International Journal of Innovative
Computing, Information and Control
Volume **19**, Number **6**, December **2023**

ICIC International ©2023 ISSN 1349-4198
pp. 1877–1890

MULTI-SOURCE INFORMATION DATA INTEGRATION METHOD FOUNDED ON K-MEDOIDS CLUSTERING ALGORITHM

Liantian Li

Department of Information Engineering
Yangjiang Polytechnic
No. 213, Dongshan Road, Jiangcheng District, Yangjiang 529566, P. R. China
li2020202207@126.com

Received December 2022; revised April 2023

Abstract. *The Internet contains a large number of information resources, and it is an important way for people to obtain the information they need. However, the amount of information data on the Internet is huge and its sources are diverse, so data integration is required. Clustering algorithm is an important way to realize data integration. Aiming at the defect of initialization sensitivity in the K-medoids, an initialization strategy founded on improved Granular computing (IGC) is proposed. Aiming at the defect of poor convergence of the K-medoids, a granular iterative search strategy is proposed, and the fitness function of the K-medoids is improved. Finally, the research builds a multi-source information data integration model founded on the improved K-medoids. Using the data set in UCI to test the model, the average clustering precision of the model reaches 95.22%, the average time-consuming is 0.129 s, the average Rand index is 0.911, and the average F-measure value is 0.829, which are remarkablely better than other models. The above data verified the precision, productiveness and robustness of the proposed data integration model. It shows that the model can help people better integrate multi-source information data, so as to obtain the required information resources more efficiently and accurately.*

Keywords: K-medoids, Clustering algorithm, Multivariate information, Data integration, GC

1. **Introduction.** Nowadays, with the highly developed Internet, there are massive information and data resources. Information and data come from different websites, and the data formats are diverse, which makes it difficult for people to obtain the information they want from the data pile [1]. Therefore, a multi-source data integration model is constructed to unify, standardize and complete multi-source information data. It can lift the productiveness of Internet information [2,3]. The K-medoids is an improved algorithm founded on the K-means algorithm. It has better clustering effect and performance, and plays a better role in multi-source information data integration. However, the K-medoids has defects such as initialization sensitivity, poor convergence, and excessive iterations. GC can simplify problems and decrease the difficulty of solving complex problems [4]. The research combines GC with K-medoids, uses GC to find the initial center point of K-medoids, and solves the sensitive problem of K-medoids initialization. However, in the process of obtaining coarse-grained sets in traditional GC, it is easy to cause repeated classification of objects and duplication of some particles [5]. The study proposes strategies to improve GC. Finally, the study builds a multi-source information data integration model founded on the K-medoids optimized by IGC to help people obtain the required information resources more efficiently and accurately. There are two main innovations

DOI: 10.24507/ijicic.19.06.1877

in the research. One is to initialize the central point of the K-medoids by using IGC to improve the clustering precision and productiveness of the K-medoids. The second is to use the optimized K-medoids to realize data integration, which improves the utilization productiveness of information data. Compared with the most advanced data integration models at present, the data integration efficiency and data integration accuracy of the data integration model proposed in the study are significantly improved, and can play a better role in obtaining the required information resources. The test results show that the average clustering accuracy of the proposed model is 95.22%, the average time is 0.129 s, the average Rand index is 0.911, and the average F-measure value is 0.829, which are significantly better than other traditional data integration models. The test proves the advantages of the proposed model. The main content of the study is divided into four parts. The first part is to summarize the relevant research results. In the second part, a multi-source information data integration method is proposed based on K-medoids clustering algorithm. The third part is to verify the performance of the proposed multi-source information data integration method. The last part is the carding and summary of the article.

2. **Problem Statement and Preliminaries.** K-medoids has good clustering effect and performance, and plays a vital role in various scientific research. Johnson et al. used K-medoids to design probe sets for targeted sequencing of nuclear genes in flowering plants. The results showed that the probe set is valid [6]. Deng et al. proposed to use K-medoids to optimize the general collaborative filtering algorithm because of the defect of data sparsity, which leads to poor recommendation effect. The optimized collaborative filtering algorithm outperforms other comparison methods on various evaluation indicators [7]. Dinata et al. chose to use the purity algorithm to optimize the K-medoids for the defect that the K-medoids needs multiple iterations to cluster large data sets. The optimized algorithm validly decreases the number of iterations of the K-medoids [8]. Abbas and others used the K-medoids to cluster the birth data of Bard City, so as to provide a theoretical basis for medical security and medical decision-making in this area. The average precision of the K-medoids is 2.00% beyond that of the K-means algorithm [9]. Hashemzadeh and Zademehdi used the imperialist competitive algorithm to optimize the K-medoids to improve its robustness. And the ICA-K-medoids is applied to fire monitoring founded on video surveillance. The experiment proves that the algorithm has high detection precision, and each index is better than the general fire detection method [10]. Tiwari et al. optimized the K-medoids founded on the idea of armed bandit technology to solve the defect that the K-medoids has poor clustering effect in large data sets. Compared with before optimization, the productiveness of the optimized algorithm is increased by 3 times [11]. Li et al. combined time series generation confrontation network (TimeGAN), K-medoids founded on soft dynamic time warping, and convolutional neural network (CNN) to construct a photovoltaic power prediction model. Experimental data show that the precision of this model is beyond the other four models [12]. Wang et al. proposed a K-medoids founded on the distribution distance measure, and applied the K-medoids to data sequence clustering. The algorithm has a relatively good clustering effect on complex distribution data sequences [13].

