Complex temporal data model and extraction technology of multi-spatial database based on clustering algorithm

Kai Xiao*
Wuhan Railway Vocational College of Technology, Wuhan430205, Hubei, China

ABSTRACT

From the 21st century, the progress of science and technology becomes more and more rapid, making the application of various macro and micro sensors such as radar, infrared, photoelectric, satellite, TV camera, electron microscopy imaging, CT imaging more and more widely, and the amount, scale, and complexity of spatial data are rapidly increasing, which has far exceeded the human ability to interpret. Because end-users are unable to analyze all the data in detail and extract the spatial knowledge of interest, the phenomenon of "spatial data explosion but lack of knowledge" occurs. Therefore, in order to improve the utilization efficiency of spatial data, it is necessary to study the complex temporal data model and extraction technology of multi-spatial database and multi-spatial database. Using spatial data mining and knowledge discovery to automatically or semi-automatically mine previously unknown but potentially useful spatial patterns from multi-spatial databases becomes necessary. In view of the above situation, this paper used clustering algorithm to verify and analyze the complex temporal data model and extraction technology of multi-spatial database. By comparing the four aspects of precision P, recall rate R, comprehensive performance F, and content extraction speed of different algorithms, the relationship between the two was obtained. The experimental research results have shown that under other conditions being the same, the precision P, recall rate R, and comprehensive performance F of the k-means clustering algorithm are basically above 97%, higher than those of the ParEx and vu algorithms. In terms of the time required to extract the content, the time of the k-means clustering algorithm is 0.35s, which is much lower than the 0.49s of the ParEx algorithm and 0.61s of the vu algorithm. It can be proved that the k-means clustering algorithm can promote the development of complex temporal data model and extraction technology of multi-spatial database, indicating the positive relationship between the two.

Keywords: Clustering algorithm, multispatial database, complex tense, data model, extraction technique.

1. INTRODUCTION

As the development of multi-spatial data extraction technology has been on the right track, the amount of spatial data is increasing. In order to fully improve the utilization of the resources in the spatial database and obtain the required information in a large amount of data, more and more people start to pay attention to a technology.

In this case, clustering analysis has gradually become a main research direction of multi-space database extraction, which has a wide range of applications and high research value, therefore, the impact of clustering algorithms on complex temporal data models and extraction techniques in multi-spatial databases has become a problem that more and more people think about. By studying the relationship between the two, it is beneficial to further improve the complex temporal data model and extraction technology of multi-spatial database.

2. RELATED WORK

Due to the rapid development of information technology, especially the rapid development of information collection, preservation, processing and other technologies, massive data are generated in the geographical field, resulting in multi-spatial databases. For the research of multi-spatial database, there have been a lot of researches. Among them: The multi-source geospatial database proposed by Zhang M could effectively manage the big data of geoscience[1]. Zhang J's research has shown that multi-spatial databases can provide important information for land-use planning and formulation of multi-scale water quality protection measures[2]. Xie's study found that grassland productivity has a significant response to changes in climatic factors and socioeconomic variables through a multi-spatial database[3]. Deng studied and established a multi-space database to promote sustainable urban development[4]. Ringelman found factors influencing the

*wtzyxk@126.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024), edited by Yang Yue, Proc. of SPIE Vol. 13395, 1339534 · © 2024 SPIE · 0277-786X · Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3049464

survival of nests by studying multi-space databases[5]. However, these studies are mostly data analysis, less practical application, and lack of scientific methods for research.

In view of the above problems, the clustering algorithm can be used to analyze the complex temporal data model and extraction technology of multi-spatial database. For this algorithm, there have been a lot of research results. Among them: Cuzzocrea found that K-means clustering algorithm can effectively support the fragmentation of large XML data warehouse[6]. Chung proposed an effective local clustering algorithm[7]. Saeed's study introduced the OASIS algorithm, software, and meta-analysis datasets of two publicly available SLEGWAS and new SLE genes[8]. Wu and his team proposed an improved low-energy adaptive clustering hierarchical clustering algorithm[9]. Research by Aminanto has shown that ant clustering algorithms can be used to classify data into different categories[10]. These studies have proved the application results of clustering algorithm in all walks of life, and laid the foundation for its application in complex temporal data model and extraction technology of multi-spatial database.

