Improved k-means-based FAKM Clustering Method for Scientific and Technical Literature

Meishu Zhao*[a], Baosheng Yin[a]
[a] Research Center for Human Computer Intelligence, Shenyang Aerospace University, 37 Daoyi South Street, Shenyang, China,110000

ABSTRACT

Research on rapid clustering technology based on bibliographic information of scientific and technical literature aims to efficiently realize the correlation analysis of scientific and technical literature, laying the foundation for discovering hot spots and trends in the research field, conducting interdisciplinary and cross-border research, and accurately recommending scientific and technical literature. Focusing on the analysis of clustering algorithms, we proposed an improved k-means-based Firefly Algorithm k-means (FAKM) clustering method, which effectively solved the problem of randomly selecting the initial center points of class cluster when using k-means algorithm for clustering in the clustering stage, which leads to local optimum, low accuracy and large gap between the division of class clusters and the real situation of clustering results. The use of FAKM clustering algorithm resulted in better clustering performance, high accuracy, and fewer iterations. The experimental results showed that the method achieved a silhouette coefficient of 0.54 and adjust mutual information of 0.69 on the same scientific and technical literature data set, which proved the good performance of the method.

Keywords: text clustering; scientific and technical literature processing; k-means clustering

1. INTRODUCTION

With the continuous development of the Internet and digital technology, the volume of scientific and technical literature is also growing worldwide. For the vast amount of literature data, traditional manual retrieval and analysis methods can no longer meet the needs of researchers. Research on rapid clustering technology based on bibliographic information of scientific and technical literature aims to efficiently realize the correlation analysis of scientific and technical literature, laying the foundation for discovering hot spots and trends in the research field, conducting interdisciplinary and cross-border research, and accurately recommending scientific and technical literature.

Scientific and technical literature clustering, in essence, is a kind of text clustering, which can usually be divided into three steps: text pre-processing, text modeling, and text clustering process[1].Among them, the selection of clustering algorithm is a key step in the entire process of scientific and technical literature clustering, and therefore has always been a hot topic under research in this field. Clustering algorithm is an unsupervised learning algorithm[2]that can categorize data points into different clusters or categories based on characteristics such as similarity. Various clustering methods for specific situations have emerged in endlessly, but the partition clustering algorithm is still the most widely used algorithm for clustering. The k-means algorithm[3]belongs to a partition clustering algorithm, and improving the k-means algorithm for specific problems[4]can get better clustering results and is of great value for research.

The k-means clustering algorithm usually has advantages such as fast computational speed and applicability to large-scale data sets, but it is sensitive to the selection of initial centers and may obtain local optimal solutions in some cases. Based on this, many scholars at home and abroad have made various improvements from different perspectives. For example, Arthur[5]et al. proposed the k-means++ algorithm based on the original k-means, which was a more efficient initialization method that was able to select better initial clustering centers, thereby improving the accuracy of clustering results to a certain extent; Li Shunyong[6]et al. proposed an improved method for determining the k value based on hierarchical thinking; Wu Yunming[7]et al. proposed a method of first using the Canopy algorithm to perform "coarse" clustering on samples, and improving the initial cluster centers to reduce the uncertainty of random selection of initial cluster centers of the k-means algorithm, so as to improve the stability and accuracy of the k-means algorithm in specific fields.

*moonsoup942@163.com; phone +86(024)89726878

2 RELATED TECHNOLOGY STUDIES

2.1 Analysis of k-means Algorithm

The k-means algorithm is a partition clustering algorithm that randomly selects k data objects during the clustering, then calculates the similarity of other data objects to these initial data objects, and categorizes them into the most similar clusters. By continuous iterative optimization of the center points of the clusters, this algorithm can ultimately obtain k categories, where each category has the same characteristics, and data objects of different categories have different text characteristics. The principle is shown in Figure 1.

