Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Construction of Personalized Service System of Digital Resource Library Based on Artificial Intelligence

Zhenyi An[1, ✉]

[1] *Library of Guizhou Minzu University, Guiyang, Guizhou, 550025, China*

ABSTRACT

Personalized service is a targeted initiative for digital resource libraries to improve the quality of service and better play the function of culture and education. This paper proposes a digital book personalized recommendation algorithm based on artificial intelligence technology. After acquiring the borrowing data and pre-processing, the reader's portrait is visualized with factor analysis and cluster analysis methods respectively. The traditional Slopeone algorithm is weighted and the collaborative filtering algorithm is improved. Combine the user profile with collaborative filtering to realize the personalized recommendation of digital books. User similarity calculates four types of readers such as pragmatic, youthful, recreational and curious. This paper's algorithm outperforms CFRA and RABC algorithms under each parameter, with the highest recommendation accuracy and novelty, and realizes personalized library services.

Keywords: artificial intelligence, Slopeone, cosine similarity, Jaccard coefficient, personalized recommendation

1. Introduction

The focus of library work should be to enable users to access knowledge quickly, accurately and conveniently in the best possible environment. China's digital resource libraries are being built at a fast pace, and based on the growing information technology, many libraries have set up digital resource libraries, and are constantly expanding their own digital resources [1-4]. However, in the process of digital resource utilization readers have encountered certain difficulties in finding the digital book resources they want quickly and accurately. This is first of all due to the rapid increase in the library's digital book resources, the number of book information is constantly increasing, which also makes the user to find a piece of information need to use more and more time [5-8]. At the same time the communication between the website system and the user is more difficult to user is difficult to precise their needs. In addition, the current library users' needs gradually to personalized, diversified

✉ Corresponding author.
E-mail address: 18285021109@126.com (Z. An).
Received 20 March 2024; Revised 05 June 2024; Accepted 20 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-529

direction of the development of the same keywords entered into the different users may also have different information needs [9-12]. Therefore, the core concept of library culture construction is all people-oriented, and digital resources library personalized service is very necessary. Under the technical support of artificial intelligence, through personalized service, users can find the resources they want more accurately and quickly to improve the utilization of library digital resources [13-16]. This also requires the library must change the service concept, the library only carry out personalized service, the user-centered service concept is determined, the digital book resources for effective integration to achieve the personalization of book resources. According to the user's information use behavior, characteristics, preferences and habits to provide personalized services for users [17-20].

In order to construct the personalized service system of digital resource library, the borrowing data of digital library readers and users are firstly obtained, and the data are pre-processed. The one-hot coding method is used to encode the basic attribute features of users. The feature word weights are calculated by TF-IDF, and the cosine similarity is used to calculate the interest attribute similarity. The Jaccard coefficient is chosen to calculate the similarity of two users' social attributes. Considering the number of co-occurrences of reader behavior on the basis of the original SlopeOne algorithm, the weighted Slopeone algorithm is used to predict the ratings and improve the data sparsity problem of the collaborative filtering algorithm. The user similarity obtained based on the user portrait and the improved collaborative filtering algorithm respectively is fused using weights to obtain the comprehensive user similarity. Finally, personalized recommendation service for digital books is provided based on similarity.

2. Digital Resource Library User Data Acquisition and Processing Methods

2.1. User data acquisition and data processing

2.1.1. User data acquisition. Library readers' data are generated in the process of interaction between readers and the library system, and the data generated are rich and diverse. For the construction of readers' portraits, multi-source data is the basic guarantee for accurate portrayal of readers' portraits, and the more readers' related data are collected, the more accurately the portrait model can reflect readers' potential characteristics. Reader data are stored in various self-service systems within the library, and reader-related data are collected in an all-round way, including readers' basic information, borrowing records, in and out of the library records, punching time, use of electronic resources, and other information that can directly determine the labeling of the reader's portrait, so as to build a clear reader's portrait.

The acquisition of library readers' data starts from the four dimensions of natural attributes, readers' preference attributes, interactive behavior attributes and social behavior attributes, which can generally be obtained from the library backstage management system, gate clocking system or library search engine.

Readers' portrait is not static, in addition to static data that can reflect readers' characteristics, dynamic data that changes with time and readers' reading habits can sense the changes in time and dig deeper into the readers' potential needs, and this kind of data is the main source of data for refining the readers' portrait, and the dynamic data can be obtained through a set of independently constructed directed network data crawling and web page parsing functions based on the Python programming language. Web page parsing function to complete.

The requests-beautifulsoup4-re technical route used to build the crawling readers' behavioral data function covers obtaining useful web content from relevant networks, analyzing web information and extracting valuable data to be stored in appropriate data structures.

2.1.2. Data pre-processing. Usually there are inevitable data errors in the process of data collection, data acquisition and data transmission, and the data quality will directly affect the correctness of the

decision-making judgment, therefore, the historical behavioral data of library readers as the original data set, if it is directly applied without data cleaning, it will largely result in an inaccurate portrait, leading to bias in the recommendation results. Therefore, it is necessary to pre-process the data to improve the data quality as well as the accuracy and performance of the subsequent mining process.

1) Data Cleaning. The purpose of data cleaning is to remove dirty data that affects judgment, including filling in vacant values, smoothing abnormal data and correcting inconsistencies in the data.

Vacant values target the absence of certain attributes, these missing values may not be filled in completely when the reader enters, or the data may be lost due to the operational failure of the equipment. In this paper, the most widely applied method for dealing with vacant values is to predict the vacant data, and to fill in and replace the possible values based on the characteristics and direction of the data, using data analysis tools that can reflect the changes in the data to predict the direction of the data.

Noisy data are data that appear abnormal, and data with noisy data can be handled by the split-box method or clustering method. The split-box method uses neighboring data to fill in and smooth out existing data values. Taking the original data set as an example, all the data is divided into multiple data sets according to the classification rules, which can be viewed as putting into multiple boxes. Noisy data in each data set is smoothed by taking the mean, median or boundary value. The rules of binning include equal-depth binning, equal-width binning, and reader-defined binning. Equal-depth binning refers to dividing all the data using the same number of processing methods to ensure that the number of data elements in each box is the same. Equal width split box is divided into different intervals, each box according to the scope of the division of the storage of all the data under the interval. Customized split box is to consider the reader's needs and business scenarios, customize the split box rules to divide the data. Clustering is a systematic analysis method that analyzes the existence of a certain relationship between variables, in which isolated points can be detected by clustering, and similar isolated values are organized into clusters by the rule of “large differences between groups and small differences within groups”. The use of clustering method to divide the data, need to judge the reasonableness of the number of clusters, through a number of times the optimal number of clusters to choose, and finally form the clustering criteria. The use of clustering method for noise smoothing can be given a larger number of clusters to find isolated points that can not enter the class clusters, manually check whether these isolated points contain valuable information, and delete the data that are not useful.

