Application of Data Mining Technology with Improved Clustering Algorithm in Library Personalized Book Recommendation System

Xiao Lin*, Wenjuan Guan, Ying Zhang
Library, Minjiang University, Fuzhou 350108, China

***Abstract*—The information construction work of university libraries is becoming increasingly perfect. However, the massive amount of data poses significant challenges to the personalized recommendation of books. Cluster analysis has always been an important research topic in data mining technology, and it has a wide range of application fields. Clustering algorithm is a fundamental operation in big data processing, and it also has good application value in personalized recommendation of library books. To improve the personalized service quality of libraries, this study proposes a clustering algorithm based on density noise application spatial clustering. This study introduced a distance optimization strategy and Warhill algorithm to the proposed algorithm, to improve the difficulties in selecting initial parameter neighborhoods and density thresholds in traditional models, as well as computational complexity. Afterwards, this study will integrate the improved algorithm with the density peak algorithm to further improve the operational efficiency of the model. The performance verification of the model demonstrated superior clustering performance. The average accuracy of the proposed model's recommendation is 98.97%, indicating superiority. The practical application results have confirmed that there is a significant similarity between the books read by the readers and the books read by the target readers, and the effectiveness and feasibility of the proposed model have been verified. Therefore, the proposed model can contribute to the personalized recommendation function of libraries and has certain practical significance.**

Keywords*—Peak density; distance optimization; warhill algorithm; collaborative filtering; book recommendations

I. Introduction

At present, the informatization work of university libraries has made great progress, but there are still problems such as too long time spent in retrieving information and documents [1]. There are a lot of library resources in university library, and there are a lot of data, which is easy to cause college students to blindly choose books in university library or do not know which books are suitable for them to read. Therefore, more and more attention is paid to the personalized recommendation of university library. At the same time, the development of artificial intelligence has brought new opportunities to the topic. Among them, the traditional intelligent data analysis methods mainly collect and analyze users' browsing records and information, excavate and analyze the collection of documents and resources, and realize the utilization of library resources [2]. However, large amounts of data often have more complex structures and types, which makes it difficult for traditional data analysis methods to meet these needs [3]. Cluster analysis is the most commonly used method in data mining [4]. Through cluster analysis, data with high similarity can be classified into the same category, so that the difference between similar data is small, and the difference between different data is large. In addition, the existing book recommendation algorithms generally have the disadvantages of low recommendation accuracy and poor applicability in university libraries [5]. Therefore, in order to help college students more accurately find their own suitable books in college libraries, this study is based on cluster analysis in data mining. This paper proposes a Density-Based Spatial Clustering of Applications with Noise, The Clustering model of DBSCAN and Density Peak Clustering Algorithm (DPC) is proposed to optimize the personalized recommendation performance of books in university libraries. The innovations of this study lie in: (1) The distance optimization strategy was proposed, which improved DBSCAN and improved the difficulty in selecting its initial parameters. (2) Warhill was introduced to reduce the computational complexity of the model. (3) The improved DBSCAN was integrated with DPC to further improve the operational efficiency of the model. (4) The constructed model was applied in the personalized book recommendation service of library. This study consists of four parts: (1) Firstly, a review was conducted on the current development status of technologies used in the article. (2) The construction process of the model was discussed in detail. (3) The model performance and practical application effects were verified. (4) The full text was summarized and prospects were made for the future.

II. Related Works

The application of CA is quite extensive and has always been a hot topic of discussion among scholars. DPC is unable to identify cluster centers and non-center point error allocation, which can cause a chain reaction. Therefore, DingS et al. proposed an improved method. This method can reduce the density difference of non-uniformly distributed datasets and use low density points as boundaries on the foundation of cluster center and surrounding points' similarity density. The proposed model can make algorithm's clustering accuracy improved while controlling running time [6]. CuiZ and other researchers proposed an improved subspace clustering model to address the shortcomings of subspace clustering methods that cannot balance the sparsity and connectivity of coefficient matrices in high-dimensional image data. This model can preserve the coefficients between the sample and its neighbors, and prune the spatial connections within the subspace. The proposed model can effectively handle noise data in the Internet of Things and has good clustering accuracy [7]. KarimM and other scholars have conducted a detailed discussion on deep learning based clustering algorithms and pointed out different clustering quality indicators. This study utilizes the clustering methods discussed for text mining in bioimaging, cancer genomics, and biomedical texts. The final conclusion helps to provide new solutions for emerging bioinformatics problems [8]. Incomplete multi view CA methods can lead to intensive computation and complex storage. In response, researchers such as LiuX have proposed an efficient incomplete multi view method. This method utilizes consensus clustering matrix to interpolate the generated incomplete base matrix to optimize the clustering performance of the model. This study validates the performance of the proposed model in terms of clustering accuracy, evolution of consensus clustering matrix, and convergence [9]. Based on the application of CA in data mining, ZouH elaborated in detail on the basic concept, classic algorithms, and implementation process of CA through literature comparison and analysis methods. The study ultimately conducted numerical simulations, confirming the strong universality of the proposed method. It can be applied to data analysis in multiple fields and has strong theoretical value [10].

Personalized recommendation services have received increasing attention in recent years, and many scholars have contributed to this. Researchers such as ArabiH have found that contextual information such as emotions, location, and time can help improve service recommendations. Therefore, they proposed a context aware recommendation system. The system implements personalized settings based on multiple user characteristics and product functions. It utilizes users' personality traits, demographic details, geographical location, and comment emotions to generate personalized recommendations. This algorithm was ultimately proven to greatly optimize recommendation performance [11]. Scholars such as HuixiangX use K-means and utilize the relationships between users, tags, and books for group recommendations. The proposed strategy first clusters users and books, and calculates the cosine similarity between the two groups. Afterwards, it sorted and clustered the books, and tested the personalized recommendation effect. The proposed model improves the recommendation effect of books and provides better book resources for the target group [12]. SarmaD et al. constructed an effective online book recommendation system to address the irrationality of existing recommendation system rating techniques. The system uses the K-means cosine distance function to measure distance to find similarity between book clusters and grade books. The experiment used 10 datasets to validate the proposed model, indicating that the constructed recommendation system has high accuracy [13]. ZhouY has designed an information recommendation book management system based on an improved Apriori data mining algorithm, which integrates borrower and book data information. After cleaning, transforming, and integrating the relevant data, the Apriori algorithm is used to generate an association rule database. The implementation process of association matching is carried out by the personalized recommendation sub module based on the borrower and the books selected in the association rule database. The proposed model has good recommendation performance [14]. KwakW and NohY analyzed their domestic and international trends, policies, and cases based on the development background of artificial intelligence, and proposed the future direction of artificial intelligence services for libraries. This study suggests that in the future, libraries will further optimize artificial intelligence and provide personalized book recommendations to users based on their usage records [15].

To sum up, cluster analysis plays an important role in personalized recommendation. Although cluster-based algorithms have always been favored and improved by scholars, there are still some problems that need to be continuously studied and perfected by scholars. In addition, the existing recommendation services widely exist in e-commerce, short video and other platforms, and there is still a large research space for personalized recommendation of college books. In addition, the existing book recommendation algorithms also have the disadvantages of low recommendation accuracy and poor applicability in university libraries. Therefore, this study proposes a clustering algorithm combining DBSCAN and DPC, and applies it to the book recommendation system, aiming at contributing to the personalized recommendation service of libraries.