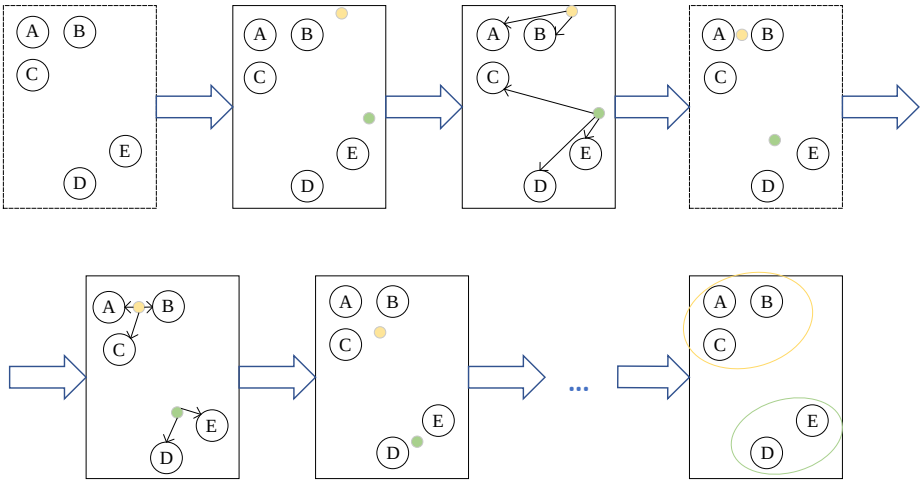

Figure 1. Principle Diagram of k-means Algorithm

Figure 1 contains a total of seven subgraphs, each of which represents a different operation of the k-means algorithm. The first subgraph represents the initial data which has not clustered; in the second subgraph, the number of clusters in the data set is set to 2, and the center points of the first iteration are represented with solid circles and assigned to each cluster; the third subgraph represents the calculation of the distance of each data object from each of these two center points, and then categorizes them based on the principle of closest distance, thus obtaining the result of the first iteration; the fourth subgraph represents the assignment of a new center point to each new category; the operation of the fifth subgraph is the same as that of the third subgraph, which yields the results of the second iteration of clustering; the sixth subgraph corresponds to the operation of the fourth subgraph; the operations of the third and fourth subgraphs are repeated, iterating the center points until they no longer change, so as to obtain the final category division, and the seventh subgraph represents the result.

The k-means algorithm is mainly used to divide a given sample set $D=\{x_1,x_2,\ldots,x_m\}$ into k categories $C=\{C_1,C_2,\ldots,C_k\}$ for the clusters obtained from clustering, and minimize the square error, with the formula as follows:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\|x-\mu_i\|_2^2 \tag{1}$$

where $\mu=\frac{1}{|C_i|}\sum_{x\in C_i}x$ is the mean embedding of cluster Ci. The Formula 1.1 calculates the sum of Euclidean distances[8] from the data in each category to the center of that category, which reflects how closely the sample points within a cluster are gathered around the cluster mean embedding. The smaller the *E* value, the higher the similarity of sample points in the cluster and the better the clustering effect.

The k-means algorithm has the advantages of simple algorithm framework structure, very clear and easy-to-understand basic ideas, and low time complexity. When the data set is large, it is fast to execute. It also has good scalability, i.e., it can be used for clustering of large-scale data sets.

But the k-means algorithm also has drawbacks. Firstly, the number of clusters k needs to be pre-set manually. Before the k-means algorithm is run, the value of the number of clusters k needs to be set in advance, but it is usually difficult to know exactly the number of clusters in the data set in advance, so the value of k is usually determined based on experience or

experiment, which may lead to inaccurate clustering results. Secondly, the clustering results are influenced by the selection of initial cluster centers. The k-means algorithm usually randomly selects cluster centers for iteration first, but each iteration is affected by the previous iteration, so the quality of the selection of initial cluster centers will have a significant impact on the clustering results. If the initial cluster centers are poorly selected, local optimal solutions and the like may occur.

2.2 Firefly Algorithm

Based on the previous description, we know that for the k-means algorithm, the optimization algorithm used to determine the optimal cluster centers may converge to the local optimum point, so an effective optimization algorithm is needed. The Firefly Algorithm is a nature-inspired, adaptive algorithm[9] ,which was proposed by the Cambridge scholar Xin-She Yang[10] in 2010 based on the luminous behavior of fireflies in nature. This algorithm improves the performance of k-means clustering by improving the initial solution, avoiding the local optimum.

The Firefly Algorithm is a heuristic algorithm based on the principle that when a firefly attracts all other fireflies by its brightness, the attraction is proportional to the brightness of the firefly. Assuming that each firefly represents a data point, the brightness of the firefly represents the adaptability of the firefly at that location, and the firefly with the highest brightness naturally attracts the firefly with lower brightness. That is to say, the brighter the firefly, the better its coordinates in the solution space. In the solution space, fireflies with lower brightness will fly towards those with higher brightness, so that they can search for better locations. The specific steps of the Firefly Algorithm are as follows.