2) Data integration and transformation. Data integration is to bring together data from multiple data sources in the same database, data integration should consider the matching of interfaces, data duplication and the impact of data conflicts. When the reading habits of library readers are deeply mined, if you want to add some data from other databases, such as adding data from the punch card system to analyze the time and characteristics of readers entering the library, you need to carry out data integration. When integrating data, it is necessary to consider the degree of matching between the unique identifiers and data transmission of the two systems, so as to minimize the occurrence of error reporting caused by data conflicts. In addition, some of the data obtained through the integration will follow the format customized by the previous system, and it is not possible to directly mine this part of the data. Therefore, it is necessary to transform the data format after data integration, and the transformation of data mainly takes the following ways:

First, the use of binning, clustering and other ways to remove the noise data.

Second, data are summarized and integrated.

Third, data generalization, using high-level concepts to replace the original low-level data.

Fourth, mapping a large amount of feature data proportionally to a small specific range, and performing rational scaling transformations of data attributes.

Fifth, creating new attributes from given attributes.

3) Data attribution. Data reduction is a parsimonious representation of the overall data set that allows the data to be reduced in size, but still comes close to maintaining the integrity of the original data and can produce similar or essentially the same analytical results. Data reduction can be done by dimensional summarization, data cube clustering, or by compressing the data values. The readers' data in college libraries have been accumulated over the years, from the establishment of the library for decades, and the datasets are generally large. And the college big readers are renewed every year, so when selecting the library readers' data, you can consider selecting the complete readers' data of the last four years for analysis, and the reference value of analyzing these data is greater.

2.2. Methods of categorizing readers

2.2.1. Factor analysis. Factor analysis is a data analysis method that downscales demand data. A common factor system is obtained by clustering the factors in the user demand preference section of the questionnaire according to their factor loadings. The general mathematical model of factor analysis can be expressed as Eq:

$$Y_j = a_{j1}F_1 + a_{j2}F_2 + a_{j3}F_3 + \cdots + a_{ji}F_i + U_i \tag{1}$$

In Eq. Y_j denotes the standardized score of the j nd variable, i.e., the factor score. F_i denotes the i th common factor, i.e., the factor that co-occurs in each variable. U_i is the special factor of Y_i, which refers to the part of the original variable that cannot be explained by the factor variables. a_{ji} represents the factor loadings, i.e. the contribution of the i th common factor to the variance of the j th variable.

2.2.2. Cluster analysis. Before segmenting users, a clustering algorithm is required to classify them into different categories based on the intensity exhibited by their interest attribute characteristics. In this study, one of the most commonly used algorithms in cluster analysis, K-Means clustering algorithm, is used to cluster groups of users with similar characteristics, so as to analyze the group characteristics of the users. K-Means clustering algorithm is an algorithm that performs the clustering operation on the data through the mean value, which divides the similar data points with the pre-set K value and the initial center of gravity of each class, and then after the division, iterates on the mean value for several times to optimize and adjust to obtain the optimal clustering results. Optimization adjustment to obtain the optimal clustering results [21]. The process is:

1) From the data set to the data object selected K clusters selected initial clustering center of mass, i.e., K Means.

2) Traverse all the data objects in the data set to calculate their similarity in terms of the length of the distance between the data objects and the center of mass, such that x represents a sample point in the data set, y represents the center of mass of the data set, m represents the number of features in each sample point, and i represents each feature composing the point x, and compute the minimum Euclidean distance $D(x, y)$ between the data object x and the center of mass y:

$$D(x, y) = \sqrt{\sum_{i=1}^{m} (x_{i-} y_i)^2} \tag{2}$$

3) Aggregate the data objects in the dataset in the same class of clusters according to the highest similarity.

4) Recalculate the center of mass of each cluster using the least error sum of squares Z:

$$Z = \sum_{j=1}^{K} (s_i - \overline{C}_i)^2 \tag{3}$$

5) Repeat steps 2, 3, and 4 until the algorithm achieves convergence.

3. Personalized recommendation algorithm for digital books

3.1. User similarity calculation based on user profiles

3.1.1. Basic attribute similarity calculation. In this paper, the age and nationality of users will be selected as the basic data for constructing users' basic attribute features. Considering each user's basic attribute features as an n-dimensional vector, one-hot coding is used to encode the user's basic attribute features and construct the user's basic attribute feature vector.

After vector representation of each word, the basic attribute feature vector of each user can be constructed. After constructing the basic attribute feature vectors of all users, the cosine similarity is utilized to calculate the user's basic attribute similarity, and the similarity of basic attributes between two users is recorded as sim_b. The formula for calculating the cosine similarity is shown below [22]:

$$sim_b = \cos(u,v) = \frac{u \cdot v}{|U||V|} = \frac{\sum_{i=1}^{n} U_i \times V_i}{\sqrt{\sum_{i=1}^{n} (U_i)^2} \times \sqrt{\sum_{i=1}^{n} (V_i)^2}} \tag{4}$$

where U and V denote the basic attribute feature vector of user u and the basic attribute feature vector of user v, respectively, and U_i and V_i denote the components of vector U and vector V, respectively.

3.1.2. Interest attribute similarity calculation. For the interest attribute dimension, user interest is shown through the user's rating behavior. Since the scoring value of this website is 0~10 points, the author believes that when the user's score for a book is greater than 5 points, it means that the user is interested in this book. Therefore, this paper selects the book name, book synopsis, publisher, author, and author's profile of each user whose rating is greater than 5 points as the basic data for constructing the user's interest attribute model, and constructs the user's interest attribute feature vector with these data. The weights of the feature words of the books rated greater than 5 by each user are calculated by TF-IDF, and each user takes the first ten words with the highest TF-IDF value as its feature words to construct the feature vector of each user's interest, and the cosine similarity is utilized to calculate the similarity between the feature vectors, so as to derive the similarity of the user's interest attributes.

