph G with N nodes, the degree centrality of node i indicates the degree of connection between the node and other N-1 nodes, which is represented as $C_D(N_i) = \sum_{j=1}^{N} x_{ij}(i \neq j), \sum_{j=1}^{N} x_{ij}(i \neq j)$. The calculation of degree centrality can simply add the cell values of the corresponding row or column of node i in the matrix. The degree is divided into in-degree and out-degree in the directed network graph, such as the follow relationship in Weibo. In this study, two nodes whose mutual distance is less than the threshold value are regarded as connected, so the relational matrix is transformed into undirected graph, which does not distinguish between degree and degree.

3.3. The improvement of the initial centre of clustering and the determination of k

Aiming at the influence of the initial center point on the clustering results, this paper proposes a way to optimize the the initial center point based on the measurement of degree centrality in the relation matrix. First, calculate the distance between the nodes in the object set N. Because Yu Q L [14] pointed out that different distance calculation methods would not have a significant impact on this algorithm, in order to reduce the amount of calculation, this paper adopts Euclidean distance and sets a threshold L. According to repeated tests, the value of L is half of the average distance between any two nodes. When establishing the symmetric matrix of the relationship between nodes, if the distance between two nodes is less than the threshold, we set the value to 1, which means that the nodes are connected to each other; If the threshold is less than the distance between two nodes, it is set to 0, that is, the two nodes are considered unconnected. Traverse the row vectors of all objects, set the node of the highest degree centrality, and the nodes connected to it are deleted from the matrix. Go on to iterate through the remaining nodes until a center point is found.

3.4. Description of improved K-Means algorithm

Let the target object set $N = \{X_1, X_2, X_3, , \ldots X_n, \}$, which the number of the objects is n, and each object has m attributes; Suppose every object $X_i = \{x_{i1}, x_{i2}, \ldots x_{im}, \}$, select k center points out of the data set (N). The calculation steps are showed below:

Step 1: preprocess the data set N;

Step 2: Calculate the Euclidean distance between any two objects $dist(X_i, X_j) = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})}$;

Step 3: Calculated the mean distance, $avg = \frac{1}{n}\sum_{\substack{X_i \in N \\ X_j \in N \\ i \neq j}} dist(X_i, X_j)$, and set the threshold $L = avg/2$, establish the distance matrix between nodes ($R = \begin{bmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \ldots & d_{nn} \end{bmatrix}$).

If the threshold L is less than the distance ($dist(X_i, X_j)$) between nodes, change the distance of the matrix to 0, which means that these nodes are not connected; If the threshold L is greater than the distance ($dist(X_i, X_j)$) between nodes, set it to 1, that means, we consider the nodes to be connected in this network. From this, the distance of matrix R becomes a matrix of the following form: R=$\begin{bmatrix} 0 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 0 \end{bmatrix}$.

Step 4: Go through every node to find the one which has the centrality of highest, and delete the node. The distance of the two centre points of the class should be far enough, and the nodes connected to each other exist in a cluster, so delete the p nodes connected to it. Finally, the matrix of n-order is turned into an matrix of n-p+1-order.

Step 5: In the newly formed matrix, the node with the highest degree centrality is further found, set as the second center, and the node to which it is connected is deleted from the matrix. The iterative algorithm is used to continue the loop until the k-th centre point is found.

IV. Research Processes and Results

4.1. Data sources

In this study, the improved K-Means clustering algorithm is used to classify Chinese entrepreneurship policy texts. The policy text comes from the professional legal website(Lexiscn.com). The research span is from January 1, 2000 to December 31, 2020. The research object is the Chinese national level laws, regulations and policy text on "entrepreneurship". In order to ensure that the data are comprehensive, reliable, detailed and true, it horizontally covers the policy-making subjects with different responsibilities and functions, and vertically exhausts the national

entrepreneurial policy documents in the past 21 years, with a total of 462 texts.