Firstly, the parameters are initialized; secondly, the locations of fireflies are randomly initialized, and the objective function values at their locations are used as the respective maximum brightness of the fireflies; thirdly, the relative brightness between the fireflies is calculated based on their distance and brightness, and the movement directions and step lengths of the fireflies are determined based on the relative brightness; fourthly, the locations of the fireflies are updated based on the movement directions and step lengths of the fireflies; fifthly, the values of the objective functions on the updated locations of the fireflies are calculated and used as the new brightness of the fireflies; and sixthly, whether the termination condition is met is determined. If the termination condition is met, the optimal solution will be output; otherwise, steps 2 to 5 will be repeated until the termination condition is met. The flow chart of the algorithm is shown in Figure 2.

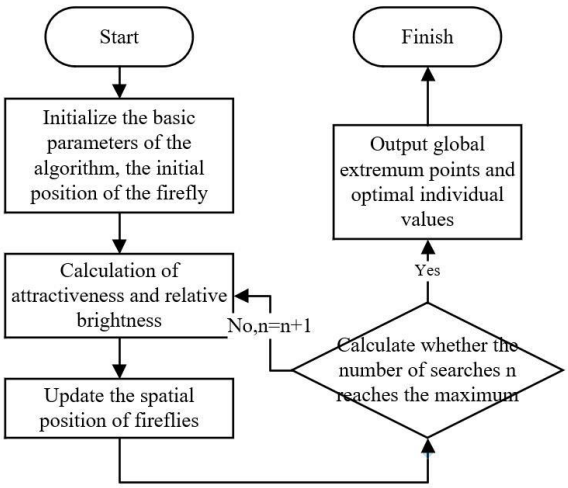

Figure 2. Flow Chart of FA Algorithm

In the Firefly Algorithm, the attraction calculation formula used for the initial cluster center selection is as follows:

$$I_{ij}=\frac{I_0}{d_{ij}^2+1} \tag{2}$$

where I_{ij} represents the magnitude of the attraction of the i^{th} firefly by the j^{th} firefly, I_0 represents the magnitude of the initial attraction, and d_{ij} represents the distance between the two fireflies. Fireflies will move towards the brighter fireflies around them, thus enabling the selection of the clustering center initialization.

It can be seen that the Firefly Algorithm uses the mutual attraction between fireflies for the selection of clustering centers, which is more likely to avoid local optimal solutions and is more adaptable to different data distributions.

3 IMPROVED K-MEANS-BASED FAKM CLUSTERING ALGORITHM

3.1 Introduction to FAKM Algorithm

In this chapter, the Firefly Algorithm will be introduced into the k-means algorithm to propose a FAKM-based clustering algorithm. The basic idea is as follows: Firstly, the Firefly Algorithm is used: calculate the attraction and relative brightness, and get the initial center points according to the result; calculate the distance between each of other data objects and the initial center points, update the locations, determine whether the data objects meet the conditions to participate in the next round of clustering, repeat the above operation until the data set becomes empty, and output the k value and the initial cluster centers; then, the k-means algorithm is used: use the result obtained by the Firefly Algorithm as the basis for clustering to get the final result, with the main parameters as follows:

(1) Degree of Attraction

$$\beta=\beta_0 e^{-\gamma r_{ij}^2} \tag{3}$$

where β is the degree of attraction, which decreases with the increasing distance; β_0 is the degree of attraction at the current location of the firefly as the light source, that is, its initial degree of attraction, which indicates the degree of attraction between two fireflies in the initial state, this initial state usually means a distance of zero, and its value is related to the brightness of the firefly; γ is the absorption coefficient of light intensity, which is usually a constant; r_{ij} represents the distance between the light source firefly and the attracted firefly.

(2) Location Update

Firefly i is attracted to move towards brighter one j, i.e.,

$$\boldsymbol{c}_i^{'}=\boldsymbol{c}_i+\beta_0 e^{-\gamma r_{ij}^2}\left(\boldsymbol{c}_j-\boldsymbol{c}_i\right)+\alpha\varepsilon \tag{4}$$

where $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ represent the locations of fireflies i and j, respectively; $\boldsymbol{c}_i^{'}$ represents the location of firefly i after its movement; ε is a random factor that follows a uniform distribution; α is the step factor, which is a constant in the interval [0,1]; $\alpha\varepsilon$ serves as a disturbance term to avoid the local optimum.