In the information age, the most common way for people to obtain information resources is the Internet. However, information data on the Internet usually has multiple sources, and their data formats are often different, resulting in poor productiveness for people to obtain the information they need. Therefore, unifying the data format of multi-source information through data integration technology will help users obtain information resources more conveniently and quickly. Argelaguet et al. studied the effect of data

integration in the analysis of single-cell multimodal data. Through data integration, single-cell multimodal data can be analyzed more validly, thus providing a tool for the study of cell heterogeneity [14]. Isaac et al. explored the application of data integration in the context of the increasing volume and type of biodiversity data collected. The results show that data integration plays a vital role in the distribution prediction of species [15]. Zipkin et al. used data integration methods to study large-scale ecological phenomena. And for the defects in data integration in the study of ecological phenomena, such as data size mismatch, and data imbalance, a targeted strategy is proposed [16]. Luecken et al. relied on data integration for joint analysis of atlas-level datasets in single-cell genomics [17]. Boehm et al. used multimodal data integration to realize the integration and analysis of cross-modal clinical data, thus promoting the development of precision oncology [18]. Canzler et al. improved the understanding of toxicology research through multi-omics data integration. They founded on data from published related literature, and it was demonstrated that multi-omics data integration can remarkablely increase confidence in detecting pathway responses [19]. He et al. proposed a coronavirus data integration, sharing and analysis ontology. This ontology covers various aspects, such as etiology, and transmission [20]. Founded on the advantages of data integration that can process information from multiple datasets at the same time, Miller et al. used data integration models to predict species distribution. This method can more accurately predict the dynamic process of species distribution [21].

The K-medoids and data integration are widely used in various fields. However, there is almost no research on combining K-medoids with data integration technology to improve the utilization productiveness of multi-source information data. To this end, the study uses IGC to optimize the K-medoids, and proposes a multi-source information data integration method founded on the optimized K-medoids to achieve data integration and improve the utilization productiveness of information data.

3. Multi-Source Data Integration Founded on Optimized K-Medoids.

3.1. K-medoids initialization founded on GC.

The development of Internet technology has resulted in an extremely large amount of data collection. To obtain valid information in huge data sets, it is very necessary to conduct data mining on data resources. Data mining can improve the utilization rate of data resources, improve users' ability to mine hidden information in data sets, and facilitate data integration. Clustering analysis is an important part of data mining technology. Several common clustering algorithms and their performance comparison are presented in Table 1.

It can be seen that each clustering algorithm has different performances in different application scenarios, and there is no absolute distinction between good and bad. K-medoids is an improved algorithm founded on K-means algorithm, which has good clustering effect and performance, and plays a vital role in various scientific researches. The clustering process of the K-medoids is shown in Figure 1. The specific process is as follows: firstly, K data are randomly selected from n data objects as the cluster center; after initializing K cluster centers, allocate the remaining data objects to each cluster; then replace the original cluster center point and calculate the new variance E. The data object with the smallest E is selected as the new cluster center. Next, whether the cluster center has changed is to be confirmed. If there is no change, output the clustering results. If there is a change, assign the data objects to each cluster, and continue to repeat the above operation.

However, the K-medoids has the defect of sensitivity to initialization. The selection strategy of the initial center point of the K-medoids is random selection. If the initial

TABLE 1. Common clustering algorithms and performance comparison

Algorithm	K-means [22]	K-medoids [23]	CLARANS [24]	BIRCH [25]	DBSCAN [26]	STING [27]	SOM [28]
Algorithm productiveness	Higher	General	Lower	High	General	High	General
Noise sensitivity	Sensitive	Insensitive	Insensitive	General	Insensitive	Insensitive	Sensitive
Data input order sensitivity	Less sensitive	Less sensitive	Very sensitive	Less sensitive	Sensitive	Insensitive	Sensitive
Cluster shape	Spherical	Spherical	Convex or spherical	Convex or spherical	Any shape	Any shape	Any shape
High dimensionality	Good	General	General	Good	General	Good	Good
Scalability	Better	Poor	Good	Poor	Better	Good	Better

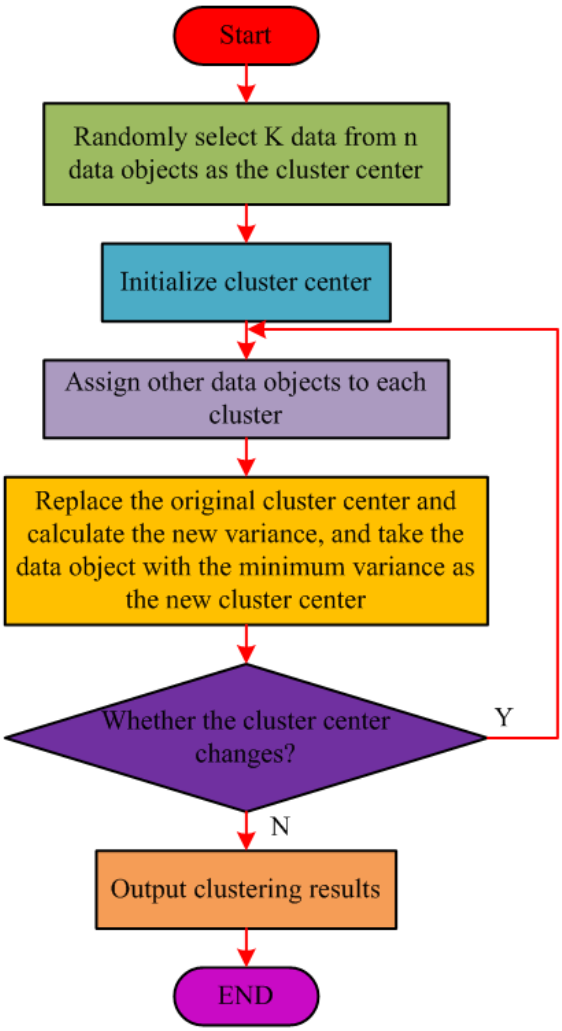

FIGURE 1. Clustering flow of K-medoids

center point is not selected properly, the algorithm may fall into a local extremum. Therefore, improving the initialization of the K-medoids can validly enhance the productiveness and precision of the K-medoids. The study proposes an initialization strategy founded on GC. Suppose there is a clustering space $T = (U, B)$, in which B represents the set of attributes and U represents the set of objects. Then the object similarity $S(x_i, x_j)$ can be expressed by Formula (1).