3.1.3. Social Attribute Similarity Calculation. The similarity of the social attributes of two users is calculated through the Jaccard coefficient and the similarity of the calculated social attributes of the users is denoted as sims and the formula for calculating the similarity of the social attributes using the Jaccard coefficient is shown below [23]:

$$sim_s = \frac{\Sigma_{k=1}^{i} |N(u) \cap N(v)|}{\Sigma_{k=1}^{i} |N(u) \cup N(v)|} \tag{5}$$

where i denotes the category of the book, $N(u)$ denotes the category to which the book rated by user u belongs, $N(v)$ denotes the category to which the book rated by user v belongs, $N(u) \cap N(v)$ denotes the intersection of the categories of the book rated by user u and user v, and $|N(u) \cap N(v)|$ is the number of the categories in the intersection set, and $N(u)UN(v)$ denotes the concatenation of the categories of the book rated by user u and user v, and $|N(u)UN(v)|$ is the number of the categories in the concatenation set.

In this paper, we consider that the basic attributes are less important than the interest attributes and social attributes in the calculation of reader similarity, then the weights occupied by the basic attributes are less than those of the interest attributes and social attributes, so the weights are given

to the basic attributes, interest attributes, and social attributes in accordance with 0.2, 0.4, and 0.4 to get the user similarity based on the user profiles sim_{up}, and the calculation formulas are shown as follows:

$$sim_{up} = 0.2 \times sim_b + 0.4 \times sim_i + 0.4 \times sim_s \tag{6}$$

3.2. Improved collaborative filtering algorithm

Currently the actual user rating records of items in the public dataset of the recommendation domain account for a very low percentage of the user-item rating matrix, and the user-item rating data is very sparse. Therefore, the number of common ratings before the item and the number of common ratings between users are very small, which affects the recommendation quality.

To address the sparsity of user-item rating matrix data faced in collaborative filtering algorithms, in this paper, we use the weighted Slopeone algorithm to predict ratings and use the predicted ratings to populate the original user-item rating matrix, thus improving the sparsity of rating matrix data faced in collaborative filtering algorithms.

3.2.1. Slopeone Algorithm. The Slopeone algorithm is a matrix filling algorithm that predicts user ratings for unrated items and fills the predicted ratings into the original ratings matrix, thus producing a better recommendation even for users who originally had very little ratings data [24].

The process of predicting user ratings for item e using the Slopeone algorithm is as follows:

Step1: Calculate the rating deviation of the target item from other items.

Calculate the rating deviation between items (two items are rated at the same time), i.e., calculate the mean of the rating difference between items. For the given two items i and $j(i \neq j)$, the formula $dev_{i,j}$ for calculating the deviation between them is defined as shown in equation (7):

$$dev_{i,j} = \frac{\Sigma_{u \in N(i) \cap N(j)} (r_{ui} - r_{uj})}{|N(i) \cap N(j)|} \tag{7}$$

where r_{ui} denotes the rating of item i by user U, r_{uj} denotes the rating of item j by user u, $N(i)$ denotes the number of users who rated item i excessively, $N(j)$ denotes the number of users who rated item j excessively, $u \in N(i)ON(j)$ denotes the number of users who rated both item i and item i excessively, and $|IN(i)/N(j)|$ denotes the number of users who rated both item i and item i excessively.

Step2: Calculate the possible ratings of users for unrated items.

Based on the rating deviation between items and the users' historical ratings, the users' ratings for unrated items are predicted. The formula for calculating the predicted ratings is as follows:

$$p_{uj} = \frac{\Sigma_{i \in N(u)} (r_{ui} - dev_{i,j})}{\Sigma_{i \in N(u)} |N(i)|} \tag{8}$$

where $N(u)$ represents the number of items rated by user u and $|N(i)|$ represents the number of items rated by user i.

3.2.2. Weighted Slopeone Algorithm. The original SlopeOne algorithm does not take into account the number of users who rate any two items simultaneously, i.e., the effect of the number of co-occurrences between items on the predicted score. Based on the original SlopeOne algorithm, weighted SlopeOne takes into account the number of co-occurrences of items, and the more the target item and any other item co-occur, the higher the confidence of its predicted score.

Considering the different contributions of the number of co-occurrences between items to the calculation of the deviation of item ratings, the weighted SlopeOne algorithm introduces the number

of co-rated users as a weight into the rating prediction formula, and weights the average deviation of the target item and other different items, and the improved formula is shown below:

$$p_{uj} = \frac{\sum_{i \in N(u)} |N(i) \cap N(j)| (r_{ui} - dev_{i,j})}{\sum_{i \in N(u)} |N(i) \cap N(j)|} \tag{9}$$

In this paper, we use the weighted SlopeOne algorithm to predict the ratings as well as backfill the user-item ratings matrix, and then calculate the user similarity based on the backfilled ratings matrix using the cosine similarity to get the similarity of collaborative filtering based on the weighted SlopeOne improvement sim_{cf}.

3.2.3. Comprehensive user similarity calculation and recommendation based on comprehensive similarity. The user similarity obtained based on the user profile and the improved collaborative filtering algorithm respectively are fused using appropriate weights to obtain the comprehensive user similarity, which is calculated using the formula shown below:

$$sim(u,v) = \alpha sim_{cf}(u,v) + (1-\alpha) sim_{up}(u,v) \tag{10}$$

In the above equation, α is the fusion coefficient that is subsequently determined through experiments and takes the value in the range of [0, 1]. When $\alpha = 0$, the similarity computation considers only the method based on user profiling. When $\alpha = 1$, the similarity calculation only considers the method based on weighted SlopeOne improved collaborative filtering.

Based on the combined user similarity, the k nearest neighbor users that are most similar to the target user can be identified. Then, through the ratings of the books by the nearest neighbor users and the comprehensive similarity between the users, the predicted ratings of the target users for all the unspecified weird books can be calculated, and the formula for calculating the predicted ratings is shown below:

$$P_{u,i} = \overline{r_u} + \frac{\Sigma_{v/s_N} (r_{v,i} - \overline{r_V}) \times sim(u,v)}{\Sigma_{v/s_N} |sim(u,v)|} \tag{11}$$

where $P_{u,i}$ denotes the predicted rating of user u for book i, $\overline{r_u}$ and $\overline{r_v}$ denote the average ratings of user u and user v for all books, respectively, S_N denotes the set of all near-neighbor users of user u, r_{vi} denotes the rating of user v for book i, and $\mathrm{sim}(u,v)$ denotes the combined similarity of user u and user v.

After calculating the predicted ratings of the target users for all unrated items, the items are sorted according to the predicted ratings from highest to lowest, and the Top-N items with the highest predicted ratings are finally selected and recommended to the target users.