The random movement of firefly i when it is the brightest firefly

$$\boldsymbol{c}_i^{'}=\boldsymbol{c}_i+\alpha\varepsilon \tag{5}$$

In the algorithm, the locations of fireflies represent data points, and the brightness represents the evaluation indicator. The algorithm first generates initial locations of the fireflies by randomly selecting data points. Then, a certain evaluation indicator of the data point represented by each firefly is calculated based on its current location, and the calculated brightness value is recorded as the brightness of that firefly.

In each iteration, the algorithm selects the firefly with the highest brightness value and records its location information. Next, the algorithm determines the degree of attraction between fireflies by calculating their brightness and distance, and moves them according to certain rules. After movement, the algorithm updates the locations of fireflies and takes them as fireflies for the next iteration in the next cycle. To ensure that the fireflies do not move beyond a limited range, the algorithm constrains the locations of the fireflies. After the termination condition is met, the algorithm will select and record the location information of the firefly with the highest brightness as the output, that is, the location information of the entity with the best evaluation indicator.

3.2 Steps of FAKM Algorithm

The FAKM algorithm is a combination of the Firefly Algorithm and the k-means algorithm, which makes clustering more effective. The FAKM algorithm includes three steps: Step 1, generating characteristic embedding based on the data set used. Step 2, predicting the k-value and initial cluster centers, the locations of which also affect the quality of clustering results. In the FAKM algorithm, some heuristic algorithms can be used to make predictions for k values and initial cluster centers. Step 3, conducting clustering experiments to obtain the final results of clustering. Specifically, in the process of adjusting the cluster center using the Firefly Algorithm, the cluster center can be regarded as an attractor, and the Firefly

Algorithm updates the cluster center by simulating the mutual attraction of fireflies to achieve better results of clustering. The flow chart of the FAKM algorithm is shown in Figure 3.

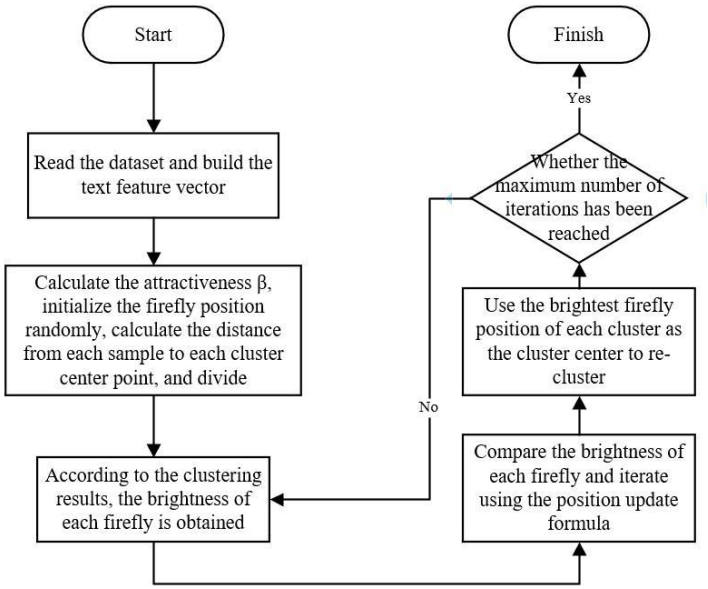

Figure 3. Flow Chart of FAKM Algorithm

3.3 Evaluation Indicators

In this study, silhouette coefficient[11](SC) and adjust mutual information (AMI)[12]were selected as evaluation indicators to evaluate the clustering results.

The SC uses the similarity measurement between objects in a data set to evaluate the quality of clustering, and is an indicator of cluster density and dispersion. The SC applies where the actual category information is unknown. The SC value is within the range of [-1,1]. The larger the value, the better the clustering effect.

Mutual information (MI) is a measure of the interdependence between two random variables, and is an indicator of the correlation between two sets of events. In order to test the effect of clustering information in the bibliography of scientific and technical literature within the cluster, AMI is introduced, which is an improvement of MI, and can better reflect the degree of coincidence of data distribution. The AMI value is within the range of [-1,1]. The closer its value is to 1, the closer the clustering result is to the true value.

