International Journal for Housing Science and Its Applications
Publish August 5, 2025. Volume 46, Issue 3 Pages 3842-3853

https://doi.org/10.70517/ijhsa463326

Research on Multimodal Data Fusion Strategy in the Construction of Personalized Service System of Digital Resource Library

Zhenyi An[1,*]
[1] Library of Guizhou Minzu University, Guiyang, Guizhou, 550025, China
Corresponding authors: (e-mail: 18285021109@126.com).

Abstract College libraries are important information centers in colleges and universities, which are important positions for knowledge storage, communication and discovery. With the development of social digitization and informatization, the quantity and variety of information are changing day by day, and the information needs of university user groups are also changing. Research from data collection and analysis, label extraction to portrait mining applicable to the library user portrait model, combined with hybrid recommendation algorithms from the data collection layer, processing layer, fusion layer, service application layer four levels to build a multimodal data-enabled smart library structure. The K-means algorithm was used to cluster and analyze library users, and four types of users were formed: "pragmatic", "youthful", "recreational" and "curious". The accuracy, recall and comprehensive value F, which are commonly used to evaluate the recommendation effect, are chosen to validate the recommendation results, and the hybrid recommendation integrating user profiles is higher than the traditional collaborative filtering recommendation in recall and F value, and the recommendation effect is good.

Index Terms user portrait, personalized recommendation service, K-means, smart library

I. Introduction

In the context of the information age, the way people acquire knowledge has also undergone new changes. The traditional library service mode has been difficult to meet the diversified and personalized needs of readers, who are no longer satisfied with simply borrowing books, but expect to obtain more accurate, efficient and convenient services [1]-[3]. In this context, digital libraries have emerged, which provide readers with knowledge acquisition channels by virtue of rich digital resources, convenient access and wide coverage [4]. However, with the increase in the number of digital library users, its management and service gradually tend to be complicated. How to effectively manage readers and meet their individualized needs has become an urgent problem for digital libraries [5]-[7].

With the rapid development of the Internet, Internet of Things, social media and other technologies, huge and diverse data are constantly generated, which cover all aspects from individual behavior to social phenomena [8]. The advent of the big data era implies the need to adopt new perspectives and technological means to understand, analyze, and utilize these massive data in order to better solve problems in the social, economic, and scientific fields [9]-[11]. Digital library management supported by big data analytics technology emphasizes the efficient management, deep mining and intelligent application of data [12], [13]. Through techniques such as advanced analytics, machine learning, and artificial intelligence, the laws, trends, and information hidden in multimodal data are discovered to provide more accurate support for carrying out personalized services [14]-[16]. Therefore, in the era of big data, digital libraries are actively exploring personalized service models, striving to better meet users' needs, thus enhancing user experience [17], [18].

This paper applies the user image to build the intelligent library field, starting from data collection, data processing, extracting labels to form the image and other aspects to gradually build the library user image model. Combined with the hybrid recommendation algorithm, the library personalized recommendation service system architecture is proposed on this basis, and a multimodal data-enabled smart library is constructed from four levels: data collection layer, processing layer, fusion layer, and service application layer. The collected data are processed to obtain different label types, and the K-means clustering method is used to achieve user classification, carry out personalized recommendation services for different users according to their characteristics, improve the utilization rate of library resources, and solve the problem of mismatch between book resources and readers' needs.

II. Construction of library user profile model under multimodal data fusion

II. A. Reader Data Acquisition and Data Processing

II. A. 1) Reader data acquisition

The research builds the requests-beautifulsoup4-re technical route used for crawling user behavioral data functions, covering the acquisition of useful web page content from relevant networks, analyzing web page information and extracting valuable data to be stored in the appropriate data structure.

Relatively stable and not easy to change static data that need to be exported from the system, attention needs to be paid to correlating user information when exporting, using the student's academic number as a unique identifier, and correlating other information of the students corresponding to the same academic number, so as to realize the data integration among various service systems. In addition, the user portrait is not static, in addition to static data that can reflect the characteristics of readers, dynamic data that changes with time and readers' reading habits can sense the changes in a timely manner 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. Dynamic data can be obtained through a set of directed web data crawling and web page parsing functions based on Python programming language.

II. A. 2) Data pre-processing

Usually there are unavoidable 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 users, as the original dataset, will largely cause inaccurate portraits and lead to bias in the recommendation results if it is directly applied without data cleansing. 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.

(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.

(3) Data generalization

Data reduction is a simple representation of the overall data set, so that the data in the capacity to reduce, but still close to maintaining the integrity of the original data, and can produce similar or basically the same analysis results. Data generalization can be used dimensional generalization, data cube clustering, or the data value compression method.

II. B. Image Mining

II. B. 1) Label Categorization Extraction

The essence of labeling is to abstract key information from a large amount of complicated data that can represent user characteristics and describe an accurate portrait. The data quality of this part of the data has a great impact on the next data mining process. Too many data dimensions will waste the time of analyzing and processing data, while too few will affect the precision of data analysis and fail to make an accurate portrait, so it is necessary to find a balance between precision and efficiency, and summarize a calculable and readable user model based on user information.