$$S(x_i, x_j) = \frac{1}{\left(1 + \sum_{l=1}^{|B|} w_l \left| x_{il} - x_{jl} \right|\right)} \tag{1}$$

In Formula (1), l is the attribute value. w_l represents the attribute resolution, which is used to describe the weight of attribute value l in the object set. Assuming that there are n objects in the object set. The average similarity $\bar{d}$ of objects can be expressed by Formula (2).

$$\bar{d} = \sum_{i,j}^{n} S(x_i, x_j) \Big/ n^2 \tag{2}$$

Supposing the particles in the object set are divided into $\{X_1, X_2, \ldots, X_n\}$, then the density of the particles in the object set is $gd(X_i)$, which is expressed as Formula (3).

$$gd(X_i) = \frac{|X_i|}{|U|} \tag{3}$$

Average density $\overline{GD}$ is represented by Formula (4).

$$\overline{GD} = \frac{\sum_{i=1}^{n} gd(X_i)}{n} \tag{4}$$

Suppose there is an object set, in which particle i contains N objects, namely $x_{i1}, x_{i2}, \ldots, x_{iN}$, then the center o_i of particle i is expressed by Formula (5).

$$o_i = \left\{ x_{ij} \left| \min_{j=1}^{N} \left| x_{ij} - \frac{1}{N} \sum_{k=1}^{N} x_{ik} \right| \right. \right\} \tag{5}$$

The basic calculation process of GC is to first calculate the similarity and average similarity of all objects according to Formula (1) and Formula (2). Suppose there is a threshold value d, and there is $92\%\bar{d} \leq d \leq 108\%\bar{d}$. When two objects are compared, $S(x_i, x_j) \geq d$. If there is a fuzzy similarity matrix between the objects $M(i, j) = 1$, it indicates that the objects j are similar objects of the objects i. The fuzzy similarity matrix between objects is in Formula (6).

$$M = \begin{bmatrix} 101 & 010 & \ldots & 110 & \ldots \\ 110 & 101 & \ldots & 011 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 101 & 110 & 011 & 001 & \ldots \end{bmatrix} \tag{6}$$

All objects similar to objects i are sorted into coarse-grained sets for numbering. A collection of any object G and its similar objects is collectively called a particle. Calculate the average density of all particles in the coarse-grained set using Formula (3) and Formula (4). The particles are called valid particles, and the collection of all valid particles is called a valid particle set H. Calculate and record the centers of all particles in the valid particle set and the Euclidean distance between any two particles according to Formula (5). The center of the maximum particle density in the valid particle is set as the center point 1, denoted as o_1, and the corresponding particle is Q_1. The center point of the particle that is farthest away from the particle and has the highest density is taken as the

second center point, denoted as o_2. The corresponding particle is Q_2. By analogy, each particle center and corresponding particle in the valid particle set are obtained. Find the distance of $o_1, o_2, \ldots, o_n$ to each particle center point $d_{i1}, d_{i2}, \ldots, d_{in}$, respectively. Make $d_i = \min(d_{i1}, d_{i2}, \ldots, d_{in})$, and calculate $\max(d_i)$. Its corresponding particle center is o_i, and the corresponding particle is Q_i, which is used as the first initial clustering point of the K-medoids. By analogy, the k center of the first particle, the corresponding particle and the initial clustering point of the K-medoids are obtained. Founded on the above content, the K-medoids initialization founded on GC is realized.

3.2. **K-medoids initialization founded on IGC.** In the process of obtaining a coarse-grained set G, it is easy to cause repeated classification of objects and duplication of some particles. In this case, there may be some valid particles with the same number of objects and the same object attributes but different particle numbers in the valid particle set. H will cause the center points of the subsequent search to appear in the same cluster, which will cause the initial center of the K-medoids to be in the same cluster. The clustering effect of the algorithm will be impacted. In response to this defect, the study improves GC. First of all, in the comparison of objects, the strategy of comparing one object with the subsequent objects one by one is not adopted. Thus, an upper triangular matrix is obtained, as shown in Formula (7).

$$M = \begin{bmatrix} 101 & 010 & \ldots & \ldots & 110 \\ 110 & 101 & \ldots & \ldots & 011 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 101 & 110 & 011 & \ldots & 001 \end{bmatrix} \tag{7}$$

Through this strategy, the phenomenon of repeated classification of particle objects can be avoided, and particles with the same number of objects and attribute values in particles can be removed. In addition, this strategy enables the valid particles to be arranged in order according to the number of objects, thus obtaining a fine-grained set $\bar{G}$. The initial centers of the final K-medoids are located in different clusters, which improves the clustering precision. The K-medoids has poor convergence and needs more update iterations to get the desired result. The reason for this situation is that the K-medoids performs a global search selection every time the center point is updated, and the amount of calculation is relatively large. Therefore, it needs to be further improved to obtain an efficient center point search update strategy. The study proposes a granular iterative search strategy. The core idea of this strategy is to search and update the center point in the particles corresponding to the initial center point. That is o_k, and the update of the center point is searched and selected in its corresponding valid particles Q_k. This strategy can validly decrease the search range of the center point, thereby decreasing the number of iterations of the K-medoids. Under this strategy, the intra-cluster distance is calculated using Formula (8).