III. A Book Recommendation Model Based on Improved DBSCAN and DPC

Traditional DBSCAN has many limitations. Therefore, this section first optimizes and improves it. Then it is integrated with DPC to better achieve the clustering performance and book recommendation effect.

A. Construction and Improvement Strategy of DBSCAN

DBSCAN is the most classic density based CA. DBSCAN adopts the concept of neighborhood and utilizes the spatial distribution characteristics of point sets to cluster the dataset [16]. The basic idea is to determine the number of neighboring data points under a certain threshold, centered around a single data point. Using sparse points as the boundary, low density data points are used as the classification boundary, and high density points are used as another class [17]. Compared with other clustering algorithms, DBSCAN can find clusters of different sizes and shapes under conditions of noise interference. The full name of DBSCAN is "Density-Based Spatial Clustering of Applications with Noise", which is characterized by its ability to effectively process noisy data. Compared with conventional CA, DBSCAN not only handles non convex data well, but also fits convex datasets well. DBSCAN includes elements such as neighborhood, density threshold, core points, boundary points, and outliers. It is assumed that the sample dataset is $S=\left\{x_1,x_2,\cdots,x_n\right\}$, $x_i \in S$, and the neighborhood is δ. All sample points in the neighborhood form a subsample set and meet the conditions shown in Eq. (1).

$$N_\delta\left(x_j\right)=\left\{x_j \in S \middle| dis\tan ce\left(x_i,x_j\right)\leq\delta\right\} \tag{1}$$

Eq. (1) represents the conditions that need to be met for the sub sample set in the hypersphere region with a radius of δ. The density threshold in DBSCAN can be manually set. According to the relevant elements of DBSCAN, there are three relationships between data points in the overall algorithm: density direction, density reachability, and density connection [18-19]. Among them, the meaning of density direction is: for two samples that exist in the domain, if one is the core object, the other sample is called density direct access by that sample. The meaning of density reachability is: for x_i and x_j, if there is a sample sequence $\left\{p_t \middle| t=1,2,\ldots,T\right\}$ that satisfies $p_1 = x_i, p_T = x_j$, and p_{t+1} is directly reachable by p_t density, then x_j is said to be reachable by x_i density. At this point, the samples in the sequence are all core points, which means that the density can satisfy transitivity but not symmetry. The meaning of density connection is: for x_i and x_j, if there is a core point x_k, so that both x_i and x_j can be reached by x_k density, then x_i and x_j are called density connections. Fig. 1 shows the schematic diagram of DBSCAN. x_1 and x_2 are the core points. x_3 and x_4 are boundary points. x_2 is directly accessible through x_1 density. x_3 is directly reached by x_2 density. x_4 can be achieved by x_1 density. x_4 and x_3 are density connected.

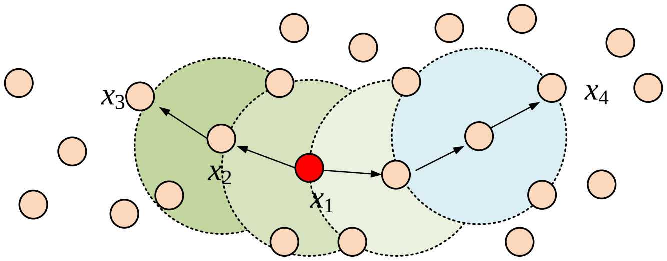


Fig. 1. DBSCAN clustering process.

Compared with other clustering methods, DBSCAN has the following advantages. Firstly, it can handle dense data of any shape. Due to its ability to find and remove noise throughout the entire search process, the clustering process is not affected by sample set noise. Its clustering does not require specifying the number of clusters in advance, and the clustering results are unbiased. However, the classic DBSCAN also has significant drawbacks that cannot be ignored. In DBSCAN, it is difficult to choose the initial parameter neighborhood and density threshold. DBSCAN is not suitable for samples with uneven distribution and large distances, in which case the clustering results are not ideal. For high-dimensional sample sets, the convergence speed of clustering algorithms is slow and cannot achieve accurate clustering results [20-21]. Therefore, this study first proposes optimization for the parameter selection process of DBSCAN, namely a DBSCAN model based on distance optimization (D-DBSCAN). D-DBSCAN can automatically select neighborhood values based on the characteristics of initial density threshold and data distribution density. Assuming the sample set is $S=\left\{x_1, x_2, \cdots, x_n\right\}$, there is Eq. (2).

$$N\left(x_i\right)=\left\{x_j \in S \middle| 0<d\left(x_i, x_j\right)<\delta\right\} \tag{2}$$

Eq. (2) indicates that for $x_i \in S$, the density of x_i is the number of data points owned within the domain δ. Eq. (3) is the distance coefficient.

$$\theta = N\left(x_j\right) / N\left(x_i\right) \tag{3}$$

In Eq. (3), θ is the distance coefficient, which specifically means that for the sample point x_j within the neighborhood of core point x_i, it is called the distance coefficient of x_j to x_i. Eq. (4) is the distance matrix between samples.

$$D=\left\{D_{ij} \middle| i, j \in R,\ i \neq j\right\} \tag{4}$$

In Eq. (4), D_{ij} represents the distance between samples x_i and x_j. The average distance of the points with the closest density threshold to x_i was calculated and added to the distance set U to calculate the overall average $\overline{U}$ of U. The neighborhood radius was set as the average distance $\overline{U}$, and all core points were identified and added to the core object set Ω. Subsequently, a core point x_i was randomly selected as the clustering center to form a new cluster C_i. All sample points in the δ neighborhood near x_i were identified, and the neighborhood radius δ was adjusted by calculating distance coefficient θ between the sample and core points. By repeating the above operation to find all sample points with achievable density and adding them to C_i, the final result was obtained in Eq. (5).

$$C=\left\{C_1, C_2, \cdots, C_n\right\} \tag{5}$$

Thus, D-DBSCAN can effectively improve the selection of initial parameters. To further reduce the complexity of D-DBSCAN calculation, Warhill was introduced to construct W-D-DBSCAN. Warhill can calculate reachable matrices to reduce the complexity of the model. It is assumed that Eq. (6) is a directed graph.

$$G=\left\langle V, E\right\rangle \tag{6}$$

In Eq. (6), V represents the node set. E refers to an adjacent points set. The node set is $V=\left\langle v_1, v_2, \cdots, v_n\right\rangle$. The matrix is $A=\left(a_{ij}\right)_{m\times n}$.

$$a_{ij}=\begin{cases}1, & v_i \ adjoin \ v_j \\ 0, & v_i \ is\ not\ adjacent\ to\ v_j\end{cases} \tag{7}$$

A directed graph can directly reflect two elements' connection. An adjacency matrix refers to nodes' connection in the directed graph. The two nodes that can be directly reached are regarded as 1. Otherwise, they are regarded as 0. It is assumed that $G = \langle V, E \rangle$ is a simple directed graph. The node set is $V = \langle v_1, v_2, \cdots, v_n \rangle$. The matrix is $F = \left(f_{ij} \right)_{n \times n}$.