II. B. 2) Label similarity calculation

The premise of forming a user portrait is to gather users with the same characteristics together, and similarity is used to find the connection between readers and books. According to the content that constitutes the user's characteristics as well as the form of data storage, the user's natural attribute similarity and content preference similarity correspond to different calculation methods.

(1) Natural attribute similarity: the user's natural attributes indicate the user's demographic characteristics, some natural attributes will intuitively affect the reader's reading behavior, define the function Attribute(i) to indicate the impact of user i's natural attribute information on the construction of the user's portrait:

$$Attribute(i) = \Psi(Gender(i), Type(i), Major(i), Department(i)) \quad (1)$$

Among them, Gender(i) indicates that the user's gender attribute will have an impact on the user's behavior in some scenarios; Type(i) indicates the student category, and Major(i) and Department(i) indicate the impact of the user's major and faculty on the construction of the user portrait, respectively. The user's natural attribute information

is expressed through mathematical notation, taking the user's gender as an example, the gender data is transformed into 0 and 1, defining the user's gender to take the value, and the formula is expressed as:

$$Gender(i)=\begin{cases}0, i="male"\\1, i="female"\end{cases} \tag{2}$$

The letters u and v represent two readers, respectively, and in performing the similarity calculation, the similarity between the two in the attribute dimension of gender is defined as 1, and 0 if they are not similar, and the gender similarity calculation formula is:

$$sim(Gender(u),Gender(v))=\begin{cases}0, Gender(u)\neq Gender(v)\\1, Gender(u)=Gender(v)\end{cases} \tag{3}$$

Similarly, data such as student category, major and faculty are converted into numbers to get their similarity formula. Based on this the similarity of natural attributes of both users u and v is obtained:

$$\begin{aligned}&sim(Attribute(u),Attribute(v))\\&=w_a\times sim(Gender(u),Gendere(v))\\&+w_b\times sim(Type(u),Type(v))\\&+w_c\times sim(Major(u),Major(v))\\&+w_d\times sim(Department(u),Department(v))\end{aligned} \tag{4}$$

where wi is the weight value of each attribute similarity, $w_a+w_b+w_c+w_d=1$

(2) Content preference similarity: in this paper, the book subject word is used as a label under the content preference dimension, and based on the number of times the book on the subject has been borrowed, it is mapped into a vector with the corresponding book subject in the position of 1, and the others in the position of 0, and the frequency of statistical appearances is used as the reading tendency as well as the weight of the label.The Pearson simple correlation coefficient is mostly used in text clustering of the vector space model as well as in the user recommendation, and it is used to measure the linear correlation between two variables to reflect the degree of similarity of two variables. It is used to measure the linear correlation between two variables, reflecting the degree of similarity between the two variables, so this paper adopts the Pearson correlation coefficient to calculate the similarity of content preference between two users, which is mathematically defined as:

$$sim(\Pr efer(u),\Pr efer(v))=r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{5}$$

where n is the number of samples, x_i and y_i are the variable values of the two variables, and the Pearson correlation coefficient, r, has a range of [-1,1], which indicates the correlation between users. The stronger the correlation, the greater the likelihood that a user's needs will change with the needs of other users, thus matching users with the same interests.

II. B. 3) User Profile Formation and Visualization

In this paper, we use K-means clustering method to divide the user group, so that there is a clear difference in each class of portraits, in order to better explain and describe the characteristics of the user [19].K-means clustering, also known as K-mean clustering, is the most widely used clustering algorithm, and the core idea is to calculate the mean center value of the samples contained in each subset of the clustering value, and this mean value will be used as a representative point of each cluster The central idea is to calculate the mean center value of the samples contained in each subset of cluster values, and this mean value will serve as the representative point of each cluster.

III. Multimodal data-enabled library personalized service system

III. A. UserCF Algorithm

III. A. 1) Feature Construction and K-Nearest Neighbor Searching

The classification form of CCTS is a tree structure with 22 major categories, expanding layer by layer from top to bottom, in order to avoid the data sparsity problem due to multiple readers' borrowing, the author takes the sum of

three values corresponding to the user's behavioral labels as the interest vector features, and chooses the second level in terms of the classification hierarchy, with a total of 222 subclasses, and assumes that the user's interest feature vector is $U=(u_1,u_2,u_3,u_4,\cdots,u_n)$, which is firstly normalized, and secondly by cosine similarity formula:

$$sim(u,v)=cos(u,v)=\frac{\sum_{i=1}^{n}u_i\times v_i}{\sqrt{\sum_{i=1}^{n}(u_i)^2}\times\sqrt{\sum_{i=1}^{n}(v_i)^2}} \tag{6}$$

The similarity between different users is computed in order to construct a similarity matrix between different users, from which the set of K neighboring users with maximum similarity to the target user, represented by U_k, is identified.