$$E(w) = \sum_{i=1}^{K} \sum_{p \in c_i} |p - o_i| \tag{8}$$

In Formula (8), c_i represents the ith cluster, p is the object in the ith cluster, and o_i is the clustering center of the ith cluster. p and o_i are multidimensional objects. The inter-cluster distance is calculated using Formula (9).

$$O(w) = \sum_{i,j=1}^{K} |o_i - o_j| \tag{9}$$

In Formula (9), o_i and o_j represent cluster centers, and both are multidimensional. The fitness function of the K-medoids is in Formula (8). The idea of the fitness function is to assess the final clustering quality of the algorithm by using the sum of intra-cluster distances. However, Formula (8) does not take account of the inter-cluster distance. The optimal result of clustering quality requires the minimum intra-cluster distance and the maximum inter-cluster distance. Formula (8) only considers the intra-cluster distance, so using it as a fitness function may make the final clustering result of the K-medoids not the optimal clustering result. For this reason, the study proposes an improved fitness function, as shown in Formula (10).

$$F(w) = \frac{O(w)}{E(w)} \tag{10}$$

As shown in Formula (10), during the experiment, the fitness function continuously adjusts the function value. When the function value reaches the maximum value, it indicates that the algorithm has obtained the optimal clustering result, and the iteration can be stopped. The basic flow of the improved K-medoids is shown in Figure 2. The specific process is as follows: input a dataset containing n objects and k clusters, initialize k cluster centers $o_1 \sim o_k$ with improved granularity calculation, and record the particles $Q_1 \sim Q_k$ corresponding to the center point. Divide the remaining objects into clusters represented by the center point of the nearest object, mark them as w, and calculate $F(w)$. Select a non-representative object O_{random} from the corresponding particles $Q_1 \sim Q_k$ to replace the

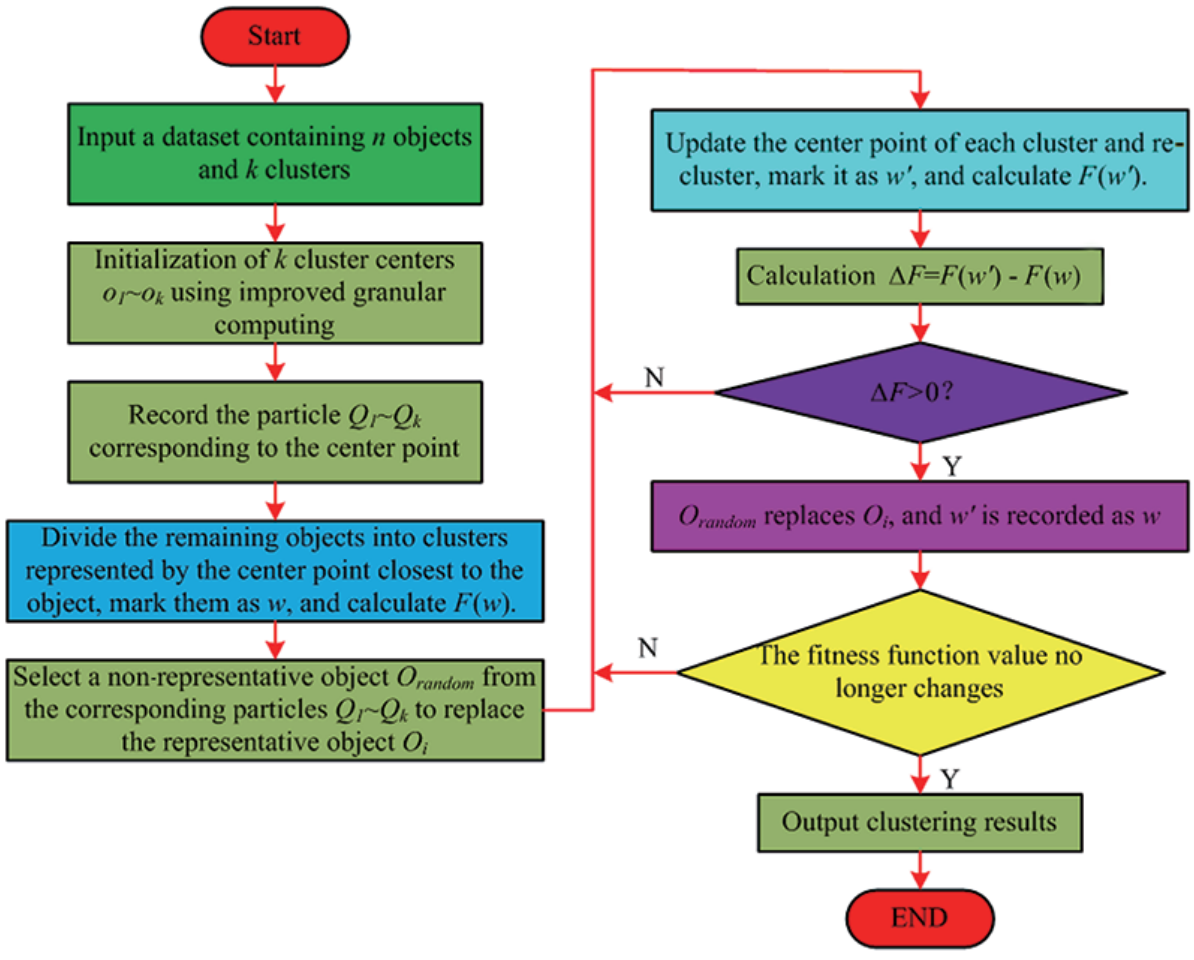

FIGURE 2. Basic flow of improved K-medoids

representative object O_i. Update the center point of each cluster and re-cluster it, mark it as w', and calculate $F(w')$. Calculate $\Delta F = F(w') - F(w)$, if ΔF is less than or equal to 0, continue to update the cluster center and calculate ΔF after re-clustering. If ΔF is greater than 0, replace O_i with O_{random} and record w' as w. Judge whether the fitness function value changes. If it changes, update the cluster center and cluster again. If it does not change, output the cluster result.

