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

Combinatorial Press

Research on Optimization Methods of Knowledge Discovery Algorithms for Library Data Mining

Xiuhua Wu[1], Guoqiang Sang[2,✉]

[1] *Library (Archives), Zhejiang College of Security Technology, Wenzhou, Zhejiang, 325000, China*

[2] *School of Physical Education and Health, Wenzhou University, Wenzhou, Zhejiang, 325000, China*

ABSTRACT

This study focuses on library data mining scenarios and proposes an optimization method for the deficiencies of existing knowledge discovery algorithms in terms of efficiency, accuracy and interpretability. The method first uses principal component analysis to downscale library high-dimensional data to extract the main features and improve the data mining efficiency. Then, the fuzzy clustering algorithm is used to cluster the dimensionality reduced data to more accurately identify the user groups, resource categories and other implicit knowledge. The clustering results are interpreted and analyzed to provide data support for knowledge discovery in library data mining. The algorithm in this paper demonstrates better performance in data dimensionality reduction at the level of memory usage as well as time consumption, and identifies three major components with cumulative contribution of more than 80%. In addition, the algorithm achieves an average purity of 95.45% for book data clustering and a clustering time consumption of 3.47s with a data stream of 300unit k, both of which are better than the comparison algorithms. The comprehensiveness weight of a university's book resources is 0.17, which is the highest performance, while the practicality and standardization are the next highest, 0.155 and 0.152, respectively. It can be seen from the clustering that the book category with the highest borrowing rate is science and technology, and the lowest one is literature, which reflects the user's demand for knowledge of a specific field.

Keywords: data mining, knowledge discovery algorithm, principal component analysis, fuzzy clustering, library

1. Introduction

In recent years, with the improvement of living standards year by year, Chinese residents have been able to better accept the “digital survival” as a way of life, due to the digital library has a fast information updating speed, large amount of information storage, not subject to the restrictions of

✉ Corresponding author.
E-mail address: sdhzsgq1978@126.com (G. Sang).
Received 25 February 2024; Revised 10 June 2024; Accepted 05 November 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-140

time and space, as well as occupying a small space and so on, so it is also more and more concerned about the people [1-4]. Although digital libraries do bring people a lot of convenience and convenience, but because of its information resources are very large and diverse, so people are also disturbed to a certain extent [5-7]. With the application of knowledge discovery algorithms in the field of digital libraries, it realizes personalized service for readers and changes passive service into active service, which has important theoretical and practical significance [8-10].

With the increasing variety of network information, knowledge discovery becomes more and more important in various industries. As a form of information processing, knowledge discovery is a research in the field of knowledge mining, which aims to discover meaningful structured or unstructured information from a large amount of unprocessed or processed information [11-14]. In library data mining, knowledge discovery can help users to obtain valuable knowledge from a large amount of information, and more easily realize the improvement of work efficiency, so as to enhance users' business performance [15-18]. With knowledge recognition, workers can keenly identify valuable information from meaningless noise, carry out in-depth research and mining in related fields, and effectively support the process of research and decision-making [19-22].

This paper collects information on book lending data from users of a university library and proposes a knowledge discovery optimization algorithm for active learning entropy sampling library data mining based on the combination of principal component analysis and fuzzy clustering. The book category variables are transformed into numerical variables by uniquely hot coding, and the deviations between the variables are eliminated by using principal component analysis. Data points with maximum information gain are obtained to improve learning efficiency through sample selection, model training, model prediction, and model updating. Combined with fuzzy clustering algorithm, the affiliation degree of each cluster of data points is calculated for category clustering. Through experiments, the main components of library knowledge discovery are identified, the algorithm's performance of high-dimensional data dimensionality reduction is verified, and multiple book cluster classes based on users' book borrowing rate are obtained to improve the accuracy and interpretability of knowledge discovery.

2. Research foundations and key methodologies

2.1. Data Mining and Knowledge Discovery Relationship Defined

2.1.1. Knowledge discovery and data mining. Knowledge discovery [23] is the extraction and refinement of new patterns from data sets. The scope of knowledge discovery is very wide it can be economic, industrial, agricultural, military, social, commercial, scientific data or satellite observed data. Data can be in the form of numbers, symbols, graphs, images, sounds, etc. Data organization also varies and can be structured, semi-structured or unstructured. The results of knowledge discovery can be expressed in various forms including rules, laws, scientific laws, equations, etc.

Currently, relational databases are widely used and have the advantages of unified organizational structure, integrated query language, equality between relations and attributes, and so on. Therefore, the research of knowledge discovery in databases (KDD) is very active.

Some scholars believe that data mining [24] is a step in the process of KDD, i.e., a specific process of applying algorithms on prepared data (after data selection, data preprocessing and data transformation), and KDD as a general term for the whole process of knowledge discovery, including data preparation, data mining, interpretation and evaluation of results. However, they are usually not strictly differentiated for the following reasons.

1) Data mining research has been not only limited to the database of data mining objects, including text, images, sound and other unstructured data formats in this sense that data mining than knowledge discovery or KDD closer to the meaning of the term itself.

2) Data mining is mainly popular in the field of statistics, data analysis, databases, and management

information systems. Knowledge discovery, on the other hand, is mainly popular in the artificial intelligence and machine learning communities. From the background of the emergence of knowledge discovery, it can be seen that knowledge discovery from the beginning is to solve the application problem, but only in the practical application of the extensive use and promotion can in turn promote the theoretical study of knowledge discovery.

2.1.2. Knowledge discovery mining process. The KDD process is shown in Fig. 1.The KDD process is a cyclic process of continuous interaction with the user. Generally speaking the process of discovering valuable knowledge from large source data is accomplished on the premise that the business requirements are already clear.