III. A. 2) Interest Extraction Factors

The UserCF algorithm needs to combine the ratings of other users with similar interests on an item to predict the user rating, and the high and low ratings are positively correlated with the user's interest. The user-item rating matrix is R=U×I. Since most libraries do not have detailed book rating data, this study adopts the college book recommendation algorithm based on interest degree and type factor, and selects the number of renewals and the duration of borrowing as the interest degree extraction factors, and in order to make the evaluation more accurate, the Douban book rating and the frequency of e-book downloads are included in the interest degree extraction factors to comprehensively obtain the user's interest in the book. In order to make the evaluation more accurate, Douban book rating and e-book download frequency are also included in the interest degree extraction factor, so as to comprehensively obtain the user's interest in the book.

III. A. 3) Calculation of interest level

Interest degree calculation to be accurate need to take into account the relevant parameters, the first parameter to consider is the borrowing time, borrowing time and user interest is usually a positive relationship. The percentage p of borrowing time of book i by user u is shown in equation (7):

$$p=\frac{T_a(u,i)-T_b(u,i)}{T_c} \tag{7}$$

where $T_a(u,i)$ is the point in time when the book is returned, $T_b(u,i)$ is the point in time when the book is borrowed, and T_c is the overdue expiration date specified by the library. The p-value is mapped into five interest degree values, and the interest degree formula (8) is shown:

$$pref_1(u,i)=\begin{cases}1,0\le p<0.25\\2,0.25\le p<0.5\\3,0.5\le p<0.75\\4,0.75\le p<1\\5,\text{Renewal times}\ge 1\end{cases} \tag{8}$$

The second is e-book downloads. In order to meet the rising demand for digital reading, the library will provide users with easily accessible e-books. Users will download the e-books after they have a strong interest in reading, and their interest score can be set relatively high, as shown in (9):

$$pref_2(u,i)=\begin{cases}3,t=1\\4,t=2\\5,t\ge 3\end{cases} \tag{9}$$

Once again, there is the Douban rating. Whether it is the borrowing time or e-book download, there is more or less uncertainty, so in order to make the UserCF algorithm more accurate, this study introduces the Douban rating, Douban reading within the user's comments on the book and star ratings are relatively more objective and fair, so the Douban rating of the book as $pref_3(u,i)$, the final user's comprehensive average interest value as shown in Equation (10):

$$pref(u,i) = \frac{pref_1(u,i) + pref_2(u,i) + pref_3(u,i)}{3} \tag{10}$$

The user reading interest level is also calculated based on UserCF algorithm as shown in equation (11).

$$pref(u,i) = \frac{\sum_{v \in u_k} pref_1(u,i) \times sim(u,v)}{\sum_{v \in u_k} sim(u,v)} \tag{11}$$

III. B. Cold start problems

Cold start includes user cold start and item cold start. Among them, the problem to be solved by user cold start is how to recommend books for new users in time, and the problem to be solved by item cold start is how to recommend new books for users in the first time. User cold start can calculate the similarity between different users based on the user's natural attributes and recommend books borrowed by other users with high similarity to the target user in real time. The most direct way of item cold start is to randomly display new books, but it is difficult to match the randomly displayed new books with the readers' needs, and the CB algorithm can properly solve this problem, the specific operation steps are:Construct a feature vector for the new book→Extract the user's interest feature vector→Calculate the similarity between the new book's feature vector and the user's interest feature vector, and recommend the book to the target user in time if the similarity degree is high.

III. C. Multimodal data-enabled smart library technology architecture construction

III. C. 1) Data acquisition layer

A multimodal data-enabled smart library technology architecture is constructed from four levels as shown in Figure 1.