3.3. **Multi-source information integration model founded on improved K-medoids.** In the information age, the most common way for people to obtain information resources is the Internet. However, information data on the Internet usually has multiple sources, and their data formats are often different, resulting in poor productiveness for people to obtain the information they need. To this end, the study builds a multi-source information integration model founded on the improved K-medoids. The multi-source information integration model founded on the improved K-medoids is mainly divided into two parts: data extraction module and data integration module. The data extraction module is mainly composed of a parser and an extractor. The data integration module mainly clusters and filters the extracted data by improving the K-medoids to obtain valid information.

In the part of the parser, since the internal structure of the HTML web page is similar to a tree structure, the HTML document can be constructed into a tree structure through special attribute tag items. The parser is used to connect to the URL. All the table nodes, tr nodes, and td nodes of the webpage through the parsing tool are parsed to save them in a tree structure, and perform structural filtering and semantic pruning on them.

After the parser obtains all the td nodes of the webpage, it transmits them to the extractor. In the extractor part, these td nodes are traversed, and the corresponding text information and link information are extracted. In the process of extraction, the initial identifier of the data source code is mainly used as a marker, and it is realized through string matching. The extracted data is stored in the corresponding memory data table after marking attributes to avoid data confusion.

The data in the data table of the memory is not all the information data required by the user, so it needs to be screened and the information required by the user needs to be integrated. However, the amount of information and data in the Internet is extremely large, and manual screening of data is basically impossible. The improved K-medoids is used to cluster the information data to obtain similar information data. The implementation method is to cluster the attributes of the extracted data founded on the improved K-medoids, and integrate the data corresponding to each matching attribute and provide it to the user. The similarity of attributes is generally calculated in two ways. The first is to compute the linguistic similarity of attributes. The second is calculated by the linear sum of the numerical type similarity of the data and the product of the weight. The flow of the data integration module is shown in Figure 3. The specific process includes: clustering attributes according to known database style sheet attributes; then the obtained clustering result information is transferred to the data storage; finally, the matching data is sent to the database according to the clustering results.

Founded on the above contents, a multi-source information integration model with the improved K-medoids is constructed. The basic flow of the multi-source information integration model is shown in Figure 4.

In Figure 4, firstly, the parser is used to connect to the URL seed, and the corresponding web page is parsed and positioned, so as to collect table points, tr points, and td points, and obtain the information data required by the user. Secondly, the extractor is used to extract the data information in each leaf node of the EC tree, and store the data