$$f_{ij} = \begin{cases} 1, \text{There is a non-zero directed path from } v_i \text{ to } v_j \\ 0, \ others \end{cases} \tag{8}$$

If there is a reachable path between two elements, they are connected and reachable, marked as 1, otherwise marked as 0. In a directed graph, the connectivity between nodes can be directly reflected by a line with an arrow. However, it is difficult to determine the connectivity between two independent nodes, and a matrix of direct connectivity must be used. Usually, Warshall is used to convert adjacency matrices into reachable matrices. $dis[i, j]$ stands for the distance. A matrix $A_{n \times n}$ was established for dataset D and $dis[i, j]$ between data objects i, j was calculated in Eq. (9).

$$A[i, j] = \begin{cases} 1, dis[i, j] \le Eps \\ 0, dis[i, j] > Eps \end{cases} \tag{9}$$

In Eq. (9), Eps represents the initial threshold. The adjacency matrix's reachable matrix obtained by Warhill is regarded as transitive closure or density connected sets which are the maximum. Fig. 2 shows the flowchart of W-D-DBSCAN. The density connected set obtained by Warhill is to achieve clustering. Compared with traditional density clustering, its clustering process is simpler.

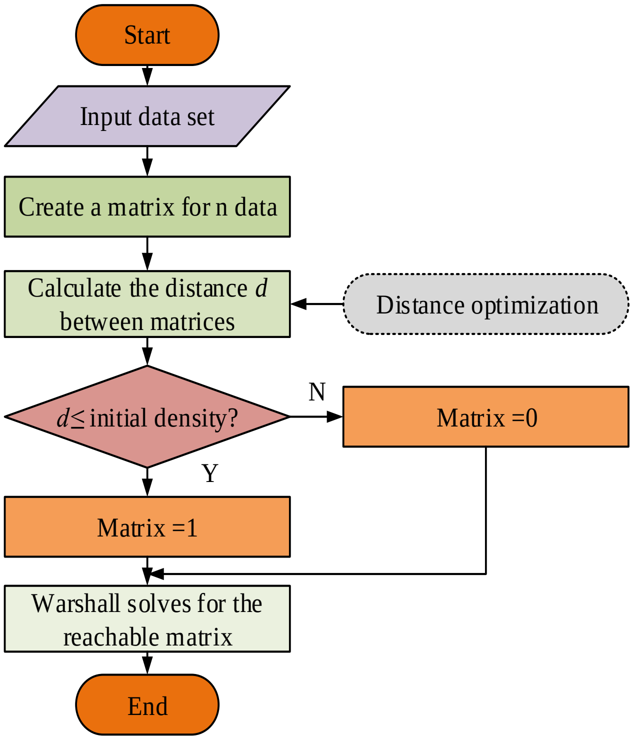

Start
Input data set
Create a matrix for n data
Calculate the distance d between matrices
Distance optimization
d≤initial density?
N
Matrix =0
Y
Matrix =1
Warshall solves for the reachable matrix
End

Fig. 2. Flow of the W-D-DBSCAN algorithm.

B. A Book Recommendation Model Integrating D-W-DBSCAN and DPC

DPC is a data clustering method based on maximum connectivity, which not only effectively distinguishes noise, but also has strong robustness for various convex and non-convex data [22-23]. DPC assumes that the centroid of data targets set is surrounded by a low-density data, and this low-density data is assigned to the nearest, less dense centroid data. DPC improves the efficiency of the algorithm and has good clustering performance by manually drawing decision maps and selecting data objects with high relative distance and density for clustering. On this basis, a new method of clustering quality centers based on DPC is proposed. This study adopts a combination of W-D-DBSCAN and DPC, namely the W-D-DBSCAN DPC model, to solve the problem that the centroid selected by DPC on the decision graph cannot be applicable to all data. This method first uses the maximum region density value as the centroid. Secondly, W-D-DBSCAN was used for clustering. Then, samples with high local density were searched out from the remaining samples, and the samples were clustered using W-D-DBSCAN as the center. The above methods were used to search for centroids and cluster them until all data were classified or local noise appeared, indicating the completion of clustering. This method can effectively solve the manual intervention problem in traditional DPC methods, while also reducing the computational complexity of W-D-DBSCAN.

Assuming any data object is i, the local density was calculated for it in Eq. (10).

$$\rho_i = \sum_j \chi\left(d_{ij} - d_c\right) \tag{10}$$

In Eq. (10), ρ_i represents local density. d_c represents truncation distance. d_{ij} represents relative distance. The relative distance between other data and data i was calculated using Euclidean distance in Eq. (11).

$$\sigma_i = \min_{J:\rho_j > \rho_i} \left(d_{ij} - d_c\right) \tag{11}$$

In Eq. (11), d_{ij} represents i and j's Euclidean distance. Through Eq. (11), the relative distance of data i refers to: if i is the largest data object of ρ_i, then σ_i represents the $\max_j \left(d_{ij}\right)$ between other data and data i. If i has no the highest local density, its relative distance is in the data with lower local density than the data, and it is closest to i. Fig. 3 shows the clustering process of W-D-DBSCAN-DPC [24].

In Fig. 3, the comprehensive model first calculates the local density. Then, they are compared to get the maximum local density, with data 7 of this maximum local density as the centroid. Starting from 7, W-D-DBSCAN was used to divide 1, 2, 3, 4, 5, and 6 into centroid data, completing the classification of the first type of cluster. Then, the local density of the remaining dataset was continued to be calculated. And the maximum local density was obtained through comparison again. Using the highest local density value of 10 as the centroid, W-D-DBSCAN was used to cluster samples 8, 9, 11, and 12, and data 8, 9, 11, and 12 were divided into centroids 10. This process continues until the remaining data does not belong to a category, marked as noise, and clustering ends [25].

This study applies W-D-DBSCAN-DPC to the clustering process of library readers. By categorizing different categories of readers, the first category in the database is the book that the reader is most interested in. Therefore, recommending the first category of books to readers is a desirable approach. The recommendation algorithm of W-D-DBSCAN-DPC effectively solves the traditional "cold start". This method will help improve the performance of recommendation systems and alleviate the pressure of big data processing. Fig. 4 shows the system diagram of the library book recommendation system. W-D-DBSCAN-DPC is used to classify readers with high similarity. By compressing the existing massive reader data into analyzing and recommending the same type of reader data, the recommendation efficiency was improved. The collaborative filtering algorithm is used to generate the Top-n nearest neighbor set of readers, thereby completing recommendations.