3. CONSTRUCTION OF CLUSTERING ALGORITHM

3.1 Knowledge of clustering

3.1.1 Definition of clustering. The distance between sample data in different clusters is always larger than the distance between sample data in the same cluster, that is, the inter-class distance of sample data is greater than the intra-class distance. The ultimate goal of clustering is to obtain independent and compact clusters[11], as shown in Figure 1:

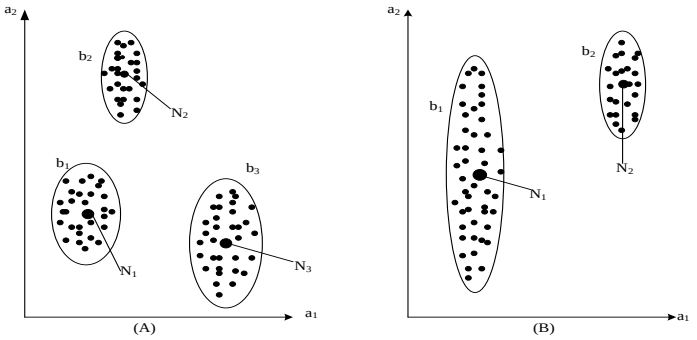

Figure 1. Clustering results

Figure 1 (A) and Figure 1 (B) are schematic diagrams of the two clustering results. In the clustering result shown in Figure (A), the distance between samples in each cluster is very close, and the distance between clusters is large, so the clustering result is ideal; in the clustering result shown in Figure (B), the distance between the sample points at the two ends of cluster b_1 and the cluster center N_1 is very large, and the clustering result is not ideal. Therefore, when this happens, it is necessary to reselect the clustering criterion and repeat the process[12,13].

3.1.2 Cluster similarity measure. Similarity measurement is a method to measure the similarity between sample data[14]. The components of vector A_1 and A_2 should be combined when measuring. However, there is no standard way of how to combine them. It needs to be determined according to the actual situation, so there are different distance measurement formulas:

(1) Euclidean distance

Let A_1, A_2 be two m-dimensional pattern samples, $A_1=[a_{11},a_{12},\cdots,a_{1m}]^Y$, $A_2=[a_{21},a_{22},\cdots,a_{2m}]^Y$. The Euclidean distance between samples A_1 and A_2 is expressed as:

$$F(A_1,A_2)=A_1-A_2=(A_1-A_2)^Y(A_1-A_2)=(a_{11}-a_{21})^2+\cdots+(a_{1m}-a_{2m})^2 \quad (1)$$

When using Euclidean distance as a criterion for measuring similarity, it should be noted that the physical quantities and units of physical quantities of each eigenvector must be consistent in the corresponding dimension[15,16].

(2) Mahalanobis distance

For m-dimensional vectors $A=[a_{11},a_{12},\cdots,a_{1m}]^Y$, $N=[n_1,n_2,\cdots,n_m]^Y$, the squared expression of Mahalanobis distance is:

$$F^2=(A-N)^Y V^{-1}(A-N) \tag{2}$$

In the formula, A represents the mode vector, N represents the mean vector, and V is the sample population covariance matrix[17].

It can be expressed as:

$$V=R\{(A-N)(A-N)^Y\}=R\begin{Bmatrix}(a_1-n_1)\\(a_2-n_2)\\\cdot\\\cdot\\\cdot\\(a_m-n_m)\end{Bmatrix}[(a_1-n_1)(a_2-n_2)\cdots(a_m-n_m)] \tag{3}$$

The advantage of this distance makes it possible to effectively exclude the influence between pattern samples[18].

(3) Mingshi distance

Let A_o and A_k be two m-dimensional pattern sample vectors, and their Mingshi distance is:

$$F_n(A_o,A_k)=\left[\sum_{l=1}^{m}a_{ol}=a_{kl}{}^n\right]^{1_n} \tag{4}$$

Among them, a_{ol} and a_{kl} represent the lth component of A_o and A_k, respectively[19].

When n=1:

$$F_1(A_o,A_k)=\sum_{l=1}^{m}a_{ol}-a_{kl} \tag{5}$$

and it is called the "neighborhood" distance; when n=2, this distance can be regarded as the Euclidean distance[20].

(4) Hamming distance

Let A_o and A_k be two m-dimensional binary (1 or -1) sample vectors, and their Hamming distance is:

$$F_g(A_o,A_k)=\frac{1}{2}\left(m-\sum_{l=1}^{m}a_{ol}\cdot a_{kl}\right) \tag{6}$$

Among them, a_{ol} and a_{kl} represent the lth components of A_o and A_k, respectively.