3.3. Book recommendation based on feature similarity

After the book attribute features and user interest features are represented by vectors, the similarity between the book attribute vectors and the user interest vectors is calculated by cosine similarity, and the Top-N books with the highest similarity are finally recommended for the user according to the similarity ranking from high to low.

This paper proposes a personalized recommendation algorithm that integrates user profiling and hybrid recommendation algorithm, the flowchart of the algorithm is shown in Figure 1, and the specific implementation steps of the algorithm are as follows:

Step1: Construct the user profile of online book website users, and provide a comprehensive description of users from basic attributes, interest attributes and social attributes. According to the constructed user profile, the user similarity of the three dimensions is calculated, and the user

similarity of the three dimensions is fused with appropriate weights to obtain the similarity sim_{up} based on the user profile.

Step2: To improve the user-based collaborative filtering algorithm for the sparsity problem, use the weighted Slopeone algorithm to fill the user-item scoring matrix, and calculate the user similarity based on the filled scoring matrix to get the similarity based on collaborative filtering sim_{cf}.

Step3: Use the similarity fusion coefficient α to assign different weights to sim_{up} and sim_{cf}, to fuse the two similarities to obtain the user's comprehensive similarity, find similar nearest neighbor users according to the user's comprehensive similarity, and get the recommendation results based on the similar nearest neighbor users.

Step4: Generate word vectors based on Word2Vec for book attribute data, use K-Mean algorithm to cluster the word vectors, construct book attribute feature vectors and user interest feature vectors respectively, calculate the similarity between book attribute feature vectors and user interest feature vectors, and generate recommendation results based on content recommendation based on the similarity.

Step5: Integrate the recommendation results obtained from Step3 and Step4, take their concatenation and get the final recommendation results.

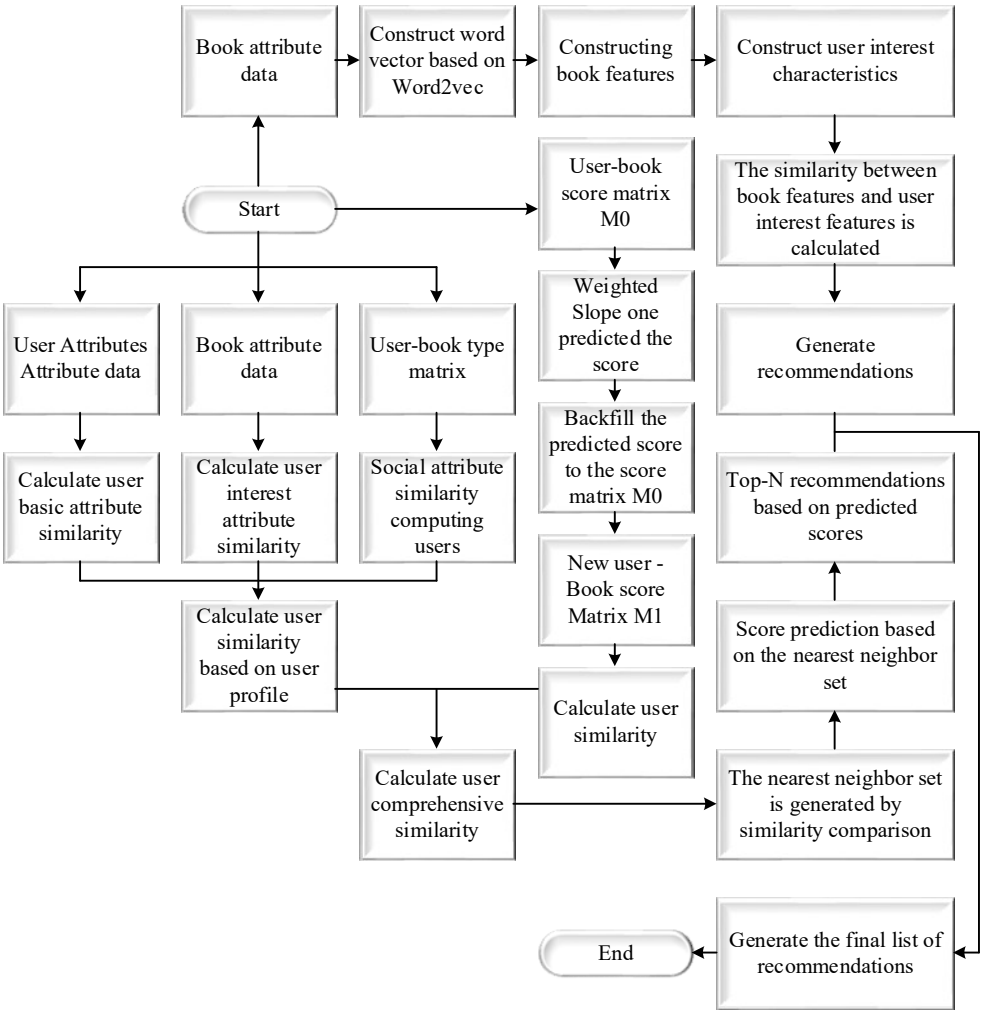

Fig. 1. Flowchart of book personalized recommendation algorithm

4. Reader Profile Visualization and Personalized Service Testing

4.1. Reader Portrait Presentation

4.1.1. Data selection and processing. In this paper, we choose the readers' borrowing data of Q City Digital Library from 2020 to 2023 as the source of analysis, and after screening the missing data and abnormal data, we finally retained 683 valid readers' data and a total of 2827 borrowing data.

The data source consists of explicit data and implicit data, and the explicit data includes readers' basic attributes and book data, where readers' basic attribute data includes readers' names, contact information, and ID numbers, and book data includes book titles, authors, and genres. The implicit data is expressed by the subject words of the books borrowed by the readers, mainly including the readers' reading theme preference in total will be divided into 25 reading themes, including poetry and prose, literature (fairy tale and fable, sociology, etc.), history, music, travel, food, art, life (parenting, sports and fitness, attitude towards life, etc.), reasoning and suspense, gender, emotion, inspirational, healing, sci-fi, magical fantasy, war, novels, Growth, Psychology, Popular Science, Teaching and Learning, Economy, Faith, Miscellaneous, Philosophy. Each subject term is used as a label under the dimension of reader behavioral information, and the type and frequency of books borrowed by readers are counted. Each book type has a corresponding subject term, and the reader's interest preference can be seen from the counted frequency, and at the same time, the frequency is set as the weight of the label.