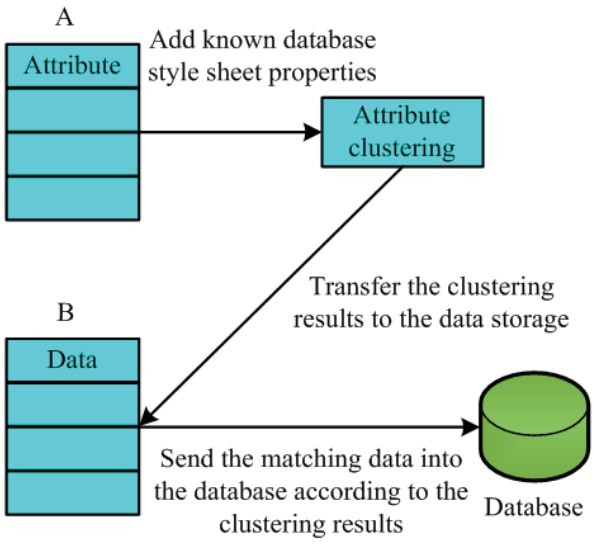

FIGURE 3. Process of data integration module

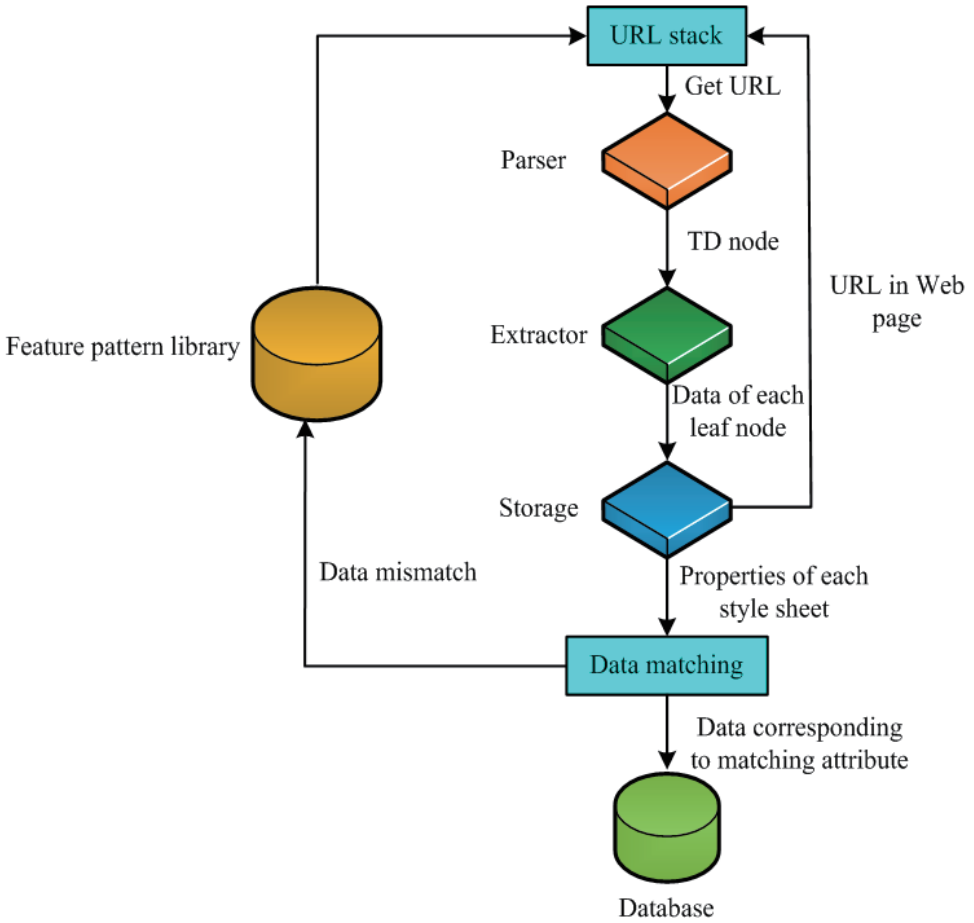

FIGURE 4. Basic process of multi-source information integration model

information in the memory. Finally, because the various data in the memory are not all the data required by the user, the improved K-medoids is used to cluster the data and integrate them into data corresponding to each matching attribute.

4. **Performance Analysis of Multi-Source Information Integration Model.** To verify the performance of the proposed multi-source information integration model founded on the improved K-medoids, a comparative experiment was designed. The experimental environment is as follows: Intel(R) Core(TM) i5-8265U CPU, memory 8G, the operating system is windows10. The integrated development softwares are Microsoft Visual C++ 6.0 and Matlab7.0. The data set used in the experiment is the data set in UCI, as shown in Table 2.

Table 2. UCI dataset

Data set name	Number of samples	Number of attributes	Cluster number
Wine	178	13	3
Bupa	345	7	2
Ionosphere	351	34	2
Iris	150	4	3
Wavaform40	5000	40	3
Heart	270	13	2
Wdbc	569	30	2

In the above data set, trial the precision of several traditional data integration models founded on clustering algorithms, including PAM (partitioning around medoid), genetic fuzzy clustering algorithm (GAFKM), K-means algorithm and the proposed model (Model 1). Each data set is trialed 10 times, and the final result is the average value of 10 trial values. The accuracies of several traditional data integration models are presented in Table 3.

Table 3. Precision of several common clustering algorithms

Data set name	Algorithm name			
	PAM	GAFKM	K-means	Model 1
Wine	0.7752	0.9314	0.7612	0.9693
Bupa	0.8214	0.9144	0.7844	0.9840
Ionosphere	0.8011	0.9008	0.7820	0.9304
Iris	0.8514	0.8941	0.8145	0.9515
Wavaform40	0.8135	0.9106	0.8233	0.9211
Heart	0.8942	0.9255	0.8536	0.9310
Wdbc	0.8709	0.9417	0.8331	0.9778
Average	0.8325	0.9169	0.8074	0.9522

As can be seen in Table 3, in each data set, the clustering precision of Model 1 is remarkablely beyond that of the other three models. The clustering effect of the proposed algorithm is better than the general clustering algorithm. Among them, Model 1 has the highest average clustering precision of 95.22%; the average clustering precision of PAM is 83.25%, which is 11.97% below Model 1. The average clustering precision of K-means is 80.74%, which is 14.48% below Model 1. The average clustering precision of GAFKM is 91.69%, which is 3.53% below Model 1. The above data sets are used to test the clustering productiveness of the models, as shown in Table 4.

As can be seen in Table 4, in each data set, the clustering time consumption of Model 1 is remarkablely below that of the other three models. The average time consumption of Model 1 on each data set is 0.129 s. The average time-consuming of PAM on each data set is 0.200 s, which is 0.071 s more than Model 1. The average time-consuming of

TABLE 4. Time consuming for clustering of several algorithms (s)

Data set name	Algorithm name			
	PAM	GAFKM	K-means	Model 1
Wine	0.233	0.125	0.144	0.110
Bupa	0.184	0.138	0.161	0.135
Ionosphere	0.182	0.133	0.156	0.124
Iris	0.210	0.151	0.175	0.145
Wavaform40	0.191	0.146	0.157	0.133
Heart	0.196	0.153	0.189	0.142
Wdbc	0.203	0.129	0.162	0.116
Average	0.200	0.139	0.163	0.129

GAFKM on each data set is 0.139 s, which is 0.010 s more than Model 1. The average time-consuming of K-means is 0.163 s, which is 0.034 s more than Model 1. Five sets of artificial data sets were generated using Matlab tools to verify the robustness of several models. Each artificial dataset is divided into 4 categories, the sample data size of each category is 500, and the attribute dimension is 10 dimensions. Add 15%, 20%, 25%, 30%, and 35% noise data to the five artificial datasets, and the final artificial datasets are presented in Table 5.