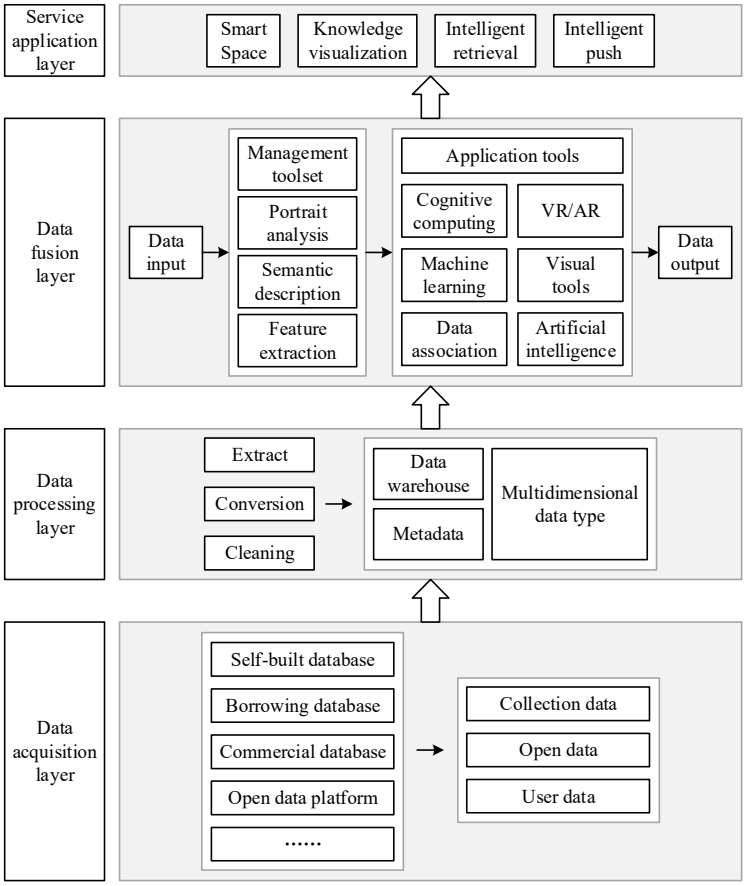

Figure 1: Intelligent library technical architecture of multi-modal data

Self-built databases, lending databases, commercial databases, open data platforms, etc. are important sources of multimodal data for smart libraries, and smart libraries mainly collect data from these databases in formats such as text, audio, and images. Due to the huge amount of data in the data collection layer, smart libraries should pay attention to the protection of user privacy and the security of collection resources in the process of data collection, strengthen data supervision, and discover security loopholes in time to prevent data leakage.

III. C. 2) Data processing layer

The role of the data processing layer is to extract, convert and clean the multimodal data, unify the data content and structure, and lay the foundation for subsequent data application. Data processing is a key step for smart libraries to improve their service performance. Smart libraries can extract data from multimodal data sets, establish unified data standards, convert and clean data formats, remove duplicated and redundant content, and store the processed data in the data warehouse.

III. C. 3) Data fusion layer

The library fuses, correlates and matches data with the help of data analysis and correlation tools. On the one hand, intelligent libraries can make use of data mining, deep learning and cognitive computing and other technologies to integrate information and deduce inferences on collection resources and open data resources to provide users with intelligent, precise and personalized information services; on the other hand, intelligent libraries can draw user profiles, dig deep into the characteristics of the user and perceive user needs, and establish the correlation between information resources and user needs with the help of data association and matching tools, laying the foundation for the development of subsequent services. On the other hand, smart libraries can draw user profiles, deeply explore user characteristics and perceive user needs, and establish the relationship between information resources and user needs with the help of data correlation and matching tools, laying the foundation for subsequent services.

III. C. 4) Service application layer

The service application layer is the top interactive layer of the multimodal data-enabled smart library technical architecture. The smart library focuses on smart services, which are concentrated in interactive interconnection, visualization services, knowledge value-added, immersive experience, etc., i.e., providing users with intelligent push, knowledge visualization, intelligent retrieval, and smart space services. The realization of multimodal data-enabled smart libraries relies on the application of information technology, and the advancement, stability, security, and reliability of information technology, professional librarians' expertise, ability to process multimodal data, and proficiency in and sensitivity to emerging technologies all affect the level of wisdom of smart libraries.

IV. Multimodal data fusion for personalized library services

IV. A. Visualization of library user profiling results

IV. A. 1) Data selection and processing

In this paper, the user lending data of a school self-service library from 2019--2023 is selected as the source of analysis, and after screening missing data and abnormal data, 658 valid user data are finally retained for a total of 2,894 lending data.

Combined with the actual borrowing situation of the self-service library, a total of 25 reading themes were categorized, including poetry and prose, literature (fairy tales and fables, sociology, etc.), history, music, travel, food, art, life (parenting, sports and fitness, attitude toward life, etc.), reasoning and suspense, gender, emotions, inspirational, healing, sci-fi, magical, war, novels, growth, psychology, science, popular science, teaching and learning, economics, faith and miscellaneous articles, Philosophy. Each topic word is used as a label under the dimension of user behavioral information, and the type and frequency of books borrowed by users are counted, each book type has a corresponding topic word, and the user's interest preference can be seen from the statistical frequency, while the frequency is set as the weight of the label.

In order to facilitate the processing of data, this paper encodes the acquired user information and the frequency of occurrence of content topics, as shown in Table 1. Excluding the user's demographic characteristics variables, 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, and the results are shown in Table 2. From Table 2, it can be seen that the KMO value and Bartlett's spherical test of the variables are ideal, the KMO value is 0785, which is greater than 0.7, and the level of significance is $P<0.05$, and there is a meaningful relationship between the 25 variables and is suitable for doing factor analysis.