4 EXPERIMENTS

4.1 Experimental Data

The data set used in this paper has over 40,000 pieces of information from scientific and technical literature, covering the fields of aerospace technology, computer technology, materials science and life science, with a wide distribution of specialties, complete types and certain representativeness. The effectiveness of the FAKM algorithm proposed in this paper was verified through this data set.

(1) Text Data Acquisition

Due to lack of publicly available data set for the analysis of Chinese scientific and technical literature in China, network data collection technology[13] was used to collect some Chinese scientific and technical literature data from the CNKI database as the experimental data set, as shown in Figure 4. The collected literature data mainly includes the title, author, abstract, literature source, publication time, keywords and citations.

name	author	content	source	datetime
冲击波作用后变形机翼模态数值模拟研究	肖良丰 周兰伟 李	为研究典型机翼在爆炸冲击	北京航空航天大学	2022.10.10
外挂物对大展弦比直机翼颤振特性的影响	祁武超 张贺铭 田	为提高带外挂物大展弦比直	计算力学学报	2021.09.29
倾转过渡状态旋翼-机翼气动干扰特性计算分	刘佳豪 李高华 王	本文针对倾转旋翼机倾转过	航空学报	2021.10.11
面向超声速民机层流机翼设计的转捩预测方法	聂晗 宋文萍 韩忠	发展工程实用的转捩预测方	航空学报	2021.12.03
撤回 考虑转动部件影响的机翼气弹耦合特性	张夏阳 周旭 赵匡	对机翼动力装置相关部件转	航空动力学报	2021.12.27
旋翼/机翼气动干扰对复合式直升机性能影响	杨克龙 韩东	为研究旋翼/机翼气动干扰对	北京航空航天大学	2022.01.26
基于变弯度后缘的机翼阵风响应减缓数值研究	尉濛恺 戴玉婷 杨	针对带有变弯度后缘的机翼	北京航空航天大学	2022.01.27
基于系统辨识的自适应变形机翼控制系统设计	谢长川 朱立鹏 孟	相对于传统飞行器的固定机	北京航空航天大学	2022.05.25
弹性机翼刚度的静气弹敏感性研究	陈恺 刘晓燕 程攀	为保证大展弦比柔性机翼在	实验流体力学	2022.06.09
复合式高速直升机旋翼下洗流对机翼的气动影	刘超凡 朱清华 刘	通过对复合式高速直升机的	航空工程进展	2022.06.11
混合层流机翼气动设计与综合收益影响研究	姜丽红 饶寒月 兰	揭示混合层流控制（Hybrid	航空学报	2022.06.30
变弯度机翼参数化气动弹性建模与颤振特性分	喻世杰 周兴华 黄	变体飞行器在变体过程中结	航空学报	2022.07.25
考虑机翼柔性的磁流变减震起落架落震动力学	祝恒佳 杨丽昆 祝	为研究飞机着陆过程中机翼	西安交通大学学报	2022.07.25
远程民机变弯度机翼后缘外形变形矩阵气动设	李春鹏 钱战森 孙	针对某远程民机变弯度机翼	航空学报	2022.08.03
倾转旋翼飞行器机翼滑流区面积可视化计算方	宋伟 王琦 何国毅	机翼滑流区面积计算是进行	北京航空航天大学	2022.08.24
大展弦比复合材料机翼非线性变形及模态分析	李金洋 王军利 王	为明确大展弦比复合材料机	战术导弹技术	2022.08.26
基于CST的三维机翼气动结构解析参数化建模	杨予成 粟华 龚春	针对概念设计阶段机翼设计	航空动力学报	2022.08.29
变弯度后缘与常规舵面机翼的颤振主动抑制对	杨永健 宋晨 张桢	后缘变弯度机翼的气动弹性	航空工程进展	2022.09.01
模块化可重构无人机机翼结构优化方法研究	罗利龙 郭文杰 常	模块化可重构无人机设计工	航空工程进展	2022.09.02

Figure 4. Excerpts from Data Set of Bibliography of Scientific and Technical Literature

(2) Text Pre-processing

The data was pre-processed after the literature data set was obtained. Firstly, the literature data was subjected to data cleaning to delete duplicate and incomplete data information, and remove non-technical literature data information such as solicitation information, conference notice, journal and magazine introduction. The remaining 42,291 pieces of scientific and technical literature data were used as the experimental data set, and data category labels were manually added. The category distribution is shown in Table 1.