(5) Tanimoto measure

The Tanimoto measure is suitable for the case of 0,1 binary features, which is specifically expressed as:

$$D(A_o,A_k)=\frac{A_o^Y A_k}{A_o^Y A_o+A_k^Y A_k-A_o^Y A_k} \tag{7}$$

3.1.3 General steps for clustering. The typical clustering process generally includes: data feature standardization and dimensionality reduction, feature selection and extraction of data, selecting or constructing distance function to measure similarity for clustering, verifying clustering results, and drawing conclusions, specifically as shown in Figure 2:

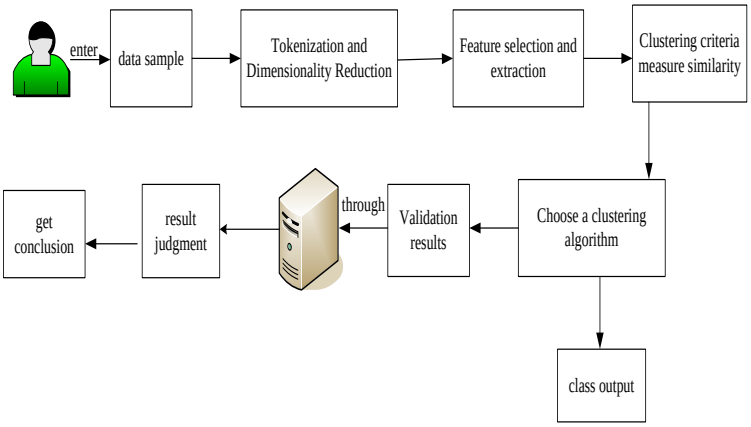

Figure 2. Clustering process.

3.1.4 k-means clustering algorithm. k-means clustering algorithm is the most basic and widely used machine learning method. Its basic meaning is to divide a given dataset $A \in T^n$ into subsets $V_1, V_2, ..., V_l$, where l is the number of cluster categories.

$$\sum_{o=1}^{l}\sum_{k=1}^{m} x_{ok}\left\|a_k - \eta_0\right\|^2 \tag{8}$$

Among them, a_k is each sample point, η_o is the cluster center, when a_k belongs to the V_o th, $x_{ok}=1$, otherwise $x_{ok}=0$, that is:

$$x_{ok} = \begin{cases} 1, & a_k \in V_o \\ 0, & otherwise \end{cases} \tag{9}$$

It is decomposed into the following matrix form:

$$\sum_{o=1}^{l}\sum_{k=1}^{m} x_{ok}\left\|a_k - \eta_o\right\|^2 = \left\|A - NX\right\|^2 \tag{10}$$

A is the sample matrix, where each column represents a sample data; N is the cluster center matrix, where N_k is the kth column of N, representing a cluster center.

Definition: The sample matrix A, each column of which is a_k, the 2-norm of A can be defined as the sum of the squares of the lengths of all sample vectors, namely:

$$\left\|A\right\|^2 = \sum_{z,k} a_{zk}^2 = \sum_{k}\left\|a_k\right\|^2 = \sum_{k} a_k^Y a_k = \sum_{k}\left(A^Y A\right)_{kk} = yt\left[A^Y A\right] \tag{11}$$

$$\begin{aligned} \sum_{o,k} x_{ok}\left\|a_k - \eta_o\right\|^2 &= \sum_{o,k} x_{ok}\left(a_k^Y a_k - 2a_k^Y \eta_o + \eta_o^Y \eta_o\right) \\ &= \sum_{o,k} x_{ok} a_k^Y a_k - 2\sum_{o,k} x_{ok} a_k^Y \eta_o + \sum_{o,k} x_{ok}\eta_o^Y \eta_o = Y_1 - 2Y_2 + Y_3 \end{aligned} \tag{12}$$

Items Y_1, Y_2, and Y_3 are simplified to get:

$$Y_1 = \sum_{o,k} x_{ok} a_k^Y a_k = \sum_{o,k} x_{ok} \|a_k\|^2 = \sum_k \|a_k\|^2 = yt\left[A^Y A\right] \tag{13}$$

$$\begin{aligned} Y_2 &= \sum_{o,k} x_{ok} a_k^Y \eta_o = \sum_{o,k} x_{ok} \sum_{z,k} x_{zk} \eta_o = \sum_{k,z} x_{zk} \sum_o \eta_{zo} x_{ok} = \sum_k \sum_z \left(A^Y\right)_{kz} \left(NX\right)_{zk} \\ &= \sum_k \left(A^Y NX\right)_{kk} = yt\left[A^Y NX\right] \end{aligned} \tag{14}$$

$$Y_3 = \sum_{o,k} x_{ok} \eta_o^Y \eta_o = \sum_{o,k} x_{ok} \|\eta_k\|^2 = \sum_o \|\eta_k\|^2 m_o \tag{15}$$

Among them, m_o belongs to the number of samples of class o.