In order to facilitate the processing of data, this paper coded the acquired reader information and the frequency of occurrence of content topics, as shown in Table 1. Excluding the demographic characteristic variables of readers, this paper conducts a reliability test on 25 topic variables with the help of SPSS software to determine the quality of the data and decide whether the data is suitable for analysis.

Table 1. Reader data processing(part)

Reader info code	Subject word coding							
Reader	Poetic prose	Literature	History	Music	Travel	Food	Art	Life
1	0	1	0	0	0	0	0	0
2	0	1	0	0	0	1	0	0
3	0	0	0	1	0	0	0	0
4	1	1	1	0	0	0	0	1
5	0	0	0	0	0	0	1	1
6	1	0	1	0	0	0	0	1
7	0	1	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0
9	0	1	0	1	1	1	0	0
10	0	0	0	0	0	0	1	0

KMO and Bartlett's spherical test are shown in Table 2, the KMO value and Bartlett's spherical test of the variables are ideal, the KMO value is 0.788, which is greater than 0.7, and the level of significance is $P<0.01$, there is a meaningful relationship between the 25 variables and is suitable for factor analysis.

Table 2. The KMO and Bartlett's Test of Sphericity

KMO		0.788
Bartlett's spherical test	Approximate chi square	6585.681
	df	400

	Sig.	0.000

4.1.2. Label Extraction and Classification. The differences in reader portraits lie in the variability of labels, and in order to obtain the types of labels with different characteristics, the 25 variables of the reader behavioral information dimension are subjected to dimensionality reduction processing to extract the characteristic factors. In SPSS software, firstly, the standardization process is carried out, and then the factor analysis in the analysis operation is selected, and the number of factors is set to be obtained by the maximum variance rotation method, and the usual extraction principle is that the eigenvalue is >1.

Table 3 shows the explained total variance of the factor analysis of variables. It is specified that the extraction of seven male factors explains a total of 62.971% of the variance of the original variable, i.e., the rotated sum of squares is loaded, and the factor analysis is more satisfactory.

Table 3. Variable factor analysis explains total variance

Factor	Initial eigenvalue			Extracted sum of squares load			Rotated sum of squares load		
	Sum.	Var.%	Acc.	Sum.	Var.%	Acc.	Sum.	Var.%	Acc.
1	5.728	22.872	22.872	5.728	22.872	22.872	3.558	14.341	14.341
2	2.639	10.421	33.293	2.639	10.421	33.293	2.786	11.400	25.741
3	2.145	8.597	41.890	2.145	8.597	41.890	2.388	9.476	35.217
4	1.518	6.116	48.005	1.518	6.116	48.005	2.139	8.616	43.833
5	1.475	5.840	53.845	1.475	5.840	53.845	1.937	7.794	51.627
6	1.252	5.020	58.866	1.252	5.020	58.866	1.629	6.539	58.166
7	1.038	4.105	62.971	1.038	4.105	62.971	1.224	4.805	62.971
8	0.999	3.925	66.896						
9	0.962	3.752	70.648						
10	0.904	3.684	74.332						
11	0.823	3.345	77.677						
12	0.777	3.192	80.869						
13	0.662	2.725	83.594						
14	0.595	2.396	85.991						
15	0.501	2.038	88.028						
16	0.454	1.803	89.831						
17	0.419	1.804	91.635						
18	0.359	1.520	93.155						
19	0.343	1.414	94.569						
20	0.312	1.370	95.939						
21	0.305	1.117	97.057						
22	0.268	0.995	98.052						
23	0.194	0.880	98.932						
24	0.210	0.694	99.625						
25	0.135	0.375	100.000						

Table 4 shows the rotated component matrix. The variables in the table are arranged according to the size of the loadings, and the loadings of the rotated factors are obviously polarized to the levels of 0 and 1, which is easy to see that the variables are categorized into the category of public factors. Combining the above two factors in the table, considering the principle of public factor selection, this paper will select seven public factors as the classification of the label.

The descriptions are as follows: Tag 1: The male factor of this label is named “Leisure”. This kind of factor is related to war, inspiration, travel, music, food, literature (fairy tales, fables, sociology), science and technology.

Tag 2: The common factor of this tag is named “Exploration”. This category is related to science fiction, mystery, coming-of-age, and magic themes.

Tag 3: The metric of this tag is named “Reflective”. This factor is related to the themes of essays, teaching aids, and emotions.

Tag 4: The male factor of this tag is named “Literature”. These factors are related to the themes of gender, philosophy, healing, art, and faith.

Tag 5: The common factor of this tag is named “Literature and History”. These factors are related to the themes of History, Poetry and Prose.

Tag 6: The common factor for this tag is named “Social”. This category is related to the topics of economy and life (parenting, sports and fitness, attitude towards life).

Tag 7: The common factor for this tag is named “Extended”. These factors are related to the themes of Fiction and Psychology.

Table 4. Rotating component matrix

	Component						
	1	2	3	4	5	6	7
War	**0.805**	0.024	-0.003	0.201	-0.019	-0.021	-0.014
Inspiration	**0.744**	-0.043	0.154	0.282	-0.121	0.108	-0.047
Travel	**0.751**	0.199	-0.035	0.225	0.103	0.089	0.153
Music	**0.714**	-0.003	-0.017	-0.104	0.227	0.016	0.184
Delicacies	**0.642**	0.211	-0.009	-0.123	0.464	-0.029	-0.008
Literature	**0.637**	0.060	0.004	0.081	0.493	0.189	-0.055
Popularization	**0.350**	0.238	0.202	-0.021	-0.211	0.219	0.112
Science fiction	0.080	**0.874**	0.009	0.228	-0.066	0.025	0.152
Inference suspense	0.148	**0.792**	0.071	-0.003	0.382	0.018	0.103
Grow	0.068	**0.730**	-0.041	0.180	-0.064	0.104	0.061
Magic	0.036	**0.674**	0.477	0.087	-0.069	0.182	0.121
Miscellaneous	-0.016	0.017	**0.856**	0.065	-0.021	-0.030	-0.024
Teaching and auxiliary	0.053	0.020	**0.838**	0.053	-0.003	-0.050	-0.006
Affections	0.014	0.176	**0.659**	-0.193	0.357	0.180	-0.090
Gender	0.024	-0.043	0.120	**0.635**	0.248	-0.137	0.262
Philosophy	0.072	0.356	-0.027	**0.611**	0.006	0.106	-0.140
Cure	0.232	0.016	0.470	**0.535**	-0.004	0.018	0.213
Art	0.165	0.154	-0.086	**0.540**	0.053	0.092	0.105
Faith	0.098	0.220	-0.036	**0.517**	0.195	0.324	-0.142
History	0.067	-0.076	0.085	0.191	**0.767**	0.014	0.083
Poetic prose	0.422	0.157	0.009	0.320	**0.642**	-0.033	0.132
Economy	0.049	0.033	0.038	0.057	-0.012	**0.860**	0.069
Life	0.145	0.167	0.010	0.144	0.057	**0.751**	0.161
Novel	0.023	0.113	-0.036	0.060	0.133	0.041	**0.700**
Mind	0.112	0.120	0.014	0.046	-0.022	0.141	**0.650**