Table 1. Breakdown of Data Categories

Category Name	Quantity
Aerospace	13,027
Materials Science	9,974
Computer	9,400
Life Science	9,890

4.2 Experimental Results and Analysis

The clustering of data set of bibliography of scientific and technical literature included the following steps: firstly, the text was subjected to pre-processing operations such as word segmentation and removal of stop words; secondly, the text was transformed into characteristic embeddings, which involved converting the text into embedding character representation that could be processed by a computer; and finally, the clustering experiment was conducted. After the above three steps, i.e., after clustering of the text, the clustering results were analyzed to prove the effectiveness of the method, and specific clustering evaluation indicators were used for analysis of the results. In this chapter, the analysis was conducted from the following four aspects.

(1) The influence of k-value on clustering results

(2) Comparison of SCs of clustering results

(3) Comparison of AMI of clustering results

(4) Comparison of the number of algorithm iterations

When the k-means algorithm is used for clustering text characteristic embeddings, the results obtained will naturally vary due to the different choices of k values. Based on the data classification of the data set used in this experiment, a number from the interval [2, 10] was selected as the k-value, and the clustering results are shown in Figure 5. When k=6, the clustering effect was the best. Therefore, the number of clustering topics selected in this chapter was k=6.

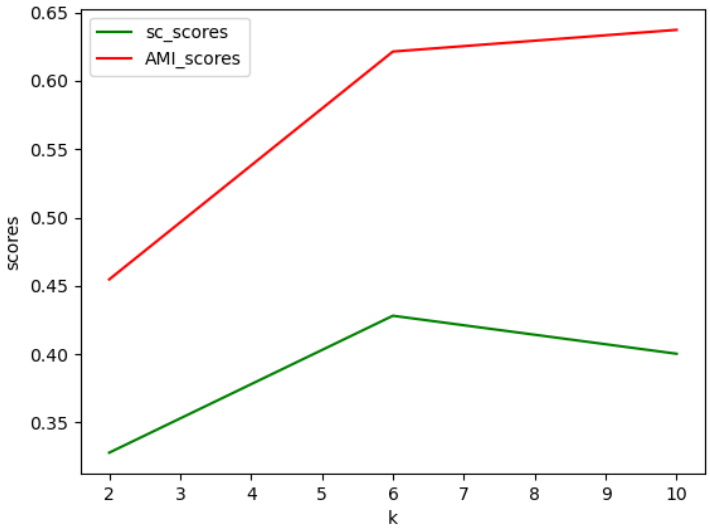

Figure 5. Clustering Results With Different k Values

In order to effectively prove the effectiveness of the clustering algorithm proposed in this chapter, the experiment was performed several times with the same data set, and the average value was calculated for the results of multiple clustering, where the frequency of clustering was selected as 10. For other parameters of the algorithm, the set values are shown in the following table:

Table 2. Algorithm Parameters

Parameter	Description	Value
num_fireflies	Number of fireflies	200
alpha	Moving step length in Firefly Algorithm	0.1
beta	Magnitude of attraction in Firefly Algorithm	1
gamma	Attenuation rate of luminous brightness in Firefly Algorithm	0.01

In order to verify the effectiveness of the FAKM algorithm proposed in this chapter for text clustering, the title and abstract data in the data set of bibliography of scientific and technical literature were spliced and used as experimental data for the clustering algorithm. The k-means algorithm was selected for the comparative experiment, and the same hardware configuration environment was used for the comparative experiment. The experimental results of SCs and AMI are as follows:

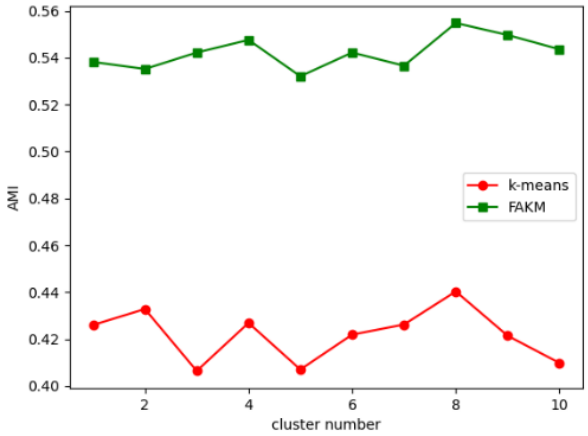

Figure 6. Fluctuations of SCs for Different Clustering Coefficients

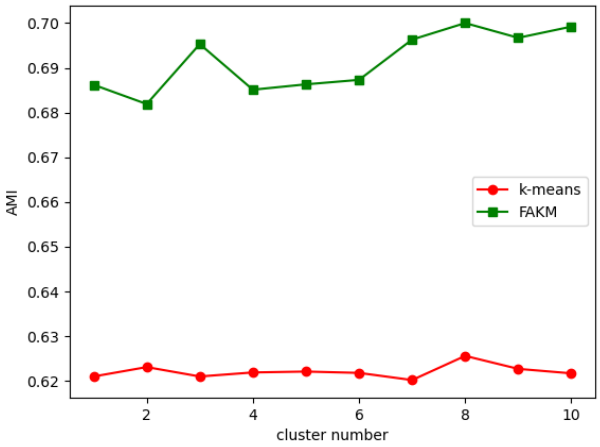

Figure 7. Fluctuations of AMI for Different Clustering Coefficients

Figure 6 and Figure 7 show the fluctuations of SCs and AMI values for various clustering results. It can be seen that the clustering results are not stable, but fluctuate up and down within a certain range. In both figures, the evaluation indicator values of the k-means algorithm are lower than those of the FAKM algorithm, indicating that the FAKM algorithm has more accurate classification of class clusters and better performance. The final results were expressed by means of mean values, in order to make the clustering results more objective and fair. The average values of SCs and AMI of the 10 clustering results were calculated. The results are shown in Table 3.3.

Table 3. Results of SCs and AMI

Clustering Algorithm	**SC**	**NMI**
k-means	0.4219	0.6221
FAKM	0.5422	0.6914

In Table 3, due to the randomness of initial selection of the center points of the k-means algorithm, the clustering result is worse than that of FAKM algorithm, with the SC of 0.4219 and AMI of 0.6221, while the CPKM algorithm proposed in this chapter increases the accuracy of initial selection of the center points, with the SC of 0.5422 and AMI of 0.6914.

The clustering experiments were performed on the data set of bibliography of scientific and technical literature, the number of iterations for each number of clusters was recorded, and the results of the experiments are shown in Figure8.

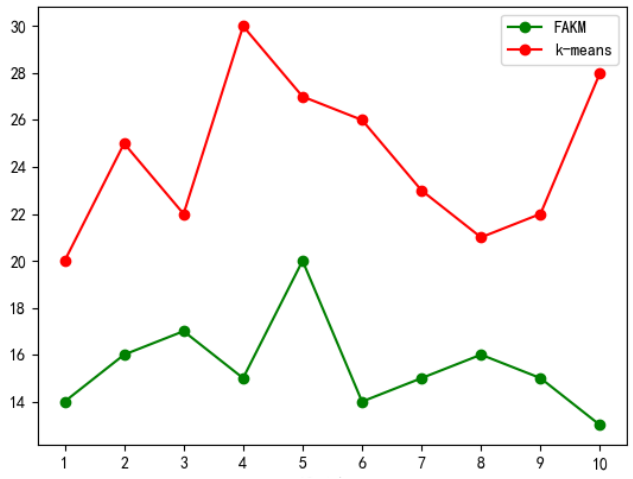

Figure 8. Comparison of Number of Iterations of Different Algorithms

The number of iterations depends on the selection of initial center points. Figure 8 shows the number of iterations of the FAKM algorithm and the k-means algorithm in multiple clustering experiments. It is not difficult to see from the figure that the number of iterations of the FAKM algorithm is smaller and the fluctuations are relatively weak compared to the k-means algorithm, indicating that the initial center points are in locations where the data objects are dense, the division of the data objects is relatively easy, so the number of iterations is also relatively small. On the contrary, the stability of the k-means algorithm is poor. It is not difficult to see from the figure that the number of iterations fluctuates wildly, because the initial center points are not reasonably selected, and there is a deviation from the true center point of each class cluster, some of the initial center points are far from the locations of dense data objects, which are isolated points. Therefore, it takes several iterations to correctly categorize the data objects, and the number of iterations naturally increases.