Table 1: User data processing(part)

Coding	Subject word coding							
User	Prose	Literature	History	Music	Travel	Delicacies	Art	Life
1	1	0	1	0	0	2	0	1
2	0	0	0	0	0	0	0	0
3	0	1	0	0	2	0	1	0
4	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0
6	1	0	0	0	0	0	2	0
7	0	0	0	0	0	0	0	0
8	0	1	0	0	0	0	0	0
9	0	0	1	0	1	0	2	0
10	0	1	0	0	0	0	0	0

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

The cayser-meyer-olkin metric of the sampling is sufficient		0.785
Bartlett's spherical test	Approximate card	6683.142
	freedom	300
	Significant value	0.001

IV. A. 2) Label Extraction and Classification

The differences in user profiles lie in the differences in labels, and in order to obtain the types of labels with different characteristics, the 25 variables in the dimension of user behavioral information are subjected to dimensionality reduction processing to extract the characteristic factors. In SPSS software, the first standardization 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, as shown in Table 3 and Table 4. As seen in Table 3, the specified extraction of seven male factors, a total of 62.612% of the variance of the original variables explained, that is, the rotation of the sum of squares loaded, the effect of factor analysis is more satisfactory. In Table 4, the variables in the table were ranked according to the size of the loadings, and the rotated factor loadings were obviously polarized to the 0 and 1 levels, which is easy to see that the variables belong to the category of the male factors. Combining the factors in the above two tables and considering the principle of public factor selection, seven public factors will be selected as the classification of labels in this paper.

Table 3: Variable factor analysis explains total variance

Component	Initial eigenvalue			Extract the sum of squares and load			Rotate the squares and load		
	Total	Variance %	Cumulative %	Total	Variance %	Cumulative %	Total	Variance %	Cumulative %
1	5.695	22.78	22.78	5.695	22.78	22.78	4.534	18.136	18.136
2	2.592	10.368	33.148	2.592	10.368	33.148	3.543	14.172	32.308
3	2.137	8.548	41.696	2.137	8.548	41.696	2.415	9.66	41.968
4	1.524	6.096	47.792	1.524	6.096	47.792	1.179	4.716	46.684
5	1.453	5.812	53.604	1.453	5.812	53.604	1.873	7.492	54.176
6	1.242	4.968	58.572	1.242	4.968	58.572	1.214	4.856	59.032
7	1.01	4.04	62.612	1.01	4.04	62.612	0.895	3.58	62.612
8	0.969	3.876	66.488						
9	0.924	3.696	70.184						
10	0.892	3.568	73.752						
11	0.822	3.288	77.04						
12	0.785	3.14	80.18						
13	0.661	2.644	82.824						
14	0.575	2.3	85.124						
15	0.49	1.96	87.084						
16	0.463	1.852	88.936						
17	0.442	1.768	90.704						
18	0.431	1.724	92.428						
19	0.365	1.46	93.888						
20	0.337	1.348	95.236						
21	0.327	1.308	96.544						
22	0.264	1.056	97.6						
23	0.236	0.944	98.544						
24	0.203	0.812	99.356						
25	0.161	0.644	100						

Table 4: Rotating component matrix

	Component						
	1	2	3	4	5	6	7
War	0.782	0.026	0.002	0.178	-0.024	-0.024	-0.006
Inspiring	0.739	-0.07	0.169	0.273	-0.105	0.108	-0.062
Travel	0.733	0.181	-0.028	0.217	0.08	0.093	0.147
Music	0.716	-0.005	-0.05	-0.115	0.199	0.011	0.178
Delicacies	0.633	0.204	-0.021	-0.142	0.454	-0.023	-0.019
Literature	0.619	0.067	0.017	0.068	0.49	0.169	-0.046
Popularization	0.332	0.214	0.195	-0.037	-0.206	0.208	0.122
Science fiction	0.063	0.864	-0.005	0.226	-0.068	0.034	0.151
Inference suspense	0.12	0.787	0.061	-0.019	0.383	0.018	0.091
Grow	0.045	0.723	-0.038	0.168	-0.063	0.104	0.053
Magic	0.027	0.679	0.474	0.076	-0.069	0.178	0.116
Miscellaneous	-0.023	0.006	0.845	0.055	-0.028	-0.039	-0.03
Teaching and auxiliary	0.029	0	0.839	0.023	-0.014	-0.042	0.001
Affections	0.011	0.184	0.651	-0.196	0.333	0.172	-0.072
Sexes	0.012	-0.041	0.128	0.629	0.248	-0.13	0.261
Philosophy	0.049	0.363	-0.021	0.597	-0.012	0.092	-0.127
Cure	0.215	0.005	0.458	0.535	-0.013	0.01	0.194
Art	0.142	0.138	-0.05	0.535	0.056	0.079	0.101
Faith	0.085	0.238	-0.037	0.508	0.185	0.3	-0.143
History	0.043	-0.087	0.1	0.194	0.746	0.017	0.104
Prose	0.413	0.161	0.014	0.303	0.629	-0.042	0.116
Economy	0.015	0.045	0.032	0.026	-0.02	0.844	0.062
Life	0.141	0.164	0.006	0.117	0.044	0.76	0.168
Novel	0.024	0.107	-0.037	0.055	0.11	0.047	0.689
Mind	0.102	0.105	0.017	0.065	-0.02	0.126	0.644

According to the loading coefficients of each variable on the factors in Table 4, the label categorization feature factors of user profiles are described as follows:

Tag 1: The male factor of this tag is named “Leisure”. This type of factor is related to the topics of war, inspiration, travel, music, food, literature (fairy tales, fables, sociology), and science.