4.1.3. Display of reader profiles. After obtaining the seven labeling metrics mentioned above, the metrics were used as variables in the cluster analysis, and the K-means clustering algorithm was chosen to cluster all the reader samples in order to establish the number of reader portraits. The use of K-means clustering method to manage readers can better explain and characterize readers, and given the classification group K (K<<n), the samples of similarity are aggregated together to form different classes or clusters, trying to achieve similarity within clusters and difference between clusters. In this paper, the number of clusters is set to 4-7 classes and analyzed under different number of clusters of 4, 5, 6 and 7 to compare the magnitude of the differences in F-value of each class, and the comparison of the differences under different clusters is shown in Table 5. It can be seen that when the number of clusters is 5, in which the significant value of “Arts and Culture” is 0.454>0.05, so it is not taken. When the number of clusters is 6 and 7, only 1 reader exists in two of them respectively, so they are not taken. When the number of clusters is 4, the Sig value of each male factor is significant, and the difference in the F value of each male factor of the clusters is obvious, therefore, this paper sets the number of clusters as 4.

Table 5. Comparison of differences under different clustering conditions

Clustering	Clustering number=4		Clustering number=5		Clustering number=6		Clustering number=7	
Variable	F	Sig.	F	Sig.	F	Sig.	F	Sig.
Leisure	281.700	0.000	225.425	0.000	212.876	0.000	188.700	0.000
Exploration	93.471	0.000	3.585	0.000	4.992	0.000	185.364	0.000
Thinking	61.381	0.000	194.198	0.000	126.679	0.000	118.193	0.000
Literature	4.869	0.000	0.957	0.454	307.894	0.000	243.289	0.000
History	57.585	0.000	365.724	0.000	286.827	0.000	241.712	0.000
Society	250.802	0.000	148.224	0.000	167.259	0.000	140.204	0.000
Extension	6.025	0.000	4.662	0.001	2.503	0.000	14.350	0.000

When the number of selected clusters is 4, the mean center value of each public factor in each category of portraits is calculated, and the mean values of the four categories of readers' portraits are shown in Table 6, and there is a significant variability in each category of portraits in different eigenfactors.

1) The “type one readers”, who are mainly driven by objective factors and competence, can be named “pragmatic readers”, who focus on the knowledge contained in books, their quality and quantity, with the main purpose of improving their own competence. Its distinctive feature is its preference for resources on economic and life topics, and its reading preference is closely related to its own occupation, identity, family status, and stress relief.

2) Type 2 readers, who are mainly subjective and ability-oriented, can be named “youthful readers”, who focus on the learning value of books, the enhancement of their professional skills and the gains they bring, and focus on the pursuit of a psychological sense of acquisition. This group of readers focuses on the learning value of books, the enhancement of their professional skills and the gains they bring, and focuses on the pursuit of psychological gains, which is characterized by a preference for resources on gender, philosophy, healing, art, faith, fiction, and psychological themes. The reading theme preference of this group of readers is closely related to their experience, vision, and stage of growth.

3) The subjective factors and preference-oriented “type 3 readers” can be named “recreational readers”, which use digital libraries mainly to soothe their emotions, expand their horizons, and spend their time, and are characterized by their preference for Their distinctive feature is their preference for resources with the themes of science fiction, reasoning and suspense, growth, magic, miscellaneous articles, teaching aids, emotions, history and poetry and prose. This category of readers pays more attention to books with suspenseful science fiction themes, and also includes miscellaneous articles, teaching aids, and literature and history.

4) The “Type 4 readers” who are mainly driven by objective factors and preferences can be named “Curious readers”, who use digital libraries less frequently and are mainly curious to learn about digital library equipment and to get a sense of experience from using digital libraries. They are mainly curious in order to learn about the digital library equipment and get the experience of using digital libraries, and they are characterized by their preference for resources on war, inspiration, travel, music, food, literature, and popular science. These patrons have a shorter average borrowing time and use digital libraries less frequently.

Table 6. The mean values of the four reader profiles

Reader portrait classification	Leisure	Exploration	Thinking	Literature	History	Society	Extension
Type 1	0.667	-0.413	0.468	1.118	0.450	**8.259**	0.406
Type 2	-0.080	-0.094	-0.075	**-0.020**	-0.069	-0.054	**-0.021**
Type 3	0.182	**3.054**	**2.622**	0.530	**2.574**	0.014	0.701
Type 4	**8.466**	-0.801	-0.049	0.656	-0.211	-0.718	1.098

4.2. Algorithm Performance Testing

In order to verify the practical application of the personalized recommendation system in this paper, the algorithm performance test is set up in this paper.

4.2.1. Test Sets and Evaluation Metrics:

1) Take the test set. The 14000 records of the book borrowing behavior data set are randomly divided into N parts, usually N is not too big, here take N=6, randomly select one part of the data set as the test set of the experiment, and the other N-1 parts of the training set. The training set is used as the basis to construct the borrowing pattern of readers, and then the data in the test set is used to realize the personalized recommendation of books. In the experiment, in order to reduce the chance of the results and improve the stability and reliability of the indicators, we conducted N trials, that is, N sets of data take turns to be the test set, and finally count the average value of each indicator as the final evaluation index of the algorithm.

2) Calculate the evaluation index. According to the obtained test set ListP and training set ListQ can start to calculate the indicators of the algorithm. Accuracy reflects the likelihood of recommendation results being selected. The novelty measure is mainly designed to solve the problem of new items not being recommended, portraying how the algorithm handles when a new book is added.