4.2. Word segmentation and stop words processing

First, the text is segmented by using the Jieba algorithm in Python, and the text after segmentation is stored in the txt file. Second, stop words needs to be removed. Text after word segmentation is presented in the form of several word sets. For professional texts such as entrepreneurship policy, there are always some words that appear frequently but fail to reflect the theme or core meaning of the text. These words are undoubtedly a kind of noise to the core keywords of the text, such as "policy", "country", "department" and "system" in the text of entrepreneurship policy, so these words need to be removed. In this study, a stop words dictionary is constructed to filter the word set, and if the words in the set appear in the stop words dictionary, they will be deleted.

4.3. Feature extraction

TF-IDF algorithm is used to find out the key words in policy texts, and noun keywords are extracted as the basis of policy text content clustering in the paper. The principle of TF-IDF can be easily explained as follows: if a word or phrase appears frequently in one article, but rarely in other articles, then it is considered that the word or phrase has good representativeness and is suitable for classification. The formula of TF-IDF algorithm is as follows:

$$TF - IDF = \frac{T_i}{N_t} \times log\left(\frac{D_n}{D_{t+1}}\right) \quad (1)$$

In which: i stands for words in the text; T_i means the number of occurrences of the word; N_t represents the total number of words in the article; D_n represents the total number of texts; D_t indicates the number of texts containing the word i.

The result returned by the algorithm is a word-text relational matrix. Among them, the row vector of the relational matrix represents vocabulary and the column vector represents text. Each row of the matrix stores the weight data of each word in a text, and the arrangement order is consistent with that of the words in the list. If a word does not appear in a text, the weight is 0.

4.4. Text cluster analysis of entrepreneurship policy

According to the improved K-Means clustering algorithm, firstly, the node which has the largest degree centrality is found as the initial clustering centre by measuring the mutual distance and average distance between the entrepreneurial policy text and the extracted keywords, so as to optimize the initial clustering center selection; Secondly, the nodes with the highest degree of centrality are eliminated in turn to find out other central points; Thirdly, iterative calculation is carried out for many times until the first center point is found. The final iteration K value of content clustering in this study is 8, that is, the content clustering results can divide the entrepreneurship policy text into eight topics. The clustering results of eight policy texts are 84, 20, 25, 92, 72, 59, 55 and 55 respectively.

V. Discussion

There is no standard classification method and classification results for reference in China's entrepreneurship policy texts. Therefore, in order to test the clustering effect of the improved K-Means clustering algorithm on entrepreneurial policy texts, and further explore the inherent characteristics and hidden information of the similar text data, so that the clustering effect can be presented more clearly and intuitively, on the basis of studying entrepreneurial policy texts, this study carries out secondary word segmentation on the original data of each type of entrepreneurial policy texts. The policy meanings of keywords are reflected as accurately as possible, and the top

100 high-frequency words in the statistical results of word frequency in each category are taken as sample data, and word cloud images are made by using Word Cloud module in Python language [15]. Figures 1-8 are the cloud charts of the eight keywords in the text content clustering of entrepreneurship policy.