Tag 2: The common factor of this tag is named “Exploration”. These factors are related to the themes of science fiction, mystery, coming of age, and magic.

Tag 3: The metric of this tag is named “Reflective”. This factor is related to the themes of Miscellaneous, Teaching and Learning, and Emotional.

Tag 4: The male factor of this tag is named “Literary”. This factor is 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 factor 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.

IV. A. 3) Visual presentation of results

After obtaining the above seven label common factors, the common factors were used as the variables of cluster analysis, and the K-means clustering algorithm was selected to cluster all user samples, so as to establish the number of user portraits. The K-means clustering method can be used to manage users, which can better explain and describe the characteristics of users, and aggregate similar samples to form different classes or clusters in the case of a given taxonomic group K($K \ll n$), so as to achieve similarity within clusters and differences between clusters. In this paper, the number of clusters is set to 4-7 categories, and the number of clusters is 4, 5, 6, and 7 to compare the size of the difference of F-value, as shown in Table 5, it can be seen that when the number of clusters is 5, the significant value of "literature and art" is 0.433>0.05, so it is not used. When the number of clusters is 6 and 7, only

one user exists in each of the two categories, so it is not taken. When the number of clusters is 4, the Sig values of each common factor are significant, and the F-value of each common factor of the clustering is obviously different, so the number of clusters is set to 4.

Table 5: Comparison of differences under different clustering conditions

Cluster variable	Clustering number =4		Clustering number =5		Clustering number =6		Clustering number =7	
	F	Sig	F	Sig	F	Sig	F	Sig
Leisure class	287.039	0	230.322	0	215.413	0	190.876	0
Exploration class	98.828	0	8.463	0	7.535	0	187.547	0
Thinking class	66.704	0	199.063	0	129.21	0	120.376	0
Arts and arts	10.201	0	5.834	0.433	310.422	0	245.465	0
Literary history	62.925	0	370.615	0	289.359	0	243.887	0
Social class	256.164	0	153.104	0	169.799	0	142.388	0
Extended class	8.378	0	9.548	0.001	5.057	0	16.541	0

When the number of selected clusters is 4, the mean center value of each male factor in each class of portraits is calculated, as shown in Table 6, there is a significant variability in different eigenfactors in each class of portraits.

Table 6: The mean values of the four user profiles

User portrait classification	Leisure class	Exploration class	Thinking class	Arts and arts	Literary history	Social class	Extended class
Type 1	0.691	-0.055	0.463	1.112	0.451	9.104	0.401
Type 2	-0.086	-0.093	-0.091	-0.029	-0.089	-0.065	-0.037
Type 3	0.187	3.179	2.637	0.531	2.567	0.012	0.718
Type 4	8.579	-0.823	-0.037	0.665	-0.199	-0.714	1.092

Taking objective factors and subjective factors as the abscissa axis, and ability-oriented and preference-oriented as the vertical axis, the four types of user portraits obtained from the analysis were divided. This is shown in Figure 2. In the figure, library users are divided into four user types: "pragmatic", "youthful", "recreational", and "curious".

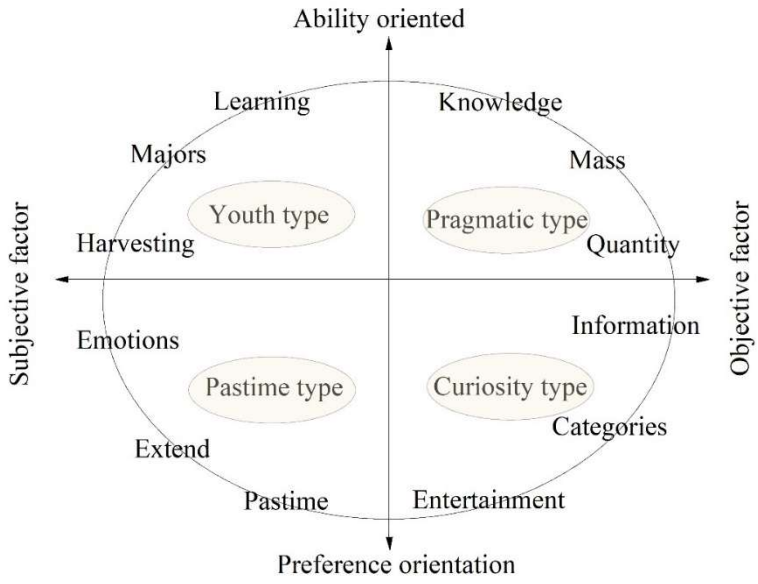

Figure 2: Classification of user profile types in Self-service library

IV. B. Digital Library Precision Recommendation Service Realization

IV. B. 1) Precision recommendation service realization process

(1) Construct user-book category matrix: according to the user, book subject, the number of times of borrowing to construct the user-book category matrix. First of all, the construction of the user U on the book I borrowing data

matrix, scoring the value of borrowing with 1, not borrowed with 0, as shown in Table 7. Then use K-means clustering algorithm on the user reading preference data clustering calculation, the user according to the reading preference attribute is divided into four categories, combined with the summary of the user's book borrowing data, for the like the same category of the user between the addition of 1 to construct the user-book category scoring matrix R shown in Table 8.