4.2.2. Analysis of experimental results. After calculating the similarity between readers through the improved similarity algorithm, the algorithm in this paper will recommend the K readers' favorite books that are most similar to his interests. In the experiments, the values of α and β are constantly going to be adjusted so that the weights of the two similarities are constantly changing to get the best experimental results. In order to get the performance of the three algorithms under different values of α and β, our experiments cover all the historical access information of the readers, and the accuracy comparison of the three recommendation algorithms is shown in Fig. 2, where the horizontal coordinate indicates the value of β and the vertical coordinate indicates the accuracy of the three algorithms. Comparison of the recommended correctness curves of the three algorithms shows that the accuracy shows an increasing trend with the increase of the value of β. We conducted several groups of experiments, and the final experimental results show that the algorithm achieves a better recommendation when $\alpha = 0.38$ with $\beta = 0.65$, and it can also be seen that our algorithm has a higher accuracy rate compared to the other two algorithms. The specific α and β combination value and accuracy rate correspondence is shown in Table 7.

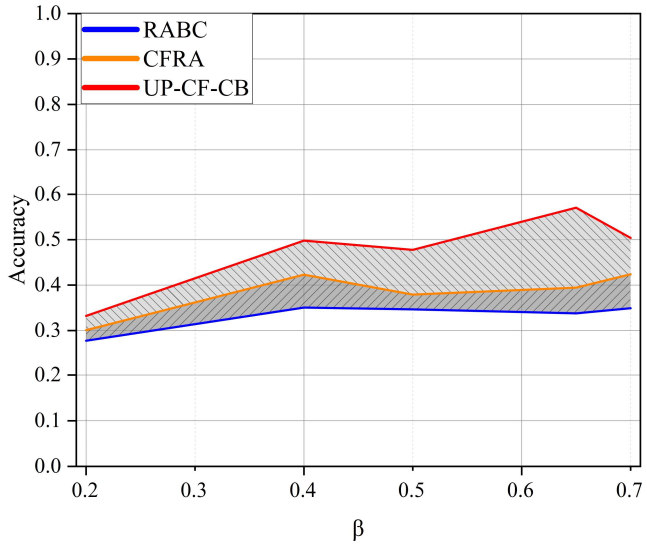

Fig. 2. Accuracy of three algorithms

Table 7. The corresponding accuracy of the combination of α and β

	α=0.8	α=0.6	α=0.5	α=0.38	α=0.3
	β=0.2	β=0.4	β=0.5	β=0.65	β=0.7
RABC	27.77%	35.06%	34.66%	33.80%	34.93%
CFRA	30.07%	42.31%	37.92%	39.45%	42.38%
UP-CF-CB	33.23%	49.81%	47.75%	57.06%	50.41%

We compare the advantages and disadvantages of the three algorithms by adding new books in batches, and after several experimental comparisons, our algorithms have obvious advantages in processing new items, as shown in Fig. 3, where the horizontal coordinate still represents the value of β and the vertical coordinate represents the novelty of the three algorithms.

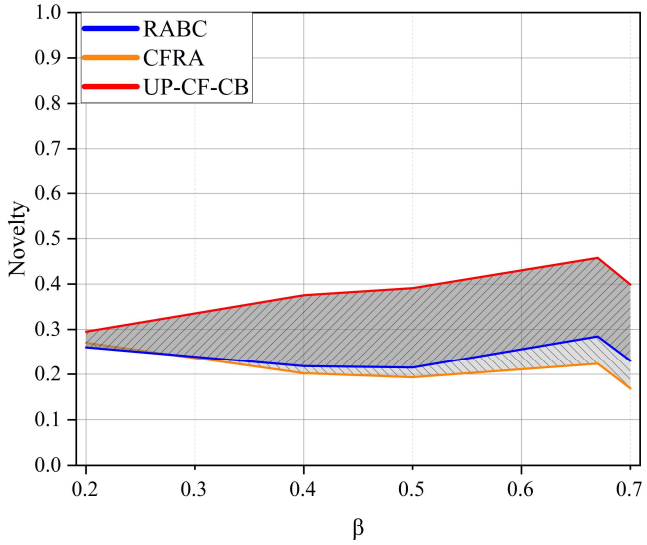

Fig. 3. The novelty of three algorithms

Firstly, 1000 book information is added as new items in the test set, and then the experiment is carried out to count the frequency percentage of new books being recommended, here α and β are the same as the previous experiments, which are taken in accordance with various combinations of

situations, and then new books are added in turn, and the experiments are repeated to obtain the results of the experiments of each group, and the results of the experiments are shown in Table 8.

Observing the experimental results, we found that: the three algorithms do not change much before and after in terms of novelty, and there will not be big ups and downs because of the increase or decrease of new items, and at the same time, we also found that, as far as the comparison of the three algorithms is concerned, our algorithm has a higher degree of novelty. Also observing under different combinations of α and β, we find that similar to the results of previous experiments, our algorithm has the best novelty when the values of α and β are 0.31 and 0.67, respectively.

Table 8. The corresponding novelty of the combination of α and β

	α=0.8	α=0.6	α=0.5	α=0.31	α=0.3
	β=0.2	β=0.4	β=0.5	β=0.67	β=0.7
RABC	26.04%	21.82%	21.52%	28.44%	23.06%
CFRA	27.06%	20.23%	19.36%	22.35%	16.88%
UP-CF-CB	29.52%	37.54%	39.10%	45.79%	39.91%

One of the innovations in this paper is the improved similarity calculation method, in order to verify the effectiveness of the algorithm, we have tested the accuracy of different similarity calculation methods. For the same batch of books, we used the inner product method and the cosine method to calculate the similarity between the books respectively, and compared the differences between the two, and finally found that the similarity of the books obtained by the inner product method is more in line with the actual situation.

In that experiment, we carried out two tests: the first one was to randomly select 1000 books and cluster them with the two similarity calculation methods respectively, and through the experiment, we found that the results of clustering were coincident. The second test was to choose 5 similar books each time, in which a book with interference factors was deliberately put in, and then carry out similarity calculation with the specified books to select the book with the highest similarity, we carried out several tests, and the final results showed that all the methods of similarity calculation using the cosine method succeeded in selecting the books that matched the actual situation, while the methods of similarity calculation using the inner product method selected many times the interfering books. The following is an example of one of the test data, in which we specify the name of the book as "Computer Network Technology Project-based Tutorial", and the specific results are shown in Table 9.

The actual situation of this experiment is that the content similarity between "Project-based Tutorial on LAN Formation and Maintenance" and the specified book is the highest, while the result obtained by using the inner product method is "Computer Network Project-based Case Tutorial". The synthesis of the above experimental results shows that the recommendation algorithm proposed in this paper has the best recommendation quality in dealing with the cold-start problem. It also has certain advantages in accuracy and novelty.