Fig. 1: Category I-Business Environment

Fig. 2: Category II-Financial and Tax Support

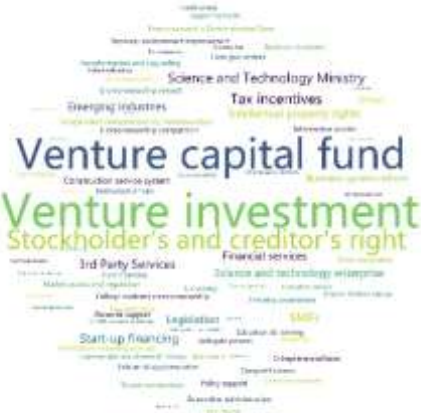

Fig. 3: Category III-Venture capital

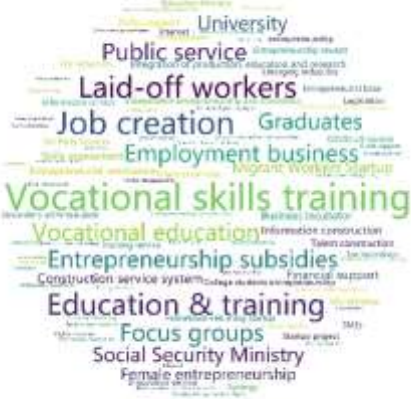

Fig. 4: Category VI-Unemployed entrepreneurship

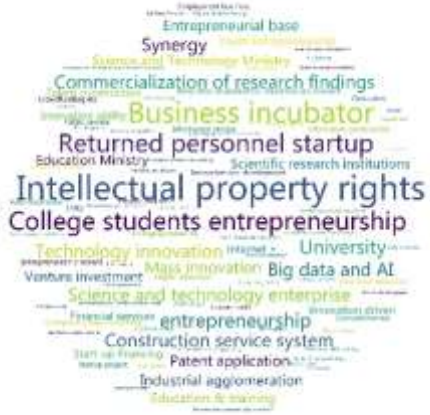

Fig. 5: Category V-Innovation and Entrepreneurship

Fig. 6: Category VI-SMEs

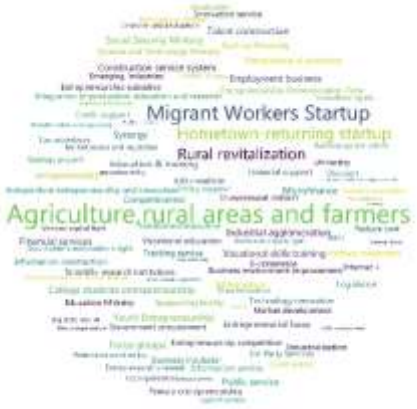

Fig. 7: Category VII-Graduate

Fig. 8: Category VIII-Rural Entrepreneurship

From the above words, we can clearly see the theme characteristics of these eight categories of entrepreneurship policy texts. In Figure 1, "business environment improvement", "construction service system ", "internet +", "start-up financing" and "financial services" are the key words of this kind of entrepreneurial policy texts, so they can be classified as "business environment". In Figure 2, "tax incentives", "income tax" and "financial support" are the key words of this kind of entrepreneurial policy text, so they can be summarized as "financial and tax support". In Figure 3, "venture investment", "venture capital fund" and "stockholder's and creditor's right" are the key words of this kind of entrepreneurial policy text, so they can be summarized as "venture capital". In Figure 4, "vocational skills training", "job creation", "laid-off workers" and "education & training" are the key words of this kind of entrepreneurship policy text, so they can be summarized as "unemployed entrepreneurship". In Figure 5, "intellectual property rights", "talent construction", "college Students Entrepreneurship", "returned personnel startup" and "business incubator" are the key words of this kind of entrepreneurship policy texts, so they can be classified as "Innovation and Entrepreneurship". In Figure 6, "SMEs"(small and medium-sized enterprises), "construction service system" and "information construction" are the key words of this kind of entrepreneurship policy text, so they can be summarized as " SMEs". In Figure 7, "graduates", "university", "employment business" and "college students' entrepreneurship" are the key words of this kind of entrepreneurship policy text, so they can be classified as "graduates entrepreneurship". In Figure 8, "agriculture, rural areas and farmers ","migrant workers startup", "hometown-returning startup" and "rural revitalization" are the key words of this kind of entrepreneurship policy text, so they can be classified as "rural entrepreneurship".

Through the above analysis, it can be seen that the improved k-means clustering algorithm is used to aggregate the Chinese entrepreneurship policy texts into eight categories, each of which has distinct thematic features, and the theme differences between different categories are obvious. These thematic features are the core focus of China's entrepreneurship policy for more than 20 years. Therefore, the new K-Means clustering method proposed in this study has got good clustering consequences in the application of topic classification of entrepreneurship policy texts.