Expand the equation to the right:

$$\begin{aligned} \|A - NX\| &= yt\left[\left(A - NX\right)^Y \left(A - NX\right)\right] \\ &= yt\left[A^Y A\right] - 2yt\left[A^Y NX\right] + yt\left[X^Y N^Y NX\right] = Y_4 - 2Y_5 + Y_6 \end{aligned} \tag{16}$$

From this, it can be found that $Y_4 = Y_1$, $Y_2 = Y_5$, so it only need to be proved: $Y_3 = Y_6$. The proof is as follows:

$$Y_6 = yt\left[X^Y N^Y NX\right] = yt\left[N^Y NXX^Y\right] \tag{17}$$

Further derivation can be obtained:

$$yt\left[N^Y NXX^Y\right] = \sum_o \left(N^Y NXX^Y\right)_{oo} = \sum_o \sum_z \left(N^Y N\right)_{oz} \left(XX^Y\right)_{zo} = \sum_o \|\eta_o\|^2 m_o \tag{18}$$

XX^Y is a diagonal matrix. It can be seen from this that the proof of $Y_3 = Y_6$ is established, so the final equation to be proved is established and the proof is completed.

The k-means clustering algorithm not only has a simple idea, but also has a relatively fast operation speed. The algorithm itself has the optimization performance of iterative correction, which can optimize the unreasonable initial clustering of samples.

The similarity measure and criterion function of the algorithm cannot cluster samples of arbitrary shape. And when the algorithm uses the criterion function to find L, the relationship between criterion function K_k and the number of clusters L is shown in the Figure 3:

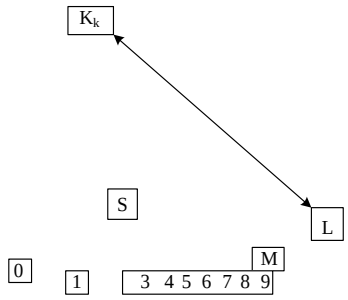

Figure 3. Example diagram of the relationship between the criterion function and L.

3.1.5 K-nearest neighbor clustering algorithm. The k-nearest neighbor clustering algorithm is a classical nonparametric clustering algorithm. Its implementation and application are shown in Figure 4:

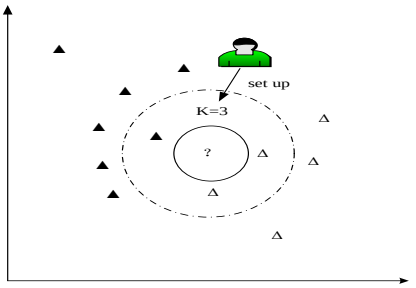

Figure 4. Implementation and application of the nearest neighbor algorithm

Assuming the sample data set $A=\{a_1,a_2,\cdots,a_m\}\in T^f$, the k-nearest neighbor clustering algorithm is specifically described as follows:

$$g:T^m\to n_o,1\le o\le z \tag{19}$$

that is,

$$\tilde{g}(a_o)=\arg\max_{n_o\in N}\sum_{o=1}^{l}\phi(n_o,g(a_o)) \tag{20}$$

Among them, is the likelihood estimation of $\tilde{g}(a_o)$ to $g(a_o)$, and the value of $\phi(n_o,g(a_o))$ is:

$$\phi(n_o,g(a_o))=\begin{cases}1 & n_o=g(a_o)\\ 0 & otherwise\end{cases} \tag{21}$$

The standard k-nearest neighbor clustering algorithm overcomes the disadvantage of the high error rate of the nearest neighbor clustering method. It has high clustering performance for sample data sets that conform to normal distribution or unknown. The algorithm can directly cluster the sample data set, which avoids the error caused when processing the sample data set.