TABLE 5. 5 sets of manual data sets

Data set	Number of samples	Number of attributes	Cluster number
15%	2300	10	4
20%	2400	10	4
25%	2500	10	4
30%	2600	10	4
35%	2700	10	4

The robustness of the data integration models was assessed using Rand index and F-measure value. The experimental results are the average of the results of each model running 10 times on each data set. The Rand exponents of several models are presented in Table 6.

TABLE 6. Rand index of several clustering algorithms

Data set	Algorithm name			
	PAM	GAFKM	K-means	Model 1
15%	0.833	0.923	0.854	0.934
20%	0.829	0.916	0.905	0.923
25%	0.825	0.894	0.854	0.910
30%	0.893	0.886	0.866	0.898
35%	0.820	0.872	0.864	0.892
Average	0.840	0.898	0.869	0.911

In Table 6, on each data set, the Rand index of Model 1 is remarkablely beyond the other three models. The average Rand index of Model 1 on each dataset is 0.911. The average Rand index of PAM on each data set is 0.840, which is 0.071 below that of Model 1. The average Rand index of GAFKM on each data set is 0.898, which is 0.013 below

TABLE 7. F-measure value of several clustering algorithms

Data set	Algorithm name			
	PAM	GAFKM	K-means	Model 1
15%	0.724	0.872	0.744	0.872
20%	0.708	0.840	0.810	0.851
25%	0.681	0.818	0.741	0.822
30%	0.677	0.786	0.747	0.803
35%	0.675	0.771	0.754	0.796
Average	0.693	0.817	0.759	0.829

that of Model 1. The average Rand index of K-means on each data set is 0.869, which is 0.042 below Model 1. The F-measure values of several models are presented in Table 7.

In Table 7, on each data set, the F-measure value of Model 1 is remarkablely beyond that of the other three models. The average F-measure value of Model 1 on each dataset is 0.829. The average F-measure value of PAM on each data set is 0.693, which is 0.136 below Model 1. The average F-measure value of GAFKM on each data set is 0.817, which is 0.012 below Model 1. The average F-measure value of K-means on each data set is 0.759, which is 0.070 less than Model 1. The results in Table 6 and Table 7 can prove that the proposed multi-source information integration model founded on the improved K-medoids is more robust. In summary, the multi-source information integration model founded on the improved K-medoids has high precision, productiveness and anti-interference ability, and provides users with a convenient, accurate and efficient data integration approach.

5. **Conclusion.** The information on the Internet comes from multiple websites, the amount of data is huge, it is difficult for people to filter out the information data that they are interested in. To this end, the study proposes an improved K-medoids, founded on which a source information integration model is built, and the performance of the model is trialed using the data set in UCI. The average clustering precision of Model 1 is 95.22%, 11.97% beyond PAM. It is 14.48% beyond K-means, 3.53% beyond GAFKM. This shows that the precision of Model 1 is ideal. The average time consumption of Model 1 on each data set is 0.129 s, which is 0.071 s less than PAM; 0.010 s less than GAFKM, 0.034 s less than K-means. This shows that the data integration productiveness of Model 1 is higher. Construct artificial datasets and add noise data of different scales to verify the robustness of the model. The average Rand index of Model 1 on each data set is 0.911, which is 0.071 beyond PAM; 0.013 beyond GAFKM, 0.042 beyond K-means. The average F-measure value of Model 1 on each data set is 0.829, which is 0.136 beyond PAM, 0.012 beyond GAFKM, 0.070 beyond K-means. This shows that the multi-source information integration model founded on the improved K-medoids proposed by the study has better robustness. It can be known that compared with the existing multi-information data integration model, the research on the multi-source information integration model founded on the improved K-medoids has higher precision, productiveness and anti-interference ability, and can provide users with more convenient and efficient data integration. The research has only verified the theoretical application of the model, and further research is needed on the practical application of the model.

REFERENCES

[1] I. K. Nti, A. F. Adekoya and B. A. Weyori, A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction, *Journal of Big Data*, vol.8, no.1, pp.1-28, 2021.