VI. Conclusion

The traditional K-means algorithm needs to artificially determine the value in advance, and randomly select k points of the clustering object as the initial clustering center, which leads to the poor stability of the algorithm and easy to produce unsatisfactory results. In this paper, a method using degree centrality and relation matrix is proposed to optimize the initial center node in K-means algorithm, and a more optimized initial center point is

obtained. Although the process of generating matrix causes a certain amount of time consumption, this algorithm reduces the number of iterations in the clustering process and obtains more stable and high-quality clustering results, so these costs are worth paying in practical application. In addition, the improved K-means algorithm finally determines the value by iteration, which greatly reduces the blindness and subjectivity of the value, so it is desirable.

The study uses the optimized K-means algorithm to cluster 462 Chinese entrepreneurship policy texts, and verifies the clustering effect through the word cloud image. At present, there hasn't been recognized standards to the classification of the Chinese entrepreneurship policy texts and theme topics, so the paper to carry out intuitionistic display to the clustering effect. It can be seen from Figure 1-8 that the improved K-means algorithm aggregates 462 Chinese entrepreneurial policy texts into eight categories with clear topics. After further studying the policy text, it can be seen that these eight categories of topics are the core focus of the Chinese entrepreneurial policy. Therefore, good results has been achieved by the improved K-means method in the clustering of long text documents with large groups of random unknown number of categories and category topics. This research not only proves the objectivity and effectiveness of the improved algorithm in practical applications, but also provides a feasible approach for the classification of Chinese entrepreneurial policy text collections. However, it should be pointed out that the optimization method proposed here still has some limitations in the performance of processing massive data, which is also the direction of further research.

Acknowledgement

The project presented in this article is supported by the Chinese national social science fund project (18BJL039).

References

[1] Shi Z. Semi-supervised model based document clustering: a comparative study. Machine Learning, 2006,65(1):3–29.
[2] Ghoul A E and Sahbi H. Semi-supervised learning using a graph-based phase field model for imbalanced data set classification. IEEE Int. International Conference on Acoustics, 2014.
[3] Chen G C et al. K-means bayes algorithm for imbalanced fault classification and big data application. Journal of Process Control, 2019(81):54-64.
[4] Erzhou Zhua and Ruhui Ma b. An effective partitional clustering algorithm based on new clustering validity index. Applied Soft Computing, 2018(71):608–621.
[5] Xie X. et al. Mixed obfuscation of overlapping instruction and self-modify code based on hyper-chaotic opaque predicates. IEEE Int. 2014 Tenth International Conference on Computational Intelligence and Security, 2014:524–528.
[6] Tunali V et al. An Improved Clustering Algorithm for Text Mining: Multi-Cluster Spherical K-Means. International Arab Journal of Information Technology, 2016,13(1):12–19.
[7] Behera C K and Bhaskari D L Self-Modifying Code: A Provable Technique for Enhancing Program Obfuscation. International Journal of Secure Software Engineering, 2017,8(3):24-41.
[8] Jiang X P et al. A modified K-means clustering for mining of multimedia databases based on dimensionality reduction and similarity measures. Cluster Computing,2017,20(10):1-8.
[9] Zhou S B et al. A new method to determine the best clustering number of K- means algorithm. Computer Engineering and Application,2010,46(16):27–31.
[10] Yu H T et al. Optimized K-means clustering algorithm based on artificial fish swarm. Computer Science, 2012,39(12):60–64.
[11] Chen G P et al. A K-means algorithm for improving initial clustering center selection. Miniature Microcomputer System, 2012,33(6):1320–1323.
[12] Yu S et al. Optimized data fusion for Kernel k-means clustering. IEEE Trans. On Pattern Analysis &

Machine Intelligence, 2012,34(5):1031–1039.
[13] Liu Y et al. Research on optimization method based on K-means clustering algorithm. Information Technology, 2019,43(1):66–70.
[14] Yu Q L. Optimization of initial clustering center selection based on K-means algorithm. Computer System Application, 2017,26(5):170–174.
[15] Zhang B J et al. Theme analysis and evolution process of national science and technology innovation policy. Science and Science Technology Management, 2019,40(11):15–31.

ISSN: 0010-8189