4. EXPERIMENTS ON COMPLEX TEMPORAL DATA MODEL AND EXTRACTION TECHNOLOGY OF MULTI-SPATIAL DATABASE

4.1 Construction of the data model

(1) Data sources

This experiment is to verify the three indicators of the algorithm (k-means clustering algorithm), the Readability algorithm, the neural network algorithm and the distributed algorithm in the precision rate (precision) and recall rate R (Recall) and the final comprehensive evaluation index F for comparison.

(2) The experimental data is mainly selected from 100 pages each from People's Daily Online, Dagong.com, Sina.com and Netease News. The 100 web pages selected from Sina.com are from the same category and at the same time. The web pages selected from NetEase News are 100 web pages from the same category. Each 100 web pages selected from People's Daily Online and Dagong.com were randomly selected.

(3) Lab environment

RedHatEnterpriseLinuxSarverrelkase7.2(Maipo) operating system,

Ine(R)Xon(R)CPUES-26300@230GHz processor, 8G memory, using Pyhon language to write code.

```
# k-means
from numpy import unique
```

```
from numpy import where
from sklearn.datasets import make_classification
from sklearn.cluster import KMeans
from matplotlib import pyplot
X, _ = make_classification(n_samples=1000, n_features=2, n_informative=2, n_redundant=0, n_clusters_per_class=1, ra
ndom_state=4)
model = KMeans(n_clusters=2)
model.fit(X)
yhat = model.predict(X)
clusters = unique(yhat)
for cluster in clusters:
row_ix = where(yhat == cluster)
pyplot.scatter(X[row_ix, 0], X[row_ix, 1])
pyplot.show()
```

(4) Experimental Results

The experimental k-means clustering algorithm, the Readability algorithm, the layout similarity algorithm and the block algorithm are compared on the three indicators of precision (precision) and recall (Recall) and the final comprehensive evaluation index F. The specific results are shown in Tables 1-3:

Table 1. Comparison of precision rates P.

Website	K-means clustering algorithm	Neural network algorithm	Distributed algorithms	Readability algorithm
Sina.com	98.54%	98.95%	99.57%	91.63%
Netease.com	98.69%	96.59%	97.22%	92.89%
People's network	97.87%	93.02%	91.22%	92.75%
Dagong network	97.95%	93.25%	90.63%	91.30%

Table 2. Comparison of recall rate R.

Website	K-means clustering algorithm	Neural network algorithm	Distributed algorithms	Readability algorithm
Sina.com	97.75%	97.04%	99.53%	92.42%
Netease.com	97.22%	96.16%	97.90%	93.12%
People's network	96.59%	93.09%	92.83%	92.80%
Dagong network	96.24%	92.74%	91.13%	94.16%

Table 3. Comprehensive evaluation F.

Website	K-means clustering algorithm	Neural network algorithm	Distributed algorithms	Readability algorithm
Sina.com	98.13%	97.99%	99.55%	92.02%
Netease.com	97.95%	96.37%	96.56%	93.00%
People's network	97.22%	93.05%	92.02%	92.77%
Dagong network	97.59%	92.99%	90.88%	92.71%

4.2 Construction of technical extraction

(1) Data sources

The experiment was divided into two groups. The first set of experiments is the comparison of these three indicators that a method for extracting the main content of a page based on the paragraph extractor ParEx, and a method for extracting news content based on the visual unit of a web page, and in the precision and recall R (Recall) and the final comprehensive evaluation index F.

In order to fully test the feasibility of this algorithm to extract the main content on thematic web pages, this experiment selects 100 web pages from Sina, Tencent, Sohu and NetEase as experimental data.

(2) Lab environment

RedHatEnterpriseLinuxSarverrelkase7.2(Maipo) operating system,
Ine(R)Xon(R)CPUES-26300@230GHz processor, 8G memory, using Pyhon language to write code.

(3) Analysis of results

The first group of experiments compares the k-means clustering algorithm (hereinafter referred to as the KM algorithm) and the method based on the paragraph extractor ParEx to extract the main content of the page (hereinafter referred to as the ParEx algorithm) and the performance of a method for extracting news content based on the visual unit of web pages (hereinafter referred to as the vu algorithm) on the precision rate and recall rate R and the final comprehensive evaluation index F. The experimental results are expressed in the form of percentages, with two decimal places reserved. The experimental results are shown in Figures 5-7:

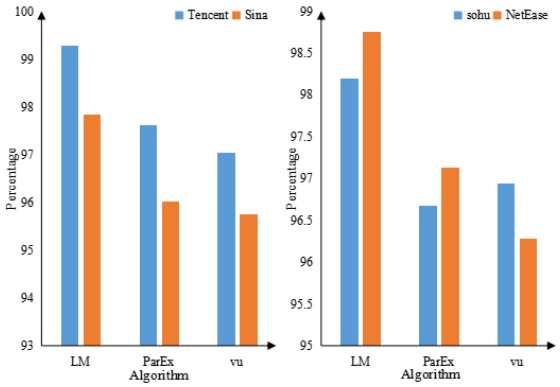

Figure 5. Comparison of precision P.