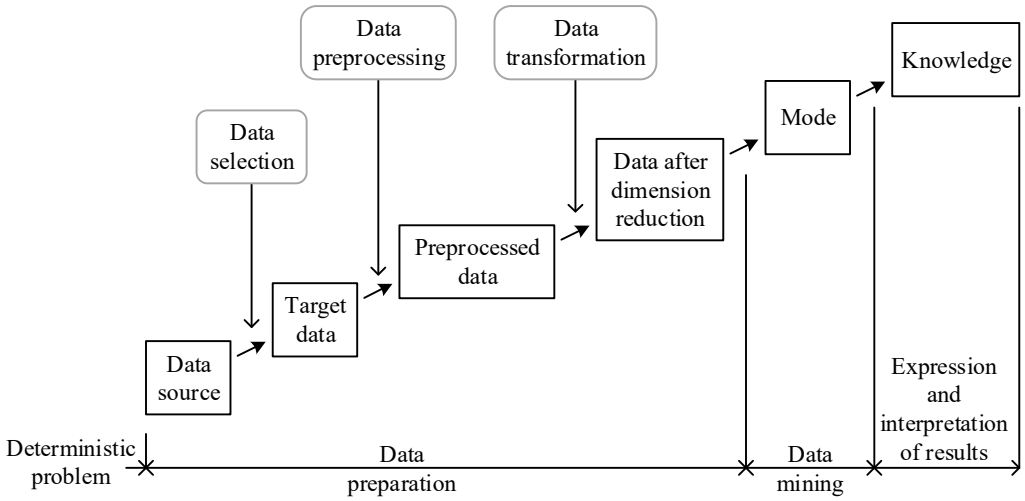

Fig. 1. KDD process

The work in the data preparation stage includes 3 aspects:

1) Data selection is mainly to determine the target data, that is, according to the user's needs from the original database to extract a set of data of interest and organize it into a suitable form of data organization for mining.

2) Data preprocessing, also called data cleaning, mainly includes the following work to be done: eliminating noisy data, the most discussed methods for dealing with noisy data are data smoothing techniques, deducing and calculating missing data, eliminating duplicate records, and completing data type conversion.

3) Data transformation mainly refers to the dimensionality reduction of data. Data mining phase is the knowledge discovery of prepared data sets using specific mining algorithms according to the task or purpose of mining. This knowledge is implicit, previously unknown, and potentially valuable for decision making extracted knowledge is represented in the form of concepts, rules, laws, and patterns.

The mined patterns are interpreted and evaluated to eliminate redundant or irrelevant patterns and present the results to the user. Data mining and KDD are closely related the general view is that data mining is a part of the KDD process but there are a few who believe that the two concepts are equivalent.

2.1.3. Deficiencies in library knowledge discovery. With the rapid development of information technology, libraries will accumulate more and more data, and the intelligent analysis and processing of massive data will be an important task in the information society. As a subcategory of knowledge discovery classification technology, although its research has made remarkable achievements and has been well applied in various fields, it still faces many problems that need to be solved. The main issues are the discovery of user-friendly knowledge, the handling of noisy data, the discovery of knowledge

based on specific data, the correctness and efficiency of classification algorithms, and so on. Therefore an optimization method for book data clustering based on active learning entropy sampling is proposed in the following section for solving the problems of knowledge discovery in library data mining.

2.2. Optimization method for book data clustering based on active learning entropy sampling

In order to further improve the accuracy and efficiency of book data clustering, this paper proposes a book data clustering optimization method based on active learning entropy sampling. The method alleviates the computational cost of data samples by invoking principal component analysis, which further reduces the error and uncertainty in the entropy sampling process. At the same time, the method also utilizes fuzzy clustering method for clustering, which can effectively solve the problem of data imbalance and noisy data.

2.2.1. Implementation of the Principal Component Analysis (PCA) method. In this paper, the borrowing data of book users were selected as the data for cluster analysis. Using Principal Component Analysis (PCA) [25] method for dimensionality reduction of book data, the first three principal components with cumulative contribution rate greater than 80% are selected for clustering, and less than 80% are not clustered, and the selected data are used as manual labeling samples for active learning, so as to reduce the cost spent on manual labeling.

Figure 2 illustrates the arithmetic flow of the PCA method. The method begins by normalizing the acquired book dataset. The mean of each variable is subtracted, which allows the data to be averaged to zero and helps to eliminate bias between variables. Each variable is divided by its standard deviation, which ensures that the variables are on the same scale and helps to eliminate scale differences between variables. For categorical variables, solo thermal coding or dummy variable coding is often required to convert them to numerical variables. The samples for principal component analysis were m $(x_1, x_2, ..., x_m)$ and each sample was described by n features. Where S denotes the standard deviation and $\tilde{X}_l$ denotes the mean. The formula for performing standardization is as follows:

$$S = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1}$$

$$\tilde{X}_l = \frac{x_i - \overline{x}_l}{s_i} \tag{2}$$

Next, the matrix of correlation coefficients is calculated, and r_{ij} is the correlation coefficient between the i nd indicator and the j rd indicator with the following formula:

$$r_{ij} = \frac{\sum_{k=1}^{n} x_{kl} \cdot x_{kj}}{n-1} \tag{3}$$

Next, this paper will calculate the eigenvalues and eigenvectors of the matrix and use these eigenvectors to construct m new indicator variable. Finally, the cumulative contribution rate of the eigenvalues will be calculated. Where b_j denotes the contribution rate and α_p denotes the cumulative contribution rate. The formula is as follows:

$$b_j = \frac{\lambda_j}{\sum_{k=1}^{m}\lambda_k} (j = 1, 2, \ldots, m) \tag{4}$$

$$\alpha_p = \frac{\sum_{k=1}^{p} \lambda_k}{\sum_{k=1}^{m} \lambda_k} \tag{5}$$

Finally, the samples with cumulative contribution greater than 80% were selected as sample labels for active learning manual labeling.

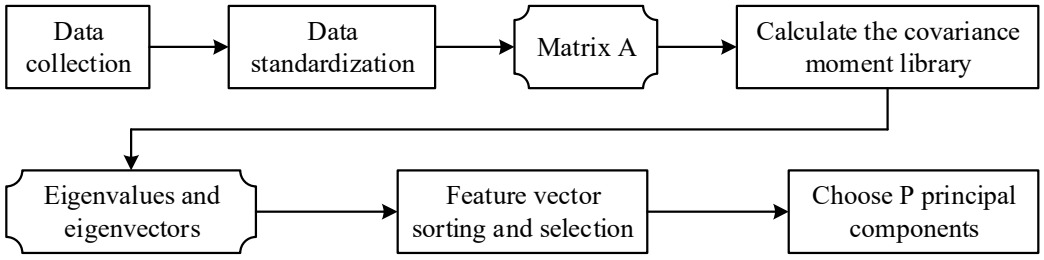

Fig. 2. PCA flowchart

2.2.2. Active Learning Models. Firstly, the samples obtained by principal component analysis method are used as the samples for active learning. Active learning [26] consists of the following processes, sample selection, model training, model prediction, and model updating, the active learning process is shown in Figure 3. It improves the learning efficiency and accuracy by allowing the model to autonomously select data points with maximum information gain. In active learning, the model actively selects new data samples for labeling based on the currently available training data and some selection criterion to obtain more information and better generalization ability. The entropy method is introduced in describing the uncertainty of a sample, entropy has a clear advantage in measuring the uncertainty of a sample, when x_H^* is larger, it means the uncertainty of the sample is larger. Therefore, those sample data with higher probability of uncertainty can be selected as the pending labeled data. Define entropy as:

$$x_H^* = argmax_x - \sum p_\theta(y_i \mid x) log p_\theta(y_i \mid x) \tag{6}$$