[2] X. Che, J. Mi and D. Chen, Information fusion and numerical characterization of a multi-source information system, *Knowledge-Based Systems*, vol.145, pp.121-133, 2018.
[3] S. Cheng, B. Zhang, G. Zou, M. Q. Huang and Z. Zhang, Friend recommendation in social networks based on multi-source information fusion, *International Journal of Machine Learning and Cybernetics*, vol.10, no.5, pp.1003-1024, 2019.
[4] Y. Yao, Three-way granular computing, rough sets, and formal concept analysis, *International Journal of Approximate Reasoning*, vol.116, pp.106-125, 2020.
[5] X. Yang, T. Li, D. Liu and H. Fujita, A multilevel neighborhood sequential decision approach of three-way granular computing, *Information Sciences*, vol.538, pp.119-141, 2020.
[6] M. G. Johnson, L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, N. Epitawalage, F. Forest, J. T. Kim, J. H. Leebens-Mack, L. J. Leitch, O. Maurin, D. E. Soltis, P. S. Soltis, G. K. S. Wong, W. J. Baker and N. J. Wickett, A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering, *Systematic Biology*, vol.68, no.4, pp.594-606, 2019.
[7] J. Deng, J. Guo and Y. Wang, A novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering, *Knowledge-Based Systems*, vol.175, pp.96-106, 2019.
[8] R. K. Dinata, S. Retno and N. Hasdyna, Minimization of the number of iterations in k-medoids clustering with purity algorithm, *Revue d'Intelligence Artificielle*, vol.35, no.3, pp.193-199, 2021.
[9] S. A. Abbas, A. Aslam, A. U. Rehman, W. A. Abbasi, S. Arif and S. Z. H. Kazmi, K-means and k-medoids: Cluster analysis on birth data collected in city Muzaffarabad, Kashmir, *IEEE Access*, vol.8, pp.151847-151855, 2020.
[10] M. Hashemzadeh and A. Zademehdi, Fire detection for video surveillance applications using ICA K-medoids-based color model and efficient spatio-temporal visual features, *Expert Systems with Applications*, vol.130, pp.60-78, 2019.
[11] M. Tiwari, M. J. Zhang, J. Mayclin, S. Thrun, C. Piech and I. Shomorony, BanditPAM: Almost linear time k-medoids clustering via multi-armed bandits, *Advances in Neural Information Processing Systems*, vol.33, pp.10211-10222, 2022.
[12] Q. Li, X. Zhang, T. Ma, D. Liu, H. Wang and W. Hu, A multi-step ahead photovoltaic power forecasting model based on TimeGAN, Soft DTW-based K-medoids clustering, and a CNN-GRU hybrid neural network, *Energy Reports*, vol.8, pp.10346-10362, 2022.
[13] T. Wang, Q. Li, D. J. Bucci, Y. B. Liang, B. Chen and P. K. Varshney, K-medoids clustering of data sequences with composite distributions, *IEEE Transactions on Signal Processing*, vol.67, no.8, pp.2093-2106, 2019.
[14] R. Argelaguet, A. S. E. Cuomo, O. Stegle and J. C. Marioni, Computational principles and challenges in single-cell data integration, *Nature Biotechnology*, vol.39, no.10, pp.1202-1215, 2021.
[15] N. J. B. Isaac, M. A. Jarzyna, P. Keil, L. I. Dambly P. H. Boersch-Supan, E. Browning, S. N. Freeman and N. Golding, Data integration for large-scale models of species distributions, *Trends in Ecology & Evolution*, vol.35, no.1, pp.56-67, 2020.
[16] E. F. Zipkin, E. R. Zylstra, A. D. Wright, S. P. Saunders, A. O. Finley, M. C. Dietze, M. S. Itter and M. W. Tingley, Addressing data integration challenges to link ecological processes across scales, *Frontiers in Ecology and the Environment*, vol.19, no.1, pp.30-38, 2021.
[17] M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danes, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché and F. J. Theis, Benchmarking atlas-level data integration in single-cell genomics, *Nature Methods*, vol.19, no.1, pp.41-50, 2022.
[18] K. M. Boehm, P. Khosravi, R. Vanguri, J. J. Gao and S. P. Shah, Harnessing multimodal data integration to advance precision oncology, *Nature Reviews Cancer*, vol.22, no.2, pp.114-126, 2022.
[19] S. Canzler, J. Schor, W. Busch, K. Schubert, U. E. Rolle-Kampczyk, H. Seitz, H. Kamp, M. V. Bergen, R. Buesen and J. Hackermüller, Prospects and challenges of multi-omics data integration in toxicology, *Archives of Toxicology*, vol.94, no.2, pp.371-388, 2020.
[20] Y. He, H. Yu, E. Ong, Y. Wang, Y. T. Liu, A. Huffman, H. H. Huang, J. Beverley, J. Hur, X. L. Yang, L. Chen, G. S. Omenn, B. Athey and B. Smith, CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis, *Scientific Data*, vol.7, no.1, pp.1-5, 2020.
[21] D. A. W. Miller, K. Pacifici, J. S. Sanderlin and B. J. Reich, The recent past and promising future for data integration methods to estimate species' distributions, *Methods in Ecology and Evolution*, vol.10, no.1, pp.22-37, 2019.

[22] M. K. Gupta and P. Chandra, Effects of similarity/distance metrics on k-means algorithm with respect to its applications in IoT and multimedia: A review, *Multimedia Tools and Applications*, vol.81, no.26, pp.37007-37032, 2022.
[23] B. Lund and J. Ma, A review of cluster analysis techniques and their uses in library and information science research: k-means and k-medoids clustering, *Performance Measurement and Metrics*, vol.22, no.3, pp.161-173, 2021.
[24] T. M. Hossain, J. Watada, I. A. Aziz, M. Hermana, S. T. Meraj and H. Sakai, Lithology prediction using well logs: A granular computing approach, *International Journal of Innovative Computing, Information and Control*, vol.17, no.1, pp.225-244, 2021.
[25] L. Sun, A strategic decision-making information fusion approach based on knowledge element relation mining, *ICIC Express Letters*, vol.15, no.12, pp.1319-1327, 2021.
[26] T. Ouyang and X. Shen, Online structural clustering based on DBSCAN extension with granular descriptors, *Information Sciences*, vol.607, pp.688-704, 2022.
[27] C. Manning, E. J. Kendon, H. J. Fowler, N. M. Roberts, S. Berthou, D. Sur and M. J. Roberts, Extreme windstorms and sting jets in convection-permitting climate simulations over Europe, *Climate Dynamics*, vol.58, nos.9-10, pp.2387-2404, 2022.
[28] Q. V. Doan, H. Kusaka, T. Sato and C. Chen, S-SOM v1.0: A structural self-organizing map algorithm for weather typing, *Geoscientific Model Development*, vol.14, no.4, pp.2097-2111, 2021.

Author Biography

Liantian Li received a bachelor's degree in Computer and Application from Guilin University of Electronic Technology, China, 2001; He received a master's degree in Computer Software and Theory from Sun Yat-sen University, China, 2009. He is currently a full-time associate professor of Yangjiang Polytechnic. His research interests include big data, algorithm analysis and artificial intelligence applications. He has published more than 20 papers in journals and conferences.