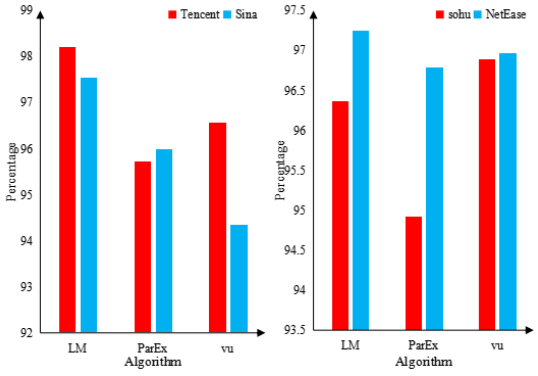

Figure 6. Recall R comparison.

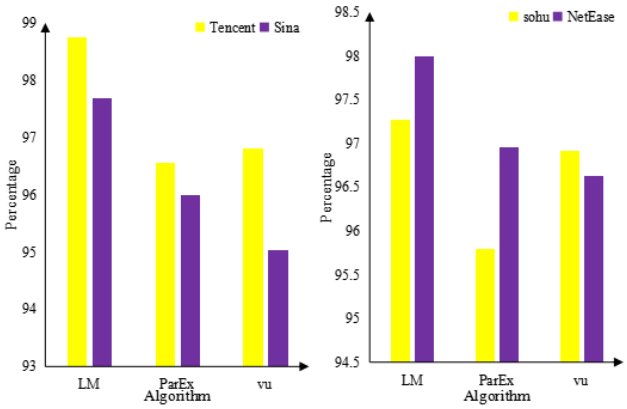

Figure 7. Composite performance f.

The second group compares the k-means clustering algorithm proposed in this chapter with the method of extracting the main content of the page based on paragraph extractor ParEx and a method of extracting news content based on the visual unit of the page in processing time, required for these three methods to extract the main content on the web pages of Sina, Tencent, Sohu, and NetEase. The experimental results are expressed in seconds, with two decimal places reserved. The results are shown in Figure 8:

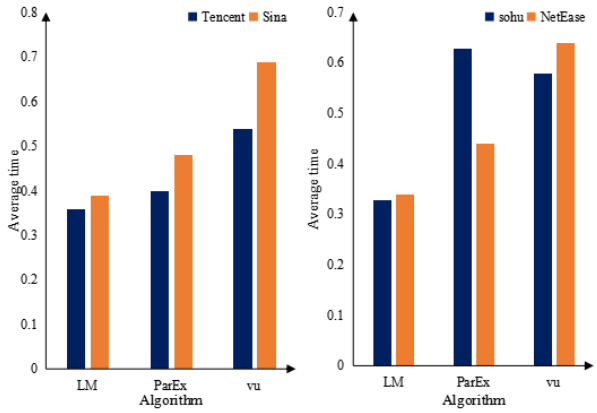

Figure 8. Time comparison of algorithms.

4. CONCLUSION

This paper mainly studies the influence between clustering algorithm and complex temporal data model and extraction technology of multi-spatial database. The content of this research is mainly to improve the utilization rate of multi-spatial database, which points out its future development. By using different types of algorithms to study the complex temporal data model and extraction technology of spatial database, and analyzing the three indicators of precision rate P, recall rate R and comprehensive evaluation index F, it can be seen that these algorithms have good performance, but the k-means clustering algorithm is more stable than other algorithms. Therefore, it is certain that the k-means clustering algorithm has a very good effect on improving the complex temporal data model and extraction technology of multi-spatial databases.

REFERENCES

[1] Zhang, M., Li X., Tang, M., Ma, L., Jia, C., Hu, X., et al., "Study on the construction of multi-source geological spatial database in 3D metallogenic prediction: A case study of Zhonggu orefield in Ningwu Basin," Scientia Geologica Sinica, 52(3):743-754(2017).