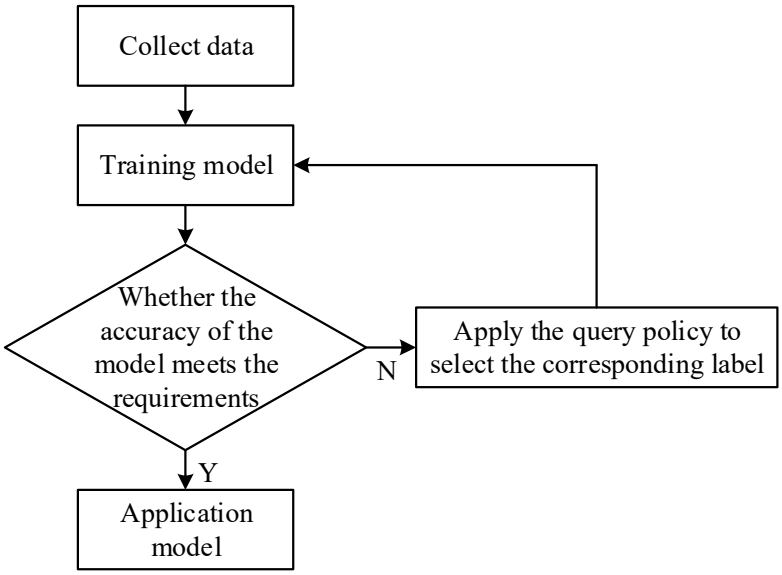

Fig. 3. Active learning flowchart

2.2.3. Clustering method implementation. FCM is a fuzzy based clustering algorithm [27] that provides more flexible clustering effect by calculating the degree of affiliation of each data point belonging to each cluster. First initialize the affiliation matrix, calculate the cluster centers and update the affiliation matrix. FCM clustering optimizes the degree to which each data point is assigned to each cluster center as an objective function. The FCM clustering process is shown in Fig. 4.The objective function of FCM is shown below:

$$J_m(U,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2} \tag{7}$$

Where U_{ij} refers to the affiliation value, the degree of affiliation of element j to category i, d_{ij} refers to the distance between element j and center point i, and the whole represents the sum of the weighted distances from each point to each category.

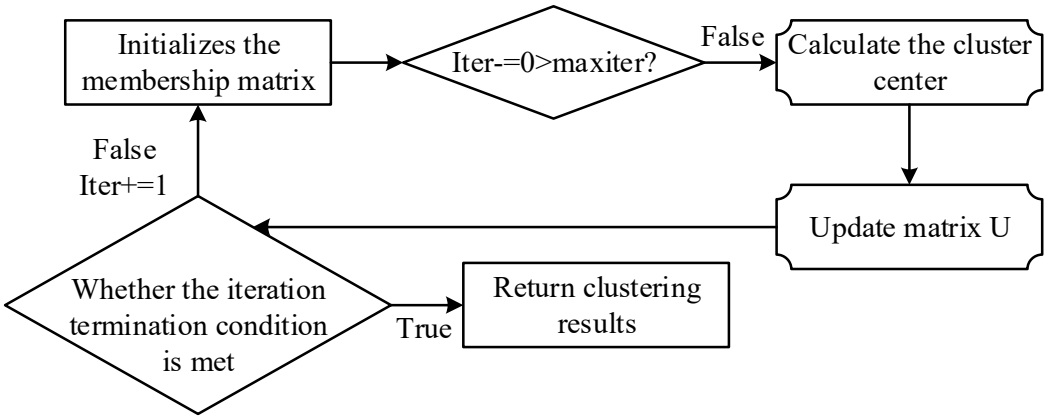

Fig. 4. Clustering flowchart

The purpose of clustering is to minimize the similarity between each cluster and maximize the similarity within clusters, so it is sufficient to make $J_m(U,V)$ obtain the minimum value. So the expression for the optimal solution:

$$\min(J_m(U,V)=\min(\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}d_{ij}^{2}) \tag{8}$$

Selection of the number of clusters c:

$$L_{(c)}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|v_i-\bar{x}\|^2/(c-1)}{\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|x_j-v_i\|^2/(n-c)} \tag{9}$$

Clustering Center:

$$p_i=\frac{\sum_{i=1}^{n}(u_{ij})^m x_i}{\sum_{i=1}^{n}(u_{ij})^m} \tag{10}$$

2.2.4. Data pre-processing. In this paper, the borrowing data of book users is selected as the data for cluster analysis. Since the initial data is usually missing or wrongly omitted, it is necessary to clean the data. For data cleaning, the following methods can be used: for data points that are high or low, they can be deleted. For data with more than 20% missing data points, they can also be deleted. For data

with no more than 20% missing data points, the average of the neighboring data can be used to supplement the data. If neighboring data is also missing, you can look backward for non-null data.

In order to improve the efficiency of the algorithm, feature extraction of the selected data is required. The value domain of the original data may have a large difference, if the original data is processed directly, it will make the data with large values and small values have a large difference and cannot be analyzed effectively. In this paper, the maximum value method is used for normalization to normalize the values to the [0,1] interval.

2.3. Applied Research on Optimization Methods for Knowledge Discovery

2.3.1. Purpose of the study. In the readership of university libraries, as different readers have different subject-specialized backgrounds and want to acquire knowledge, such as professors, lecturers, liberal arts students, science students, and other employees of the university, these readers have different differences in their needs. In order to provide library decision makers with strong support for the rational construction of the collection of literature, it is necessary to group different types of readers according to the differences in reader demand, use the appropriate model for each group to make predictions, and then compare the results of the demand prediction of various types of readers, so as to grasp the differences in the future demand of various types of readers.

2.3.2. Objects of study. The data source used in this research paper comes from the book loan circulation data and bibliographic search data of a university library. The total number of books borrowed and returned by a university library for the year 2024 is 2.16 million times, and the data of May 2024 is chosen as the experimental data in this paper.

2.3.3. Raw Data Mining and Description:

1) Data records in library bibliographic search system. There are mainly the following four kinds of records: records of readers' borrowed book information, records of readers' borrowed book history information, records of readers' retrieval history, and records of readers' customized literature categories of interest.