Table 7: User - book borrowing matrix(portion)

Book / User	I1	I2	I3	I4	I5
U1	1	0	0	1	0
U2	1	1	0	1	0
U3	1	0	0	1	1
U4	0	1	1	0	0
U5	1	0	0	1	0

Table 8: User - book category matrix(portion)

Categories / User	I1	I2	I3	I4	I5
U1	0	1	0	0	0
U2	2	2	1	0	0
U3	1	0	0	0	0
U4	1	0	1	3	3
U5	0	0	0	2	0

(2) Calculate nearest neighbors: the calculation of nearest neighbors is an important step in recommendation, and the set of nearest neighbors is generally determined by applying Top-N or threshold. In this paper, we choose to use the Top-N method to obtain, that is, in generating the set of nearest neighbor users of the target user, the set of nearest neighbor users is sorted by the similarity with the target user, and the top N users with the highest similarity are obtained. Sim(UI, U2) is used to represent the similarity between user U1 and user U2, $K(u_1)$ is the topic preference interest aggregation of user U_1, and $K(u_2)$ is the topic interest preference aggregation of user U_2. Based on the user-one-book category matrix constructed in the previous step, the cosine similarity is utilized to calculate the similarity between two users, and obtain the user ranking with higher similarity.

$$Sim(U1,U2) = \frac{|K(u_1) \cap K(u_2)|}{\sqrt{|K(u_1)| \times |K(u_2)|}} \tag{12}$$

(3) produce recommendations: for the target user recommended before the previous step in the matrix found with the target user U most similar to the M users, expressed in S (u, i), S in the user's favorite book theme i are listed, and eliminate the target user interested in the theme, for each listed theme i, are quantified using the formula to calculate the degree of interest of the user on the book theme i, will be sorted in the front of the theme of the book recommended to the target user. Where P_{ui} denotes the rating prediction of book i by user UI, R_u denotes the mean value of book ratings of user U1, and Sim denotes the similarity between users U1 and U2.

$$P_{ul} = R_u + \frac{\sum_{V=1}^{N} Sim(U1,U2) * (P_{vi} - R_u)}{\sum_{V=1}^{M} Sim(U1,U2)} \tag{13}$$

IV. B. 2) Precision recommendation result analysis

In order to verify the accuracy of the recommendation results, the accuracy rate, the recall rate and the composite value F, which are commonly used to evaluate the effectiveness of recommendations, are used to evaluate the recommendation results. Among them, the accuracy rate refers to the proportion of the number of hit recommendations to the total number of recommendations, the recall rate refers to the proportion of the number of recommended hit items to the theoretical number of recommended hits, and the comprehensive index F value is the weighted reconciliation average of the accuracy rate and the recall rate. In this paper, collaborative filtering

recommendation using fusion user profiles is compared with traditional filtering algorithms, and the accuracy rate, recall rate, and F value of comprehensive indexes are calculated under different numbers of recommendations, and the results are shown in Figures 3, 4, and 5, respectively. The results show that when the number of recommendations rises from 3 to 18, the accuracy rate of the two methods of recommendation decreases sequentially, while the recall rate and F-value gradually rise, when the number of recommendations is the same, the hybrid recommendation with fused user image is higher than the traditional collaborative filtering recommendation in the recall rate and F-value, and the recommendation effect is good.