[2] Zhang, J., Li, S., Dong, R., Jiang, C., Ni, M., "Influences of land use metrics at multi-spatial scales on seasonal water quality: A case study of river systems in the Three Gorges Reservoir Area, China," Journal of Cleaner Production, 206 (PT.1-1156): 76-85(2019).

[3] Xie, Y., Crary, D., Bai, Y., Cui, X., Zhang, A., "Modeling Grassland Ecosystem Responses to Coupled Climate and Socioeconomic Influences in Multi-Spatial-And-Temporal Scales," Journal of Environment Informatics, 33(1):37-46(2019).

[4] Deng, W., Cheshmehzangi, A., Ma, Y., Peng, Z., "Promoting sustainability through governance of eco-city indicators: a multi-spatial perspective", International Journal of Low-Carbon Technologies, 15(2):1-12(2020).

[5] Ringelman, Kevin, M., Walker, Johann, James, k., et al., "Temporal and multi-spatial environmental drivers of duck nest survival," The Auk: Ornithological Advances, 135(3):486–494(2018).

[6] Cuzzocrea, A., Darmont, J., Mahboubi, H., "Fragmenting very large XML data warehouses via K-means clustering algorithm," International Journal of Business Intelligence & Data Mining, 4(3/4):301-328(2017).

[7] Chung, F., Simpson, O., "Computing Heat Kernel Pagerank and a Local Clustering Algorithm," European Journal of Combinatorics, 68(7):96-119(2017).

[8] Mohammad, S., "Novel linkage disequilibrium clustering algorithm identifies new lupus genes on meta-analysis of GWAS datasets," Immunogenetics, 69(5):295-302(2017).

[9] Wu, W., Xiong, N., Wu, C., "Improved clustering algorithm based on energy consumption in wireless sensor networks," IET Networks, 6(3):47-53(2017).

[10] Aminanto, M. E., Kim, H. J., Kim, K. M., Kim, K., "Another Fuzzy Anomaly Detection System Based on Ant Clustering Algorithm," Ieice Transactions on Fundamentals of Electronics Communications & Computer Sciences, 100(1):176-183(2017).

[11] Hu, Y., Niu, Y. G., Zou, Y. Y., "A zone-based unequal multi-hop clustering algorithm in WSNs," Kongzhi yu Juece/Control and Decision, 2017, 32(9): 1695-1700.

[12] Atashi, A., Nazeri, N., Abbasi, E., Dorri, S., Alijani-Z, M., "Breast Cancer Risk Assessment Using adaptive neuro-fuzzy inference system (ANFIS) and Subtractive Clustering Algorithm," Multidisciplinary Cancer Investigation, 1(2):20-26(2017).

[13] Neamatollahi, P., Naghibzadeh, M., "Distributed unequal clustering algorithm in large-scale wireless sensor networks using fuzzy logic," The Journal of Supercomputing, 74(6):2329-2352(2018).

[14] Wang, Z., Wang, K., Pan, S., Han, Y., "Segmentation of Crop Disease Images with an Improved K-means Clustering Algorithm," Applied Engineering in Agriculture, 34(2):277-289(2018).

[15] Fahim, A., "A Clustering Algorithm based on Local Density of Points," International Journal of Modern Education and Computer Science, 9(12):9-16(2017).

[16] Babichev, S. A., Gozhyj, A., Kornelyuk, A. I., Lytvynenko, V. I., "Objective clustering inductive technology of gene expression profiles based on sota clustering algorithm," Biopolymers & Cell, 33(5):379-392(2017).

[17] Kong, X., Hu, Q., Dong, X., Zeng, Y., Wu, Z., "Load Data Identification and Correction Method with Improved Fuzzy C-means Clustering Algorithm," Dianli Xitong Zidonghua/Automation of Electric Power Systems, 41(9):90-95(2017).

[18] Memon, K H., Lee, D H., "Generalised fuzzy c-means clustering algorithm with local information," Fuzzy Sets & Systems, 11(1):1-12(2018).

[19] Zhang, T., Ma, F., "Improved rough k-means clustering algorithm based on weighted distance measure with Gaussian function," International Journal of Computer Mathematics, 94(1-4): 663-675(2017).

[20] Yin, C., Lian, X., Sun, Z., Sun, R., Jin, W., "Improved clustering algorithm based on high-speed network data stream," Soft Computing, 22(4): 1-11(2017).