2) Data Source. Data source: book circulation log and bibliographic search log of a university library:

Owner: A university library.

Organization responsible for maintaining the data: Technical Services Department of a university library.

Data storage form: oracle database.

Number of fields, records: 8 fields, 54,618 search history records.

Timeframe: May 1, 2024 to May 31, 2024

3) Data description. In order to be able to clearly describe the complex reader's bibliographic retrieval data in a bibliographic retrieval system, a formalized system that can express some kind of entity or information clearly is needed, as well as a number of rules describing the interrelationships and roles of the different components of the system, where the concept of ontology is introduced, and the reader's bibliographic retrieval data is described through a combination of semantics and numerical values.

As a semantic model in a bibliographic retrieval system, ontologies for describing readers' retrieval of complex data from multiple sources construct a basic body of knowledge in the field of library Web data mining by normalizing the concepts, terms, and interrelationships within the field.

3. Results and analysis

3.1. Library Knowledge Discovery Principal Component Loadings Calculation

In this section, SPSS software is used to perform principal component analysis on the collected user

book lending data. The initial eigenvalue mainly reflects how much the public factor contributes power to the original variable, which is the degree of explanation of the variation of the data by each principal component.

Table 1 shows the total variance statistics of the eight principal components of knowledge discovery for library data mining. For example, the initial eigenvalue in the first principal component of library knowledge discovery is: 6.163, and the cumulative is: 55.41%, the meaning is that the initial eigenvalue of the first principal component of library data mining knowledge discovery on library data mining knowledge discovery of all the indicators data in the data interpretation degree of 55.41%. The loaded sum of squares reflects the degree to which the public factor explains the variance of the original variable, which is the cumulative contribution. Initial eigenvalues greater than 80% were selected to determine the three principal components.

Table 1. The total variance statistics of 8 main components

Principal components	Initial eigenvalue			Sum of squares of extraction load		
	Total	Variance (%)	Cumulation (%)	Total	Variance (%)	Cumulation (%)
1	6.163	55.41	55.41	6.163	55.41	55.41
2	2.016	16.28	71.69	2.016	16.28	71.69
3	1.637	10.75	82.44	1.637	10.75	82.44
4	0.824	6.73	89.17			
5	0.435	4.28	93.45			
6	0.217	3.19	96.64			
7	0.093	1.93	98.57			
8	0.046	1.43	100			

A fragmentation plot is a graphical method used to select the number of principal components, showing the eigenvalue or variance contribution of each principal component in descending order. Figure 5 shows the eigenvalue fragmentation plot for 8 principal components, where the horizontal axis is the serial number of the principal components and the vertical axis is the eigenvalue or variance contribution rate. From the figure, it can be seen that the library data mining knowledge found that there are three eigenvalues more than 1, which can be used as principal components. The inflection point in the figure is the point at which the eigenvalue or variance contribution rate changes significantly, that is, from a steeper decline to a smoother decline. Principal components before the inflection point have larger eigenvalues or variance contributions, indicating that they explain most of the variation in the data, and principal components after the inflection point have smaller eigenvalues or variance contributions, indicating that they explain less of the variation in the data. Thus, the inflection point can be used as a natural cut-off point to separate the important principal components from the unimportant ones. In general, the principal components prior to the inflection point are chosen as the final principal components to be retained in order to achieve a balance between reducing the dimensionality of the data and retaining the information in the data.

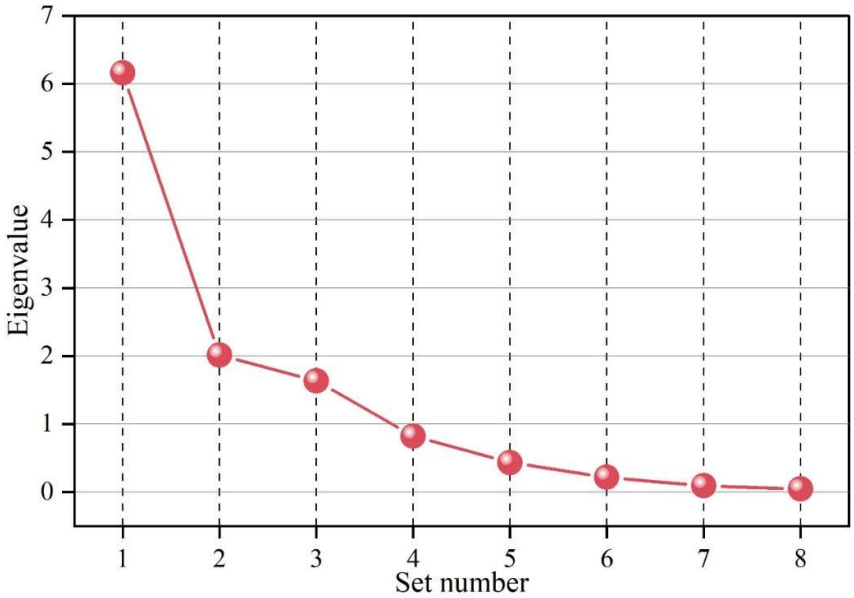

Fig. 5. The characteristics of 8 main components are found

In addition, this section establishes library knowledge discovery evaluation indexes centered on book lending rate, electronic resource downloads, seat utilization rate, equipment utilization rate, and

the number of library services by reviewing related literature. In order to evaluate the contribution of the three main components to the library data mining knowledge discovery indexes, the contribution coefficients of the three main components to the evaluation indexes were calculated as shown in Figure 6. From the figure, it can be seen that in the extraction of the first principal component, the coefficient of the number of library services is the highest, amounting to 0.772. The values of the number of electronic resource downloads and the book lending rate are higher, amounting to 0.769 and 0.65, respectively. Which indicates that these two indexes have a greater impact on the knowledge discovery of libraries. In the second principal component extraction, the number of library services has the highest value of 0.755, while the value of e-resource downloads in the third principal component extraction reaches 0.879, which is the highest performance. The above shows that the most illustrative directions for library data mining knowledge discovery found in the data space through principal component analysis are the number of library services, the number of e-resource downloads, and the book lending rate indicators with the highest distribution, which ultimately produce the highest value of contribution to the composite score of the principal components.

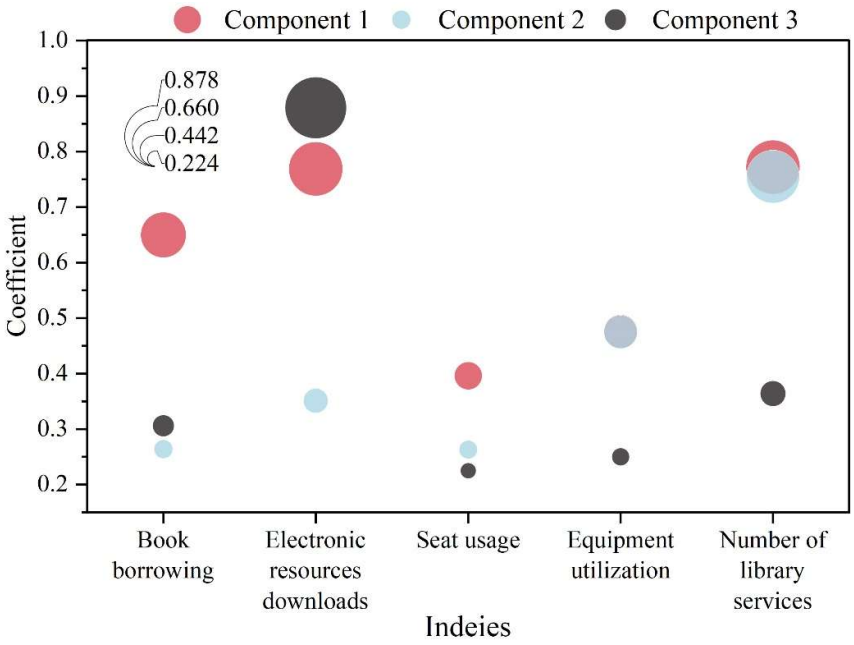

Fig. 6. The contribution coefficient of the main component to the evaluation index

3.2. Memory vs. time-consuming overhead for dimensionality reduction performance

There are some problems when applying PCA algorithm to dimensionality reduction of high dimensional feature data, such as high memory consumption and long time. When computing the attribute covariance matrix, the time and space complexity will increase dramatically with the increase in the number of dimensions. In order to the performance of clustering optimization algorithm based on active learning entropy sampling for memory consumption proposed in this paper, the memory occupied by the covariance matrix is calculated using Eq:

$$Mem(Cov) = \frac{N \times N \times 8}{1024^2} \tag{11}$$

First, this work evaluates the proposed algorithm in this paper against the traditional PCA algorithm, the Randomized SVD algorithm, and the Gaussian Random algorithm in terms of memory occupancy in the following test environment:

The overall test environment is a 4-core 16GB hardware, and the test of memory is performed by obtaining the RSS capacity corresponding to the process ID (this test value is slightly lower than the

actual value because the output time point cannot be sampled continuously, but it can be used as a side-by-side comparison of the methods). The test process used 25000 samples, respectively, from the number of dimensions 2500 to 25000 of the isotropic series for the test, the neighboring dimensions of the difference is 2500.

The results of different algorithm memory tests are shown in Fig. 7, from which it can be seen that with the increase in the number of dimensions of data attributes, the use of memory by all types of algorithms increases significantly. Among them, the traditional PCA consumes higher memory, the growth of Randomized SVD algorithm increases significantly when the number of dimensions exceeds 12500, and the memory consumption exceeds that of Gaussian Random algorithm, while the algorithms in this paper consume the lowest level of memory at all times. When the number of dimensions of data attributes reaches 25000, the memory consumption of PCA, Randomized SVD, and Gaussian Random algorithms reaches 5784MB, 4763MB, and 3261MB, respectively, while the memory consumption of the improved PCA proposed in this paper is 2670MB.It can be seen that, compared to the comparison algorithms, the proposed active learning-based entropy sampling clustering optimization algorithm has a much improved memory footprint.

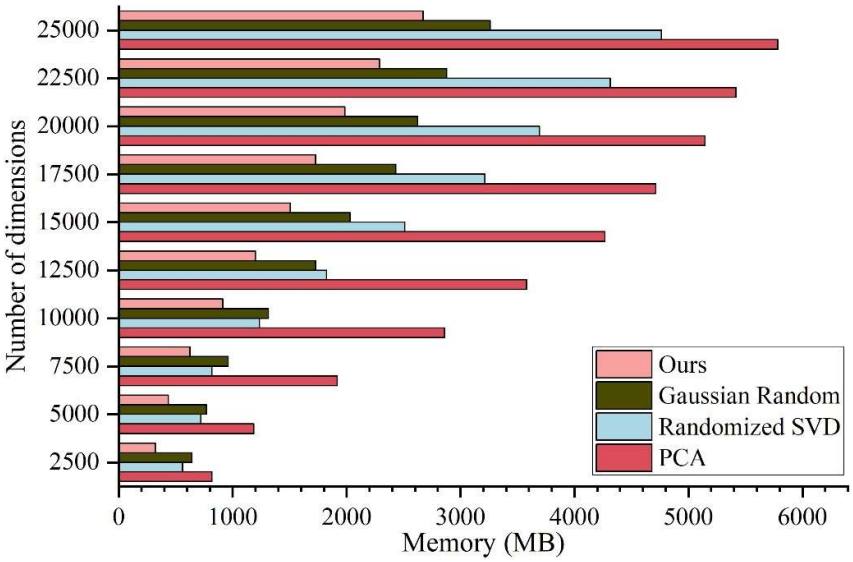

Fig. 7. Different algorithm memory test results

Secondly, this work evaluates the comparison of the algorithms in terms of time overhead and the results of the comparison of time overhead of the algorithms are shown in Table 2. The data in the table shows that the PCA algorithm performs the worst in terms of time consumption, and the best performer is the algorithm of this paper. The algorithm in this paper consumes the shortest time, when the number of dimensions reaches 25000, the comparison algorithms all consume more than 300 seconds, while the algorithm in this paper consumes only about 128s. This shows that the clustering optimization algorithm based on active learning entropy sampling proposed in this paper is greatly optimized in terms of time complexity. The algorithm performs dimensionality screening in practice, which is equivalent to the optimization of dimensionality reduction under the acceptance of a certain degree of information loss.

Table 2. The time overhead of the proposed algorithm is compared

Number of dimensions	Time consumption(s)			
	PCA	Randomized SVD	Gaussian Random	Ours
2500	78.34	64.76	69.29	46.38
5000	123.87	96.44	87.39	55.71
7500	187.50	132.92	123.87	69.29
10000	305.71	173.93	187.50	73.81
12500	387.73	219.46	228.51	78.34
15000	492.36	278.56	255.66	87.39
17500	606.05	333.14	274.04	105.77
20000	715.21	415.16	292.14	114.82
22500	774.32	487.84	328.62	123.87
25000	829.19	569.85	360.29	128.39

3.3. Clustering quality evaluation based on clustering purity and elapsed time

In this chapter, in order to test the overall performance of the clustering optimization algorithm based on active learning entropy sampling, clustering purity, and runtime evaluation metrics are used to test the applicability and effectiveness of the algorithm.

The experiment uses the real dataset of user book borrowing and lending (BLDS) in a university library collected above. The data has a total of 5436498 data records with 44-dimensional attributes. For the convenience of experimental operation, the experiment was tested using 20% of the data in the sample dataset. The program accesses the ASCII file stream in a sequential way to simulate the effect of data flow. In order to facilitate algorithmic calculations, all experimental data were normalized and the values of each dimension of the data were mapped between [0,1].

In the experiments, the clustering purity was used to evaluate the clustering quality. The higher the clustering purity, the more similar data in the cluster and the greater similarity between the data. The formula for clustering purity is as follows:

$$purity = (\sum_{i=1}^{K} \frac{\left|C_i^d\right|}{\left|C_i\right|}) / K * 100\% \tag{12}$$

Where K denotes the number of clusters in the clustering result, $\left|C_i^d\right|$ denotes the number of major class labels in the i rd cluster, and $\left|C_i\right|$ denotes the number of all data in the i th cluster.

The clustering quality of this paper's algorithm is measured experimentally using BLDS, and the algorithm is compared with the clustering quality of the classical data stream clustering algorithm (CluStream), the center-weight-based streaming data clustering algorithm (CW-Stream), and the density- and extended-network-based data streaming algorithm (DEGDS).

Fig. 8 depicts the clustering purity comparison results of the four algorithms. From the experimental results, it can be seen that the average clustering purity of this paper's algorithm can reach more than 95%, which is significantly better than that of CluStream, CW-Stream and DEGDS algorithms. The algorithm in this paper automatically divides the sample points into the clustering center, and the data points are more reasonably assigned to the micro-clusters, which reduces the error caused by grid division, and thus has a higher clustering purity.The CluStream algorithm is unable to form clusters of any shapes, and it cannot identify the outliers well, which reduces the accuracy of clustering.

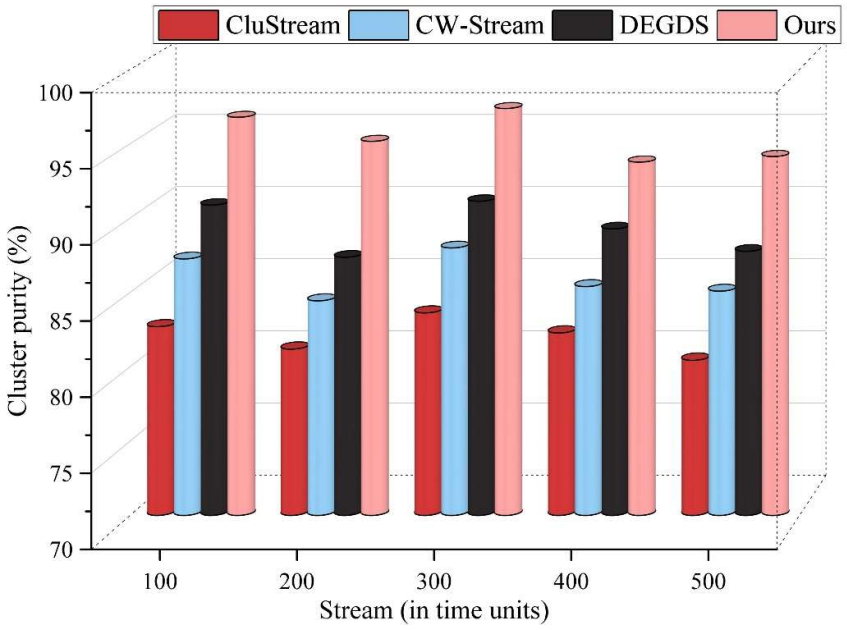

Fig. 8. The comparison of the clustering purity of four algorithms

The data stream clustering algorithm must have a high execution efficiency to keep up with the arrival rate of the data stream. To test the running speed of the algorithms, the experiments were conducted using the BLDS dataset as the execution efficiency. With the same size of each dataset, the running time of each algorithm is recorded and the average value is counted.

The results of the comparison of the running time of each algorithm are shown in Fig. 9. From the figure, it can be seen that when the test dataset is small, the execution time of the comparison algorithms is slower than that of the algorithms of this paper on a single node. This is because the parallel computing under the Spark framework partitions the streaming data, which interferes with the clustering results. However, with the gradual growth of the data stream, the running time of the comparison algorithm grows too fast, while the time consumption of this paper's algorithm is relatively flat. When the data flow exceeds 150unit k, the running time of this paper's algorithm is smaller than that of the comparison algorithm until 300unit k, which is shortened by 33.33% to 49.82% compared with the comparison algorithm. This is because the algorithm in this paper can distribute the data to different Worker nodes for parallel computation, and the intermediate results can be cached in memory, which saves a lot of time.

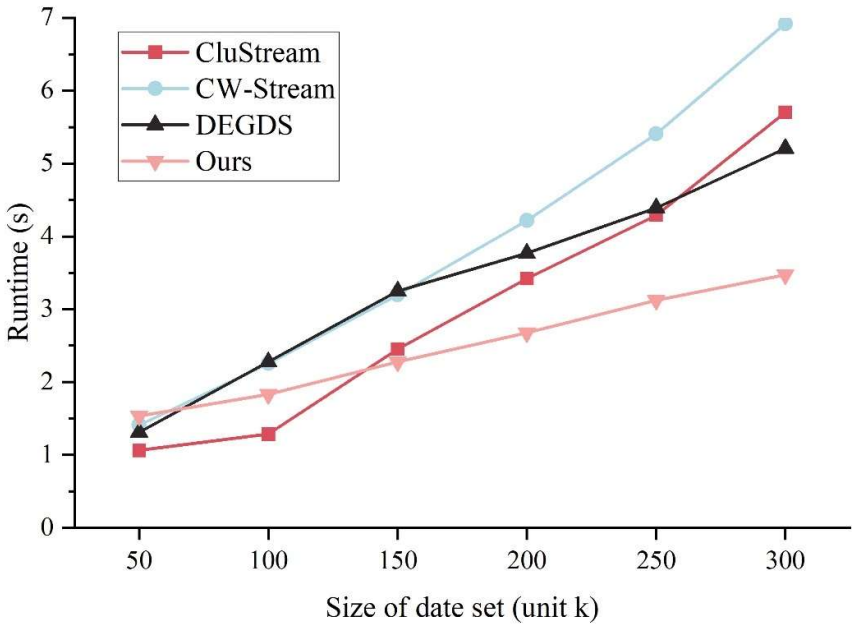

Fig. 9. The algorithm runs the result of the time comparison

3.4. Experimental Evaluation of Library Resource Integrity and Clustering

In order to further analyze the library knowledge discovery, the experiment is validated in the above dataset, and the evaluation indexes of resource integrity are established from the timeliness, comprehensiveness, authority, utility, distinctiveness, standardization, and continuity of the book resources in colleges and universities.

The information entropy, utility value and weight of the evaluation indexes of book resources are shown in Fig. 10, in which the comprehensiveness of book resources has the highest weight of 0.176, because the comprehensiveness of book resources best represents the integrity of resources. Firstly, comprehensive book resources can provide more knowledge information to meet the user's demand for multi-domain and multi-type information. This is followed by the practicality and standardization of book resources, which have weights of 0.155 and 0.152 respectively.

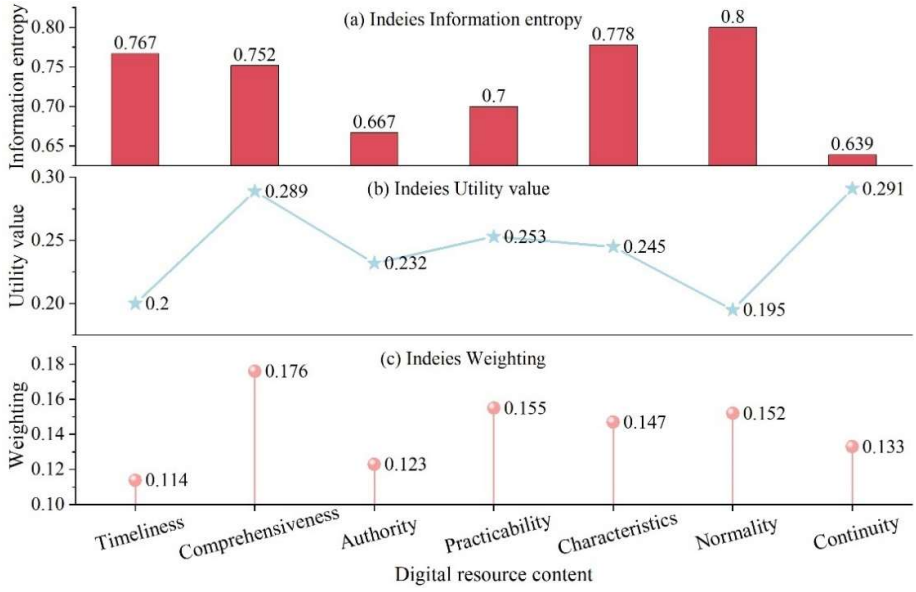

Fig. 10. Information entropy, utility value and weight of resource evaluation index

Applying the research model, the resource integrity of this library was scored based on the provided sample dataset in terms of three evaluation indicators: data coverage, data accuracy and data updating. Figure 11 shows the scoring results of the library resource integrity indicators. Firstly, the data coverage, whose scoring results are shown as the red line in the figure, it can be seen that the data coverage of this library reaches the highest score when the number of clusters is 5, and the average score of data coverage is 7.97. The data of the blue line in the figure shows that the data accuracy rate also reaches the highest value of 9.53 when the number of clusters is 5. The highest score of the data updating rate is less than 9, as can be seen from the black line in the figure. The average data coverage score of this library is 0.39 and 1.20 higher than the average scores of data accuracy and data updating, respectively. It can be seen that this library has wider data coverage and wider data accuracy, but the updating speed of the data needs to be further improved. In summary, it can be seen that the algorithm in this paper can effectively weight the library resource integrity score, provide a realistic basis for libraries to improve the comprehensiveness of resources and can promote the user's demand for knowledge discovery.

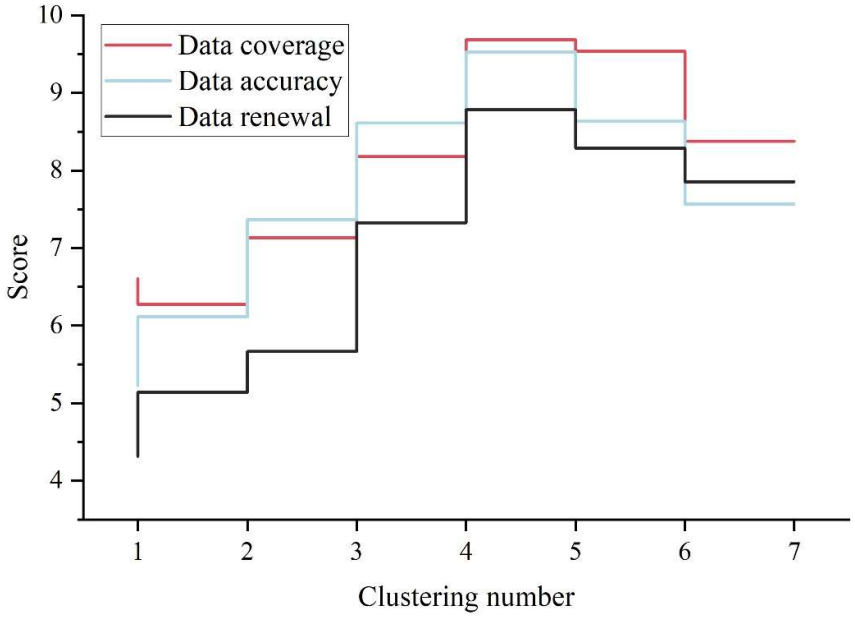

Fig. 11. Index scores of book resource integrity

For the above dataset, the clustering of book classification numbers based on users' book borrowing rate divides the book classification numbers into five major categories, which are distinctly different from each other in guiding book purchases. Figure 12 shows the results of book clustering analysis based on user book borrowing rate. Among them:

Cluster 1 (red): literature, the lowest borrowing rate. It can be found that there are many kinds of books, the number of categories reaches 28, mainly novels, essays, poems and so on. The utilization rate of these 28 types of books is low, that is, the borrowing of these 28 types of books is very low, but their collection is very high, and the library can appropriately reduce the purchase of these 28 types of books in terms of procurement.

Cluster 2 (blue): science and technology category, the borrowing rate is the top 10%. The number of categories can be found to be 8, mainly computer science, engineering technology, natural science and so on. The utilization rate of this type of books is high, that is, these 8 types of books borrowing volume is very high, but its collection of books is low, easy to cause “a book is difficult to find” embarrassing situation, the library in the procurement of these 8 types of books can be increased purchasing volume.

Cluster 3 (orange): social sciences, moderate borrowing rate. Mainly includes politics, economics,

law, education and other 29 kinds of books, the utilization rate of this type of books is generally normal, that is, the collection of these 29 kinds of books is very suitable for the current lending volume, the library in the procurement can continue to purchase according to previous standards, generally do not have to change the purchase of these 29 kinds of books.

Cluster 4 (gray): art, the cluster includes 19 kinds of books, the lending rate is low, but higher than cluster 1, the library can appropriately reduce the procurement of these books.

Cluster 5 (yellow): History, including world history, regional history, biographies, etc. The borrowing rate of this category is slightly lower than that of Cluster 2. Occasionally, there is an oversupply of books in this category, and the library may increase the purchase of books in this category as appropriate.

The borrowing rate reflects readers' interest in a specific topic or field, and a high borrowing rate implies a high demand for knowledge in that field. The borrowing rate can predict the trend of knowledge demand and help libraries optimize the allocation of resources.

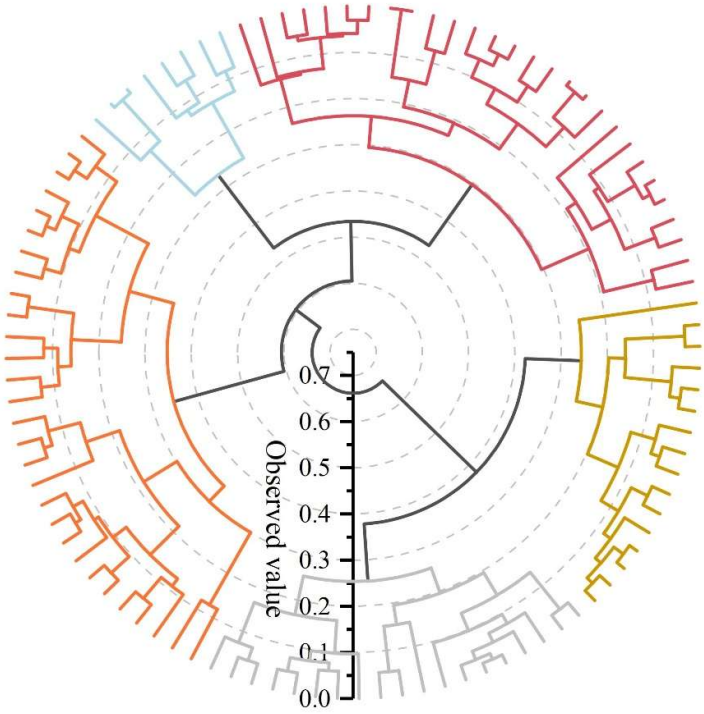

Fig. 12. The book borrowing rate is the result of the book clustering analysis

4. Conclusion

The article proposes a book clustering optimization method based on active learning entropy sampling. The method takes principal component analysis and fuzzy clustering algorithm as the core, and realizes the dimensionality reduction processing of massive library data through principal component analysis, and actively selects new data samples for labeling to obtain more information. Using fuzzy clustering algorithm, the clustering center of book data points is calculated and the affiliation matrix is updated to get accurate and reliable book clustering results. The optimization effect of knowledge discovery algorithm is studied by mining the borrowing data of users in a university library, combined with several experiments. The research results are obtained as follows: when dimensionality reduction is performed on data with a dimension of 25000, the memory occupation of this paper's algorithm is 2670MB, and the time consumption is only 128s, which is improved to different degrees compared with the comparison algorithm. The cumulative contribution of the three main components identified in the study is 82.44%, with the number of library services, the number of e-resources

downloaded, and the book borrowing rate contributing the most to the principal component score. The algorithm in this paper demonstrated better clustering quality, with an average cluster purity of more than 95%. In addition, the average value of book data coverage in a university library is 7.97, with a wide coverage of book knowledge. And according to the user book borrowing rate, the books in this library are categorized into literature, science and technology, social science, art and history. This study verifies the effectiveness and superiority of the proposed method through experiments, which is of great significance to improve the level of library knowledge services.

Funding

This work was supported by The National Social Science Fund of China (Project Number: 24BTY044).

About the Author

1) Xiuhua Wu was born in Wenzhou, Zhejiang, P.R. China, in 1982. She obtained a master's degree from East China Jiaotong University in China. Currently, she is working at Zhejiang College of Security Technology. Her main research direction is the construction of smart libraries and data governance.
2) Guoqiang Sang was born in Heze, Shandong, P.R. China, in 1978. He obtained a doctoral degree from Fujian Normal University. Currently, He is working at Wenzhou University. His main research directions are sports system engineering and adapted activity.

References

[1] Fox, E. A., & Ingram, W. A. (2020, August). Introduction to digital libraries. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020 (pp. 567-568).

[2] 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.

[3] Siregar, N. Z., Hasibuan, A., & Syam, A. M. (2024). The Influence of Digital Library Service Quality On Student Satisfaction. PERSPEKTIF: Journal of Social and Library Science, 2(2), 40-48.

[4] Kim, J. (2018). Evaluating author name disambiguation for digital libraries: A case of DBLP. Scientometrics, 116(3), 1867-1886.

[5] Xie, I., & Matusiak, K. (2016). Discover digital libraries: Theory and practice. Elsevier.

[6] Bisma, B., Nahida, N., & Loan, F. A. (2019). National digital library of India: an overview. Library Philosophy and Practice, 2019.

[7] Ali, S., Habes, M., Youssef, E., & Alodwan, M. (2021). A cross-sectional analysis of digital library acceptance, & dependency during Covid-19. International Journal of Computing and Digital System.

[8] Dong, N. (2024, March). Research on Knowledge Discovery Service in Digital Libraries Based on Deep Learning. In 2024 13th International Conference on Educational and Information Technology (ICEIT) (pp. 328-333). IEEE.

[9] Weng, C. H. (2016). Knowledge discovery of digital library subscription by RFC itemsets. The Electronic Library, 34(5), 772-788.

[10] Rahman, M. H., Zakaria, S., & Ahmad, A. (2021). Knowledge Discovery from the Digital Library's Contents: Bangladesh Perspective. In Towards