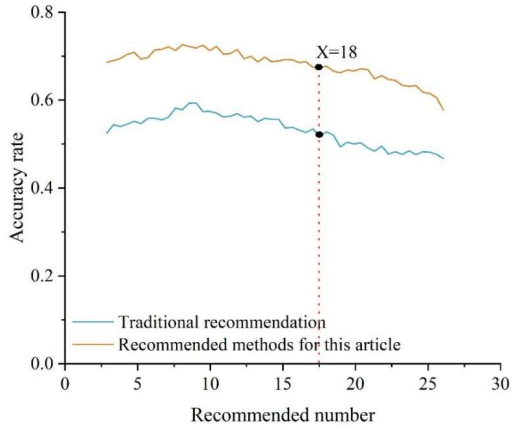

Figure 3: Comparison of accuracy of different recommendation Numbers

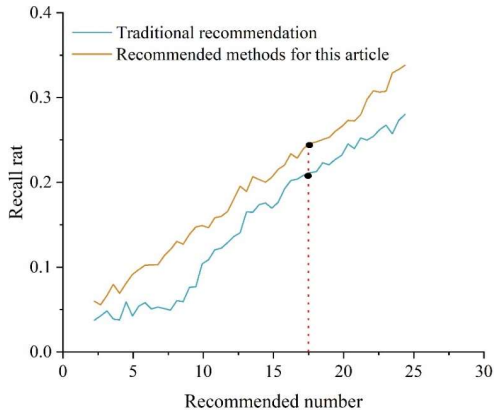

Figure 4: The recall rate was compared with different recommendations

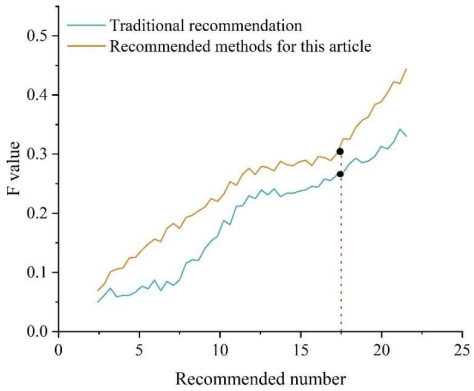

Figure 5: Comparison of f values in different recommendation Numbers

V. Conclusion

In this study, user portraits are applied to the construction of library personalized service system, and a hybrid recommendation algorithm is applied to accurately outline users' reading behaviors and reading tendencies. The K-

means clustering method was used to classify library users, and four user types were formed: "pragmatic", "youthful", "recreational" and "curious", which was convenient for libraries to accurately grasp the basic characteristics and reading needs of users, and then provide them with personalized book recommendation services. When the number of recommendations increased from 3 to 18, the accuracy of the two methods decreased sequentially, while the recall rate and F value gradually increased.

References

[1] Yaqin, M. A. (2022). Strategy of library development towards digital library. Khatulistiwa: Jurnal Pendidikan Dan Sosial Humaniora, 2(2), 52-69.

[2] Paskali, L., Ivanovic, L., Kapitsaki, G., Ivanovic, D., Surla, B. D., & Surla, D. (2021). Personalization of search results representation of a digital library. Information Technology and Libraries, 40(1).

[3] Fu, J., Yan, S., & Chen, X. (2025). A Mobile Technology-Based Framework for Digital Libraries: Bridging Accessibility and Personalized Learning. International Journal of Interactive Mobile Technologies, 19(4).

[4] Ifijeh, G., Idiegbeyan-Ose, J., Segun-Adeniran, C., & Ilogho, J. (2019). Disaster management in digital libraries: Issues and strategies in developing countries. In Emergency and Disaster Management: Concepts, Methodologies, Tools, and Applications (pp. 1556-1571). IGI Global.

[5] Khavidaki, S., Rezaei Sharifabadi, S., & Ghaebi, A. (2022). Status of Personalized Service in Digital Academic Libraries in Iran. Academic Librarianship and Information Research, 56(1), 21-42.

[6] Khavidaki, S., Sharifabadi, S. R., & Ghaebi, A. (2023). User Profile in Personalized Service of Academic Digital Libraries: A Delphi Study. Library & Information Sciences (1680-9637), 26(3).

[7] Zhuang, Y. (2021). Optimization of the personalized service system of university library based on internet of things technology. Wireless Communications and Mobile Computing, 2021(1), 5589505.

[8] Roy, S. G., Sutradhar, B., & Das, P. P. (2017). Large-scale metadata harvesting—tools, techniques and challenges: A case study of National Digital Library (NDL). World Digital Libraries-An international journal, 10(1), 1-10.

[9] Li, S., Jiao, F., Zhang, Y., & Xu, X. (2019). Problems and changes in digital libraries in the age of big data from the perspective of user services. The Journal of Academic Librarianship, 45(1), 22-30.

[10] He, M. (2023). Application of Big Data Analysis in Personalized Service Management of University Libraries. In Information and Knowledge Management.

[11] 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.

[12] Simović, A. (2018). A Big Data smart library recommender system for an educational institution. Library Hi Tech, 36(3), 498-523.

[13] Xiu, Y. (2025). Research on Resource Integration and Intelligent Recommendation Mechanism of Digital Library under Big Data Environment. J. COMBIN. MATH. COMBIN. COMPUT, 127, 1083-1100.

[14] Liu, Y., Yang, L., Sun, J., Jiang, Y., & Wang, J. (2018). Collaborative matrix factorization mechanism for group recommendation in big data-based library systems. Library Hi Tech, 36(3), 458-481.

[15] Adetayo, A. J., Adeniran, P. O., & Gbotosho, A. O. (2021). Augmenting traditional library services: Role of smart library technologies and big data. Library Philosophy and Practice, 6164, 1-15.

[16] 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.

[17] Tian, Y., Zheng, B., Wang, Y., Zhang, Y., & Wu, Q. (2019). College library personalized recommendation system based on hybrid recommendation algorithm. procedia cirp, 83, 490-494.

[18] Yi, K., Chen, T., & Cong, G. (2018). Library personalized recommendation service method based on improved association rules. Library Hi Tech, 36(3), 443-457.

[19] Liu Meidan. (2025). A Digital Classification Method of Rural Public Affairs Governance Based on K-means Clustering Algorithm. International Journal of High Speed Electronics and Systems,(prepublish).