Table 9. The comparison of two different calculation methods

Name of books	Similarity (inner product method)	Similarity (cosine)
Computer network project case tutorial	0.22331	1.24818
LAN building and maintenance project tutorial	0.14176	1.28744
Computer network establishment and maintenance	0.14228	0.99955
Website construction and management	0.15905	0.98681
Network operating system	0.12351	1.01249

5. Conclusion

The data of Q city digital library is chosen to portray the readers' portraits, and seven common factors are extracted from the factor analysis, which are further clustered to obtain four types of readers, namely: pragmatic, youthful, recreational and curious readers. After completing the reader profile, the performance test of personalized service recommendation is carried out. This paper's algorithm consistently outperforms the CFRA and RABC algorithms, and the recommendation accuracy of this paper's algorithm is highest when $\alpha = 0.38$ and $\beta = 0.65$. With the addition of new books, this paper's algorithm has a high degree of novelty, and the values of α and β are 0.31 and 0.67, respectively, when the degree of novelty is the highest. Experiments show that the cosine method is used to calculate the similarity more in line with the actual situation. Therefore, the improved method in this paper is feasible and can provide readers with book recommendation service, which is conducive to constructing the personalized service system of digital resource library.

References

[1] Li, S., Hao, Z., Ding, L., & Xu, X. (2019). Research on the application of information technology of Big Data in Chinese digital library. Library Management, 40(8/9), 518-531.

[2] Satheesh, S. N., Antony, V. R., Anantharaman, S., & Ashraf, R. C. (2024). Artificial Intelligence in Library Services: Enhancing Access, Operational Efficiency, and User Experience. Library Progress International, 44(3), 5838-2843.

[3] Sun, L., Li, Y., & Lu, Y. (2022). Construction of Cloud Library Intelligent Service Platform Relying on Artificial Neural Network. Mobile Information Systems, 2022(1), 6259127.

[4] Tian, Z. (2021, June). Application of artificial intelligence system in libraries through data mining and content filtering methods. In Journal of Physics: Conference Series (Vol. 1952, No. 4, p. 042091). IOP Publishing.

[5] Perdana, I. A., & Prasojo, L. D. (2020, February). Digital Library Practice in University: advantages, challenges, and its position. In International Conference on Educational Research and Innovation (ICERI 2019) (pp. 44-48). Atlantis Press.

[6] Kato, A., Kisangiri, M., & Kaijage, S. (2021). A review development of digital library resources at university level. Education Research International, 2021(1), 8883483.

[7] Alphonce, S., & Mwantimwa, K. (2019). Students' use of digital learning resources: diversity, motivations and challenges. Information and Learning Sciences, 120(11/12), 758-772.

[8] Sivasankari, R., Suriya, S., Sindhu, S., Devi, J. S., & Dhilipan, J. (2024). AI-Powered Recommendation Systems and Resource Discovery for Library Management. In Applications of Artificial Intelligence in Libraries (pp. 223-244). IGI Global.

[9] Barsha, S., & Munshi, S. A. (2023). Implementing artificial intelligence in library services: A review of current prospects and challenges of developing countries. Library Hi Tech News, 41(1), 7-10.

[10] Wang, C., & Sha, Z. (2021, November). Research on intelligent information system of library under big data and digitization technology. In Journal of Physics: Conference Series (Vol. 2083, No. 4, p. 042063). IOP Publishing.

[11] Jing, Z. (2021, June). Application of data mining based on computer algorithm in personalized recommendation service of university smart library. In Journal of Physics: Conference Series (Vol. 1955, No. 1, p. 012008). IOP Publishing.

[12] Xie, J. (2023). The application of artificial intelligence technology in public library information retrieval. Transactions on Computer Science and Intelligent Systems Research, 1, 119-127.

[13] Lin, G., & Miao, A. (2022, April). Construction of Library Personalized Intelligent Service System Based on Data Mining. In Proceedings of the 3rd Asia-Pacific Conference on Image Processing, Electronics and Computers (pp. 857-862).

[14] Kang, X. (2022, April). Personalized recommendation system of smart library based on deep learning. In International Conference on Multi-modal Information Analytics (pp. 1011-1016). Cham: Springer International Publishing.

[15] Wang, J. (2022). Personalized information service system of smart Library based on multimedia network technology. Computational Intelligence and Neuroscience, 2022(1), 2856574.

[16] Sa'ari, H., Sahak, M. D., & Skrzeszewskis, S. (2023). Deep learning algorithms for personalized services and enhanced user experience in libraries: a systematic review. Mathematical Sciences and Informatics Journal (MIJ), 4(2), 30-47.

[17] Wang, C., Sha, Z., & Jia, L. (2021, June). Research on digital resource system construction of smart library based on computer network and artificial intelligence. In Journal of Physics: Conference Series (Vol. 1952, No. 4, p. 042018). IOP Publishing.

[18] Okunlaya, R. O., Syed Abdullah, N., & Alias, R. A. (2022). Artificial intelligence (AI) library services innovative conceptual framework for the digital transformation of university education. Library Hi Tech, 40(6), 1869-1892.

[19] Liu, M. (2022). Personalized recommendation system design for library resources through deep belief networks. Mobile Information Systems.

[20] Pang, N., & Dou, C. (2023). Artificial intelligence inspired computer-aided design of library service system. Computer-Aided Design and Applications, 20, 53-63.

[21] Su Wutyi Hnin,Jessada Karnjana,Youji Kohda & Chawalit Jeenanunta. (2024). A hybrid K-means and KNN approach for enhanced short-term load forecasting incorporating holiday effects. Energy Reports5942-5959.

[22] GulELaraib ,Zaman Sardar Khaliq uz,Maqsood Tahir,Rehman Faisal,Mustafa Saad,Khan Muhammad Amir... & Elmannai Hela. (2023). Content Caching in Mobile Edge Computing Based on User Location and Preferences Using Cosine Similarity and Collaborative Filtering. Electronics(2),284-284.

[23] Li Zhongbing,Zheng Xinyu,Chen Guihui,Wei Yuli & Lu Kai. (2023). A Matching Pursuit Algorithm for Sparse Signal Reconstruction Based on Jaccard Coefficient and Backtracking. Circuits, Systems, and Signal Processing(10),6210-6227.

[24] Hu Qiang,Qi Haoquan,Huang Wen & Liu Minghua. (2023). A method to recommend cloud manufacturing service based on the spectral clustering and improved Slope one algorithm. Journal of Cloud Computing(1).