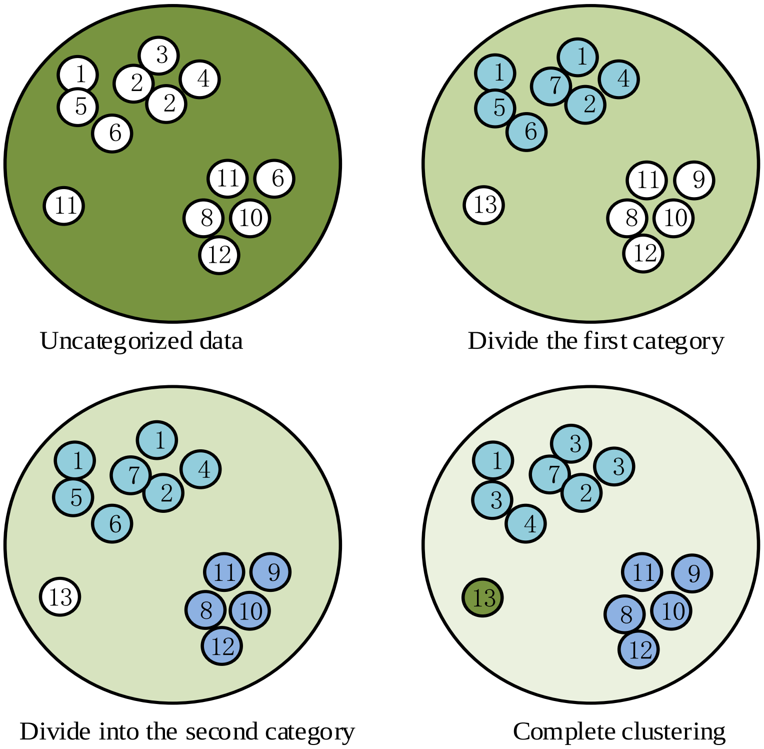


Fig. 3. Clustering process of W-D-DBSCAN-DPC.

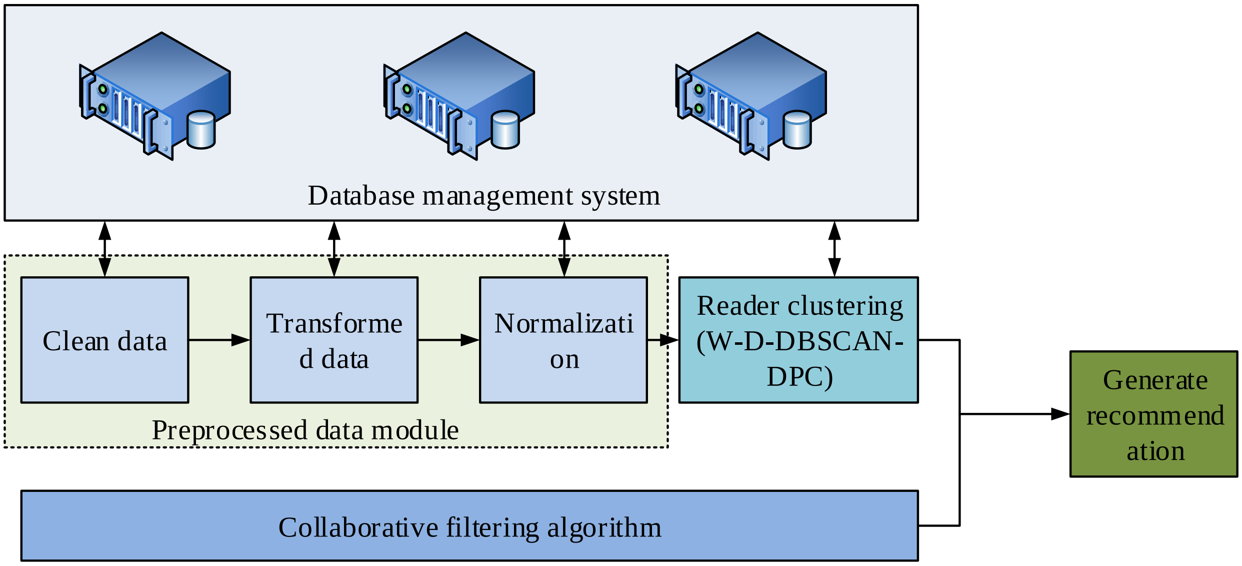

Database management system
Clean data
Transforme d data
Normalizati on
Preprocessed data module
Reader clustering (W-D-DBSCAN-DPC)
Generate recommend ation
Collaborative filtering algorithm

Fig. 4. Library book recommendation system.

This study used W-D-DBSCAN-DPC to cluster readers, identify the most popular books, and recommend them to new readers to solve the "cold start" problem. Adopting this method for users with a reading history reduces the range of data that the recommendation algorithm needs to process and reduces the problems that users may encounter during the reading process. If the target reader has a historical borrowing record, then based on its clustering results, a collaborative filtering algorithm is used to obtain the Top-n neighbor set for other readers of the same type and recommend them. The similarity between the target reader and the reader is calculated by Eq. (12).

$$sim(i,j)=\frac{\sum_{u\in U_{ij}}\left(R_{u,j}-\overline{R_i}\right)\left(R_{u,j}-\overline{R_j}\right)}{\sqrt{\sum_{u\in U_{ij}}\left(R_{u,j}-\overline{R_i}\right)}\sqrt{\sum_{u\in U_{ij}}\left(R_{u,j}-\overline{R_j}\right)}} \tag{12}$$

In Eq. (12), $sim(i,j)$ represents the similarity between readers i,j. U_{ij} represents the book categories that readers i,j are both interested in. $\overline{R_i}$, $\overline{R_j}$ represent the average values of readers i,j's interest in all books. By using Eq. (12), the similarity between the target reader and the reader can be obtained, and the similarity sequence can be obtained by sorting their similarity. Then, the Top-n neighbor set is generated from the high similarity readers. Therefore, the book recommendation model based on W-D-DBSCAN-DPC has been constructed.

To evaluate the clustering effect, Accuracy (ACC), Purity, and contour coefficient were introduced as evaluation indicators in this study. Eq. (13) is the calculation of ACC.

$$ACC=\left(\sum_{i=1}^{n}\delta\left(\hat{C}_i map\left(C_i\right)\right)\right)/n \tag{13}$$

In Eq. (13), C_i represents the category label of the proposed algorithm. $\hat{C}_i$ represents the true label of data object. $map(x)$ represents a mapping function. A high ACC value indicates high clustering quality. Eq. (14) represents the calculation of Purity.

$$purity=\frac{1}{N}\sum_{i=1}^{k}x_i \tag{14}$$

In Eq. (14), N represents the number of datasets. k represents the number of clusters in the dataset. x_i represents the number of correctly clustered data objects. Purity is within 0-1, which is closer to 1, the data clustering accuracy is higher. Eq. (15) is the calculation of contour coefficient.

$$S(X)=\frac{b(x)-a(x)}{\max\left(a(x),b(x)\right)} \tag{15}$$

In Eq. (15), $a(x)$ represents the average distance of other samples within the same cluster of sample x. $b(x)$ represents the average distance between sample x and all sample points within the nearest cluster. The range of contour coefficient values is $[-1,1]$, which is closer to 1, the clustering effect is better.

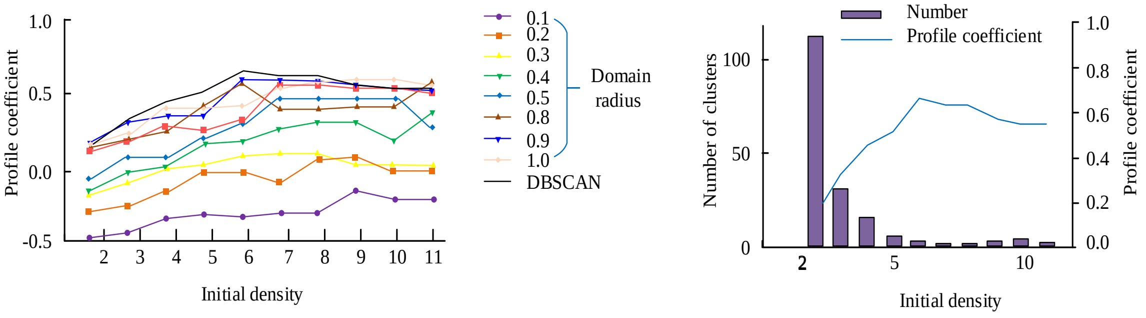
IV. Performance Verification of Book Recommendation Model Application Based on W-D-DBSCAN-DPC

This study first analyzes the clustering performance of W-D-DBSCAN-DPC. For this purpose, sufficient datasets were selected for the study and detailed discussions were conducted. In addition, the actual application effect of book recommendation was verified using a certain university as an example.

A. Performance Verification of W-D-DBSCAN-DPC Model

This study first verifies the clustering performance of the W-D-DBSCAN-DPC model. This study selected three datasets, namely Spiral, Lineblobs, and Aggregation, for validation. Spiral is a set of non-convex spiral datasets. Lineblobs is a set of smiling face datasets. Aggregation includes both spherical and non-spherical datasets. The proposed W-D-DBSCAN-DPC is implemented in C language and visualized using MATLAB. The proposed algorithm was evaluated and analyzed by comparing clustering results at different scales. To evaluate the clustering effect and determine the optimal number of clusters, this study used traditional DBSCAN and D-DBSCAN to cluster the dataset, and obtained the contour coefficients corresponding to different number of clusters in Fig. 5.

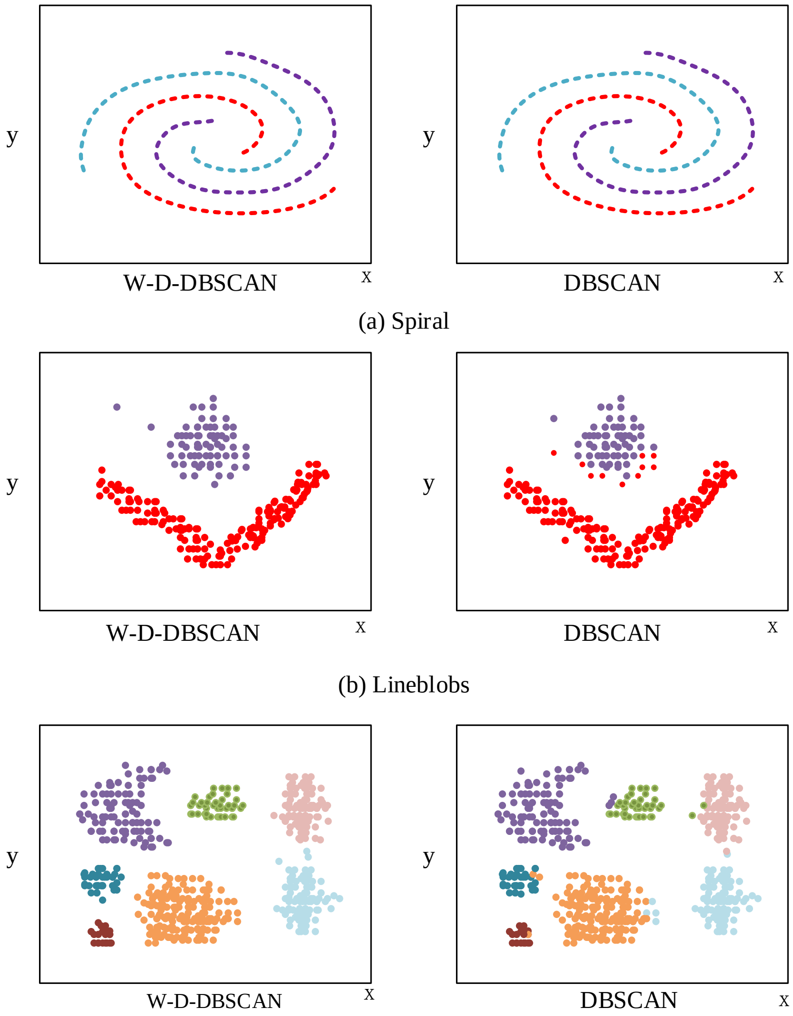
In Fig. 5 (a), DBSCAN requires continuous search of two parameters, namely initial value and neighborhood radius, in order to find the optimal solution. D-DBSCAN only needs to find the optimal solution under different initial density conditions, without adjusting the neighborhood radius. Fig. 5 (b) shows the number of class clusters and contour coefficients corresponding to different initial densities. At an initial density of 6, the D-DBSCAN contour coefficient reached its optimal value of 0.668. The contour coefficient of the DBSCAN method is 0.612. Compared with DBSCAN, the clustering performance of D-DBSCAN has improved by 9.17%. These results verify the effectiveness of distance optimization.



(a) DBSCAN profile coefficients at different initial densities and domain parameters

(b) D-DBSCAN Specifies the value of d-dbscan at different initial densities

Fig. 5. D-DBSCAN performance verification results.



(a) Spiral

(b) Lineblobs

(c) Aggregation

Fig. 6. Clustering results of D-DBSCAN and W-D-DBSCAN.

In Fig. 6, they are the clustering results of three datasets using two methods, D-DBSCAN and W-D-DBSCAN. In the figure, two algorithms' clustering effects on three types of datasets are relatively similar. Because W-D-DBSCAN continues the advantages of D-DBSCAN and can achieve clustering on any dataset.

Fig. 7 shows the comparison results of the runtime between D-DBSCAN and W-D-DBSCAN algorithms on a spiral dataset. The spiral dataset was generated into datasets with different shapes, sizes, and densities. And experiments were conducted on these datasets using D-DBSCAN and W-D-DBSCAN. In Fig. 7, W-D-DBSCAN has a shorter runtime. As a result, W-D-DBSCAN effectively reduces the complexity of the model and improves its running speed.

To highlight the excellent performance of W-D-DBSCAN-DPC, this study selected four clustering algorithms: DPC, FCM, and K-means for comparative analysis. Fig. 8 shows the clustering performance of four algorithms on three datasets. In Fig. 8(a), K-means cannot cluster non-convex datasets, while the spiral shape of Spiral dataset is non-convex, resulting in clustering errors. In Fig. 8(b), the clustering results of W-D-DBSCAN-DPC and DPC are basically correct because they consider the transfer relationship between data. In Fig. 8(c), the clustering performance of W-D-DBSCAN-DPC is superior to other three algorithms.

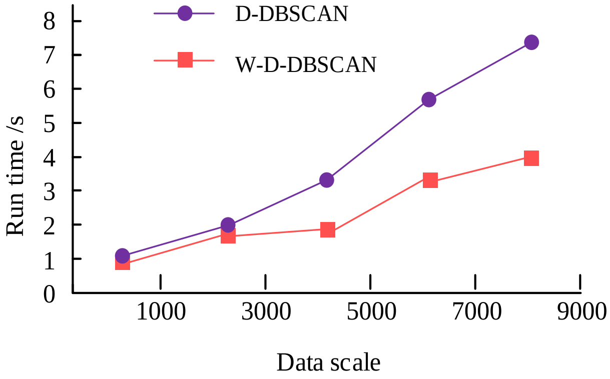
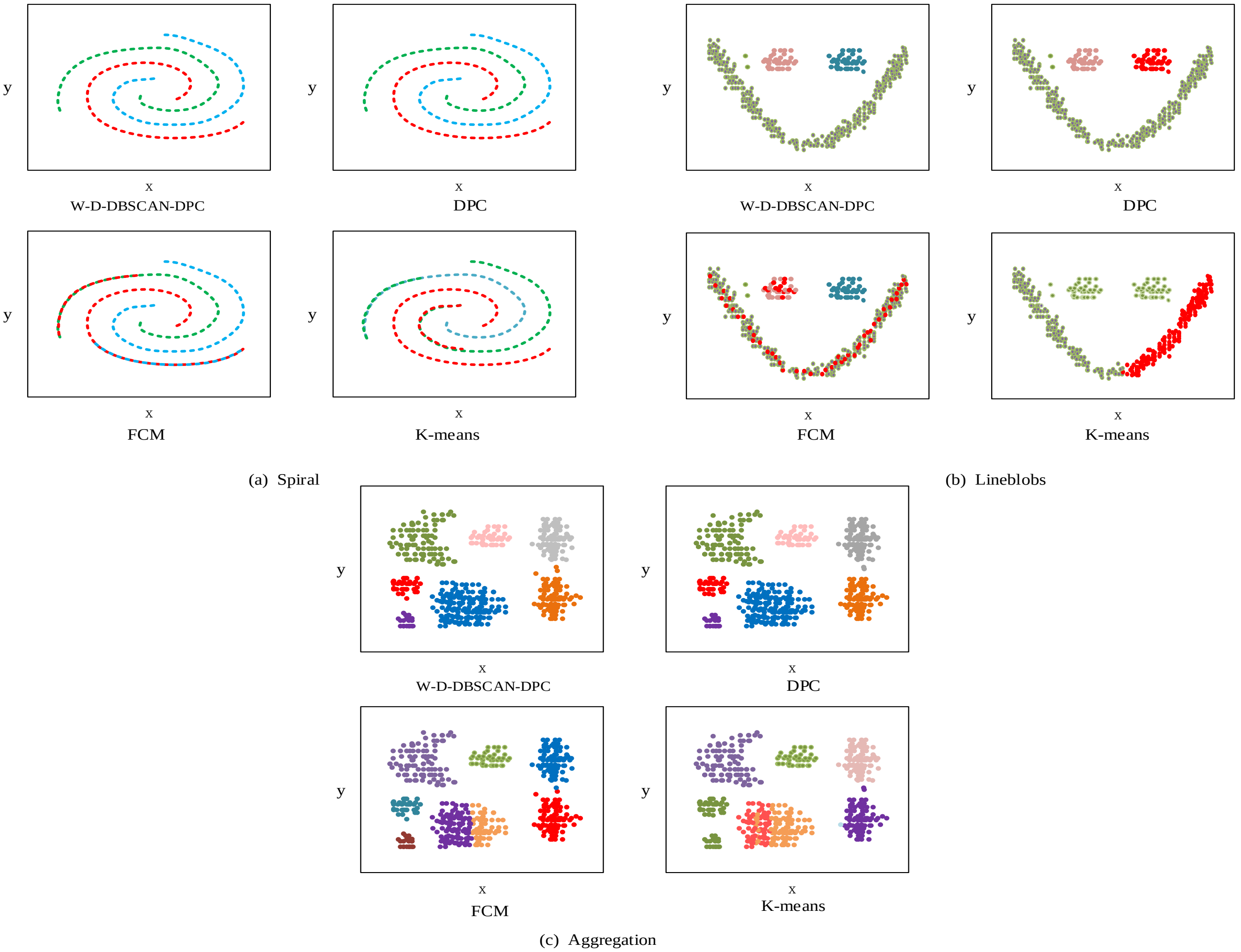


Fig. 7. Comparison of running time of the two algorithms on spiral data set.



(a) Spiral

(b) Lineblobs

(c) Aggregation

Fig. 8. Clustering effect of four algorithms on three data sets.

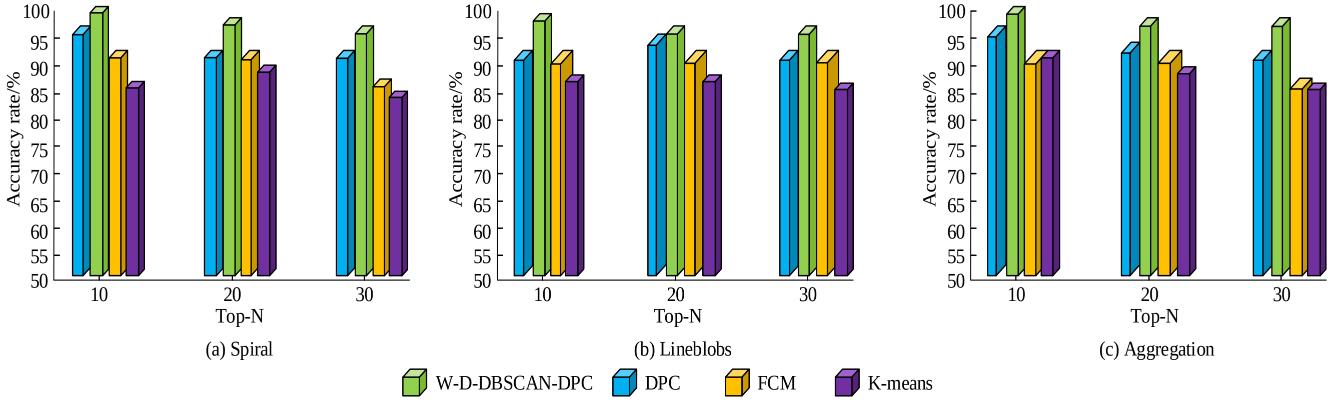
TABLE I. PURITY VALUES OF THE FOUR ALGORITHMS

Data set	W-D-DBSCAN-DPC	DPC	FCM	K-means
Iris	0.94	0.91	0.89	0.79
Tae	0.66	0.64	0.55	0.56
Cmc	0.78	0.73	0.63	0.56
Seeds	0.92	0.89	0.88	0.78

To further test the performance of W-D-DBSCAN-DPC, quantitative analysis was conducted on the aforementioned artificial dataset and UCI dataset. Four sets of data were randomly selected from UCI database for experiments on DPC, FCM, and K-means. Table I is four algorithms' purity values. The purity values on four datasets, W-D-DBSCAN-DPC, are the highest, indicating better clustering performance. Its clustering performance in Tae and Cmc datasets is relatively poor because one of these datasets has a large feature value, which affects the clustering results.

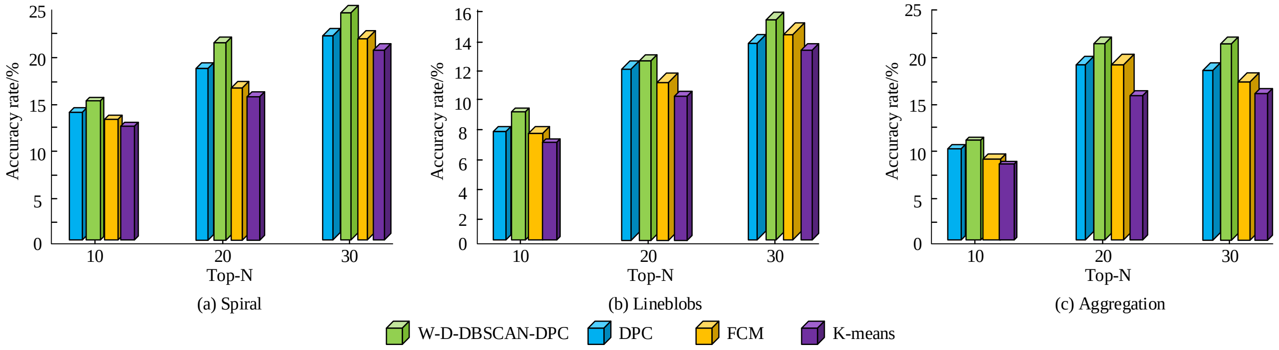
To demonstrate the effectiveness of the proposed W-D-DBSCAN-DPC recommendation algorithm in recommendation, the above three algorithms are used as comparative recommendation algorithms for this experiment. Fig. 9 shows the recommendation accuracy obtained by four algorithms on three types of datasets. The average accuracy of W-D-DBSCAN-DPC is 98.97%, while DPC, FCM, and K-means are 95.67%, 93.23%, and 90.34%, respectively. Thus, the superiority of W-D-DBSCAN-DPC was verified.

Fig. 10 shows the comparison results of recall rates obtained by four algorithms on three types of datasets. The average performance of W-D-DBSCAN-DPC is also superior to other algorithms, further verifying its superior performance.



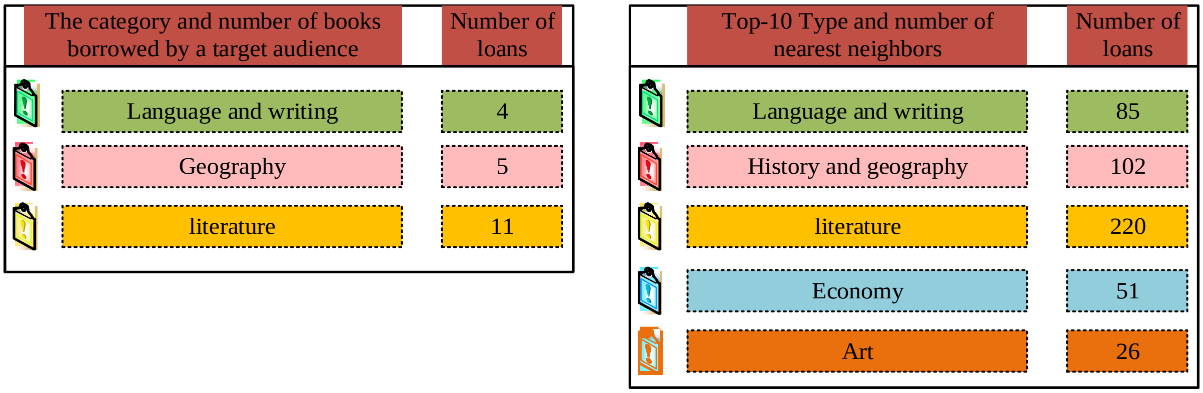
(a) Spiral (b) Lineblobs (c) Aggregation

Fig. 9. Recommendation accuracy rates of four algorithms on three types of data sets.



(a) Spiral (b) Lineblobs (c) Aggregation

Fig. 10. Comparison results of recall rates of four algorithms on three types of data sets.



The category and number of books borrowed by a target audience	Number of loans
Language and writing	4
Geography	5
literature	11

Top-10 Type and number of nearest neighbors	Number of loans
Language and writing	85
History and geography	102
literature	220
Economy	51
Art	26

(a) The category and number of books borrowed by a target audience

(b) Top-10 Type and number of nearest neighbors

Fig. 11. W-D-DBSCAN-DPC recommended result.

B. Practical Application Effect of Book Recommendation Model Based on W-D-DBSCAN-DPC

This study divides the existing book materials in the library into 22 categories and uses them as reader feature vectors. The experiment extracted borrowing information from the data center of a certain university library for 21st year college students, of which 8011 were borrowing information from the university library. By processing these materials, some reader's interest and preference data can be obtained.

This experiment proved that W-D-DBSCAN-DPC was used for clustering, combined with recommendation algorithms, and finally the Top-n algorithm was used for classification. Fig. 11(a) is a summary of the types of books borrowed and the number of target readers. According to the distribution, when recommending four books to the target audience, two literary books, one local science book, and one language and text book can be recommended. Fig. 11(b) shows the borrowing classification and quantity of readers in the Top-10 neighbor set. These results confirm that among the top 10 neighboring readers of the target audience, in addition to books related to language, geography, literature, etc., there are also books related to history, economy, art, etc. Therefore, according to Fig. 11(b), books on history, economics, art, etc. can be appropriately recommended to target readers. In addition, there is a significant similarity in the types of books read by Top-10 readers and the books read by the target readers, thus verifying the effectiveness and feasibility of this algorithm.

V. Discussion

The problem of "blind selection" or "unable to find suitable books" is common among contemporary college students in the library. Therefore, this topic tries to combine the clustering algorithm with the recommendation system organically, and build a system suitable for university book recommendation. Firstly, the data are collected and classified. On this basis, the user's historical reading records are clustered and classified, and finally the user's Top-n nearest neighbor set is calculated, and books are recommended to it, so as to help students find books suitable for themselves. Density clustering has always been a hot topic in the field of data mining, and the density peak clustering algorithm has pushed the study of density method to a hot trend [26-28]. Starting from cluster analysis and recommendation system, this study introduced the theoretical analysis, research status and common methods of related methods in detail, so as to make sufficient preparation for the follow-up work. This study mainly discusses the density peak and density clustering algorithm. It is found that in the density clustering algorithm, data objects are grouped by density linkage, but the calculation process of density linkage is complicated, and it needs to determine the core point, density reachability, direct density reachability, etc. Density peaks the method of selecting the center of mass on a decision graph is not suitable for all data. The main contents of this research are as follows: (1) Propose distance optimization strategies to improve DBSCAN and improve the difficulty in selecting its initial parameters. (2) Warshall algorithm was introduced to reduce the computational complexity of the model. (3) Merge the improved DBSCAN with DPC to further improve the operating efficiency of the model. (4) Apply the model to the personalized book recommendation service of the library.

Experimental results show that, compared with the traditional DBSCAN algorithm, W-D-DBSCAN-DPC algorithm can find clusters with different shapes and adaptively select appropriate neighborhood radius, which is more suitable for complex student book preferences [29-30]. Finally, according to the results of student book recommendation, different student groups show strong differences in daily borrowing activities. It can be concluded that the reader feature vector based on the borrowing information of college students in the library has a strong correlation with the reading preference of students, which provides more reference and research ideas for college library managers.

VI. Conclusion

The increasing amount of information data in university libraries has brought a lot of inconvenience to the daily lives of teachers and students. This study proposes a W-D-DBSCAN-DPC to enhance the personalized service quality of libraries. The performance of the model was verified: the optimal contour coefficient for D-DBSCAN was 0.668, while for DBSCAN, it was 0.612. Compared with DBSCAN, D-DBSCAN clustering performance improved by 9.17%, verifying the effectiveness of distance optimization. W-D-DBSCAN has a shorter runtime compared to D-DBSCAN, verifying the effectiveness of Warhill. The comparative analysis of four clustering algorithms, W-D-DBSCAN-DPC, DPC, FCM, and K-means, shows that W-D-DBSCAN-DPC's clustering results are basically correct, and its clustering effect is better than other three algorithms. W-D-DBSCAN-DPC has the highest purity values on the four datasets, indicating better clustering performance. The average accuracy recommended by W-D-DBSCAN-DPC is 98.97%, while DPC, FCM, and K-means are 95.67%, 93.23%, and 90.34%, respectively, confirming the superiority of W-D-DBSCAN-DPC. The practical application has confirmed that there is a significant similarity between the books read by Top-10 readers and the books read by the target readers, verifying the effectiveness and feasibility of this algorithm. The drawback is that the threshold in Warhill has not been controlled, resulting in some interference that can be optimized in the future.

Acknowledgment

The research is supported by Fund Project of Steering Committee for Academic Libraries of Fujian Province: Research on Innovative Service of Augmented Reality and Mixed Reality Technology in University Libraries (Project number: FJTGW202228); Major topic of the research project on education and teaching reform of undergraduate colleges and universities of the Education Department of Fujian Province: the construction of smart platform for reading promotion activities of University Library Alliance in Fuzhou University Town (Project number: FBJG 20210319).

REFERENCES

[1] Ez-Zahout A, Gueddah H, Nasry A, Madani R, Omary F. A hybrid big data movies recommendation model based knearest neighbors and matrix factorization. Indonesian Journal of Electrical Engineering and Computer Science, 2022, 26(1): 434-441.

[2] Gupta M, Kumar P. Recommendation generation using personalized weight of meta-paths in heterogeneous information networks. European Journal of Operational Research, 2020, 284(2): 660-674.

[3] Hobbs R. Propaganda in an age of algorithmic personalization: Expanding literacy research and practice. Reading Research Quarterly, 2020, 55(3): 521-533.

[4] Blin K, Shaw S, Kloosterman A, Charlop-Powers Z, Wezel G, Medema M, Weber T. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic acids research, 2021, 49(W1): W29-W35.

[5] Soffer O. Algorithmic personalization and the two-step flow of communication. Communication Theory, 2021, 31(3): 297-315.

[6] Ding S, Du W, Li C, Xu X, Wang L, Ding L. Density peaks clustering algorithm based on improved similarity and allocation strategy. International journal of machine learning and cybernetics, 2023, 14(4):1527-1542.

[7] Cui Z, Jing X, Zhao P, Zhang W, Chen J. A new subspace clustering strategy for AI-based data analysis in IoT system. IEEE Internet of Things Journal, 2021, 8(16): 12540-12549.

[8] Karim M, Beyan O, Zappa A, Costa I, Rebholz-Schuhmann D, Cochez M, Decker S. Deep learning-based clustering approaches for bioinformatics. Briefings in bioinformatics, 2021, 22(1): 393-415.

[9] Liu X, Li M, Tang C, Xia J, Xiong J, Liu L, Kloft M, Zhu E. Efficient and effective regularized incomplete multi-view clustering. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(8): 2634-2646.

[10] Zou H. Clustering algorithm and its application in data mining. Wireless Personal Communications, 2020, 110(1): 21-30.

[11] Arabi H, Balakrishnan V, Shuib N. A Context-Aware Personalized Hybrid Book Recommender System.Journal of Web Engineering (JWE), 2020, 19(3/4):405-428.

[12] Huixiang X, Xiaomin L, Yueyan L. Group Recommendation Based on Attribute Mining of Book Reviews. Data analysis and knowledge discovery, 2020, 4(2/3): 214-222.

[13] Sarma D, Mittra T, Hossain M S. Personalized book recommendation system using machine learning algorithm. International Journal of Advanced Computer Science and Applications, 2021, 12(1):2121-219.

[14] Zhou Y. Design and Implementation of Book Recommendation Management System Based on Improved Apriori Algorithm.Intelligent Information Management, 2020, 12(3):75-87.

[15] Kwak W, Noh Y. A study on the current state of the library’s AI service and the service provision plan. Journal of Korean Library and Information Science Society, 2021, 52(1): 155-178.

[16] Bindhu V, Ranganathan G. Hyperspectral image processing in internet of things model using clustering algorithm. Journal of ISMAC, 2021, 3(2): 163-175.

[17] Hu L, Zhang J, Pan X, Luo X, Yuan H. An effective link-based clustering algorithm for detecting overlapping protein complexes in protein-protein interaction networks. IEEE Transactions on Network Science and Engineering, 2021, 8(4): 3275-3289.

[18] Oyewole G J, Thopil G A. Data clustering: Application and trends. Artificial Intelligence Review, 2023, 56(7): 6439-6475.

[19] Liu S, Jin S. 3-D gravity anomaly inversion based on improved guided fuzzy C-means clustering algorithm. Pure and Applied Geophysics, 2020, 177(2): 1005-1027.

[20] Chen D. Automatic vehicle license plate detection using K-means clustering algorithm and CNN. Journal of Electrical Engineering and Automation, 2021, 3(1): 15-23.

[21] Li P, Xie H. Two-stage clustering algorithm based on evolution and propagation patterns. Applied Intelligence, 2022, 52(10): 11555-11568.

[22] Nitu P, Coelho J, Madiraju P. Improvising personalized travel recommendation system with recency effects. Big Data Mining and Analytics, 2021, 4(3): 139-154.

[23] Zhao G, Liu Z, Chao Y, Qian X. CAPER: Context-aware personalized emoji recommendation. IEEE Transactions on Knowledge and Data Engineering, 2020, 33(9): 3160-3172.

[24] Guo Y, Mustafaoglu Z, Koundal D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. Journal of Computational and Cognitive Engineering, 2023, 2(1): 5-9.

[25] Sarkar A, Biswas A, Kundu M. Development of q-Rung Orthopair Trapezoidal Fuzzy Einstein Aggregation Operators and Their Application in MCGDM Problems. Journal of Computational and Cognitive Engineering, 2022, 1(3): 109-121.

[26] Wang M, Zhang Y Y, Min F, Deng L, Gao L. A two-stage density clustering algorithm[J]. Soft Computing, 2020, 24(23): 17797-17819.

[27] Sibille L, Civera M, Zanotti Fragonara L, Ceravolo R. Automated Operational Modal Analysis of a Helicopter Blade with a Density-Based Cluster Algorithm[J]. AIAA Journal, 2023, 61(3): 1411-1427.

[28] Hassan B A, Rashid T A, Mirjalili S. Formal context reduction in deriving concept hierarchies from corpora using adaptive evolutionary clustering algorithm star[J]. Complex & Intelligent Systems, 2021, 7(5): 2383-2398.

[29] Anand S K, Kumar S. Experimental comparisons of clustering approaches for data representation[J]. ACM Computing Surveys (CSUR), 2022, 55(3): 1-33.

[30] Zubaroğlu A, Atalay V. Data stream clustering: a review[J]. Artificial Intelligence Review, 2021, 54(2): 1201-1236.