5 CONCLUSIONS

This paper firstly introduces the process of clustering scientific and technical literature, investigates the current status of domestic and international research on clustering methods, and conducts an analysis. It was found that in the stage of text clustering, partition-based clustering algorithms such as k-means clustering were simple, fast, and widely used. However, the algorithm required predetermination of number of clusters and initial cluster center points, which might lead to poor or unstable clustering results during the algorithm implementation, such as local optimum.

In response to the above problems, research on clustering algorithms was carried out and the FAKM algorithm improved based on k-means was proposed, whose basic idea was to introduce the Firefly Algorithm on the basis of the k-means algorithm to achieve the optimization of cluster center points. Experiments were conducted on data set of bibliography of scientific and technical literature, and the optimal number of clusters was obtained through experiments on the impact of different k values on clustering effectiveness; through the comparison of clustering results, it was concluded that FAKM algorithm was superior to k-means algorithm, with a certain degree of improvement in both the internal evaluation indicator SC and the external evaluation indicator AMI.

Clustering research on scientific and technical literature data is a very valuable job, which can explore the internal connections of scientific and technical literature. However, there are still some shortcomings in this paper, which should be continuously studied and improved in future research. The subsequent research will focus on the following aspects: firstly, in this study, only bibliographic information of scientific and technical literature was used as the clustering data set for experiments, and the citation relationship of scientific and technical literature was not deeply explored. Next, we will further explore the citation relationship of scientific and technical literature to improve the effectiveness of our research; secondly, only the data set of Chinese scientific and technical literature was collected in this paper, and the analysis of scientific and technical literature in multiple languages may be more meaningful.

ACKNOWLEDGMENT

Funding this work was supported by the Scientific Research Project of Liaoning Provincial Department of Education (LJKMZ20220536).

REFERENCES

[1] Shang W. Research on Feature Extraction Algorithm in Science and Technology Literature Clustering [D]. Beijing: Beijing University of Posts and Telecommunications, 2017.
[2] Hou H, Ding S, Xu X. Research progress of deep clustering based on unsupervised representation learning [J]. Pattern Recognition and Artificial Intelligence, 2022,35(11):999-1014.
[3] Hartigan J A, Wong M A. Algorithm AS 136: A k-means clustering algorithm[J]. Journal of the royal statistical society. series c (applied statistics), 1979, 28(1): 100-108.
[4] Wang S, Liu C, Xing S. A Review of K-means Clustering Algorithms [J]. Journal of East China Jiaotong University, 2022,39(05):119-126.
[5] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding[R]. Stanford, 2006.
[6] Li Shunyong, Zhang Yujia, Peng Xiaoqing, etc. A fast clustering algorithm for big data based on stratified sampling [J]. Computer Application and Software, 2020, 37(10): 256-261+277.
[7] Wu Y, Wang L, Wei Z, etc. Research on Aggregation, Mining and Analysis of Mobile Commerce User Demand Based on Canopy-Kmeans [J]. Information Science, 2022, 40(10): 97-106.
[8] Al Radhwani A M N, Algamal Z Y. Improving K-means clustering based on firefly algorithm[C]. Journal of Physics: Conference Series. IOP Publishing, 2021, 1897(1): 012004.
[9] Wang Jidong, Gu Zhicheng, Ge Leijiao, etc. Analysis of Distribution Network Load Clustering Characteristics Based on the Combination of Improved Firefly Algorithm and K-means Algorithm [J]. Journal of Tianjin University (Natural Science and Engineering Technology Edition), 2023, 56(02):137-147.
[10] Yang X S. Firefly algorithm, stochastic test functions and design optimization[J]. International journal of bio-inspired computation, 2010, 2(2): 78-84.
[11] Yin Shizhuang, Wang Tao, Xie Fangfang, etc. Clustering Results Evaluation Method Based on Mutual Information and Silhouette Coefficient [J]. Journal of Ordnance Equipment Engineering, 2020, 41(08):207-213.
[12] Yeung, K. Y. and W. L. Ruzzo. Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. Bioinformatics 17(9) (2001): 763–774.
[13] Songtao H. Practical Combat of Python Web Crawlers [M]. Beijing: Tsinghua University Press, 2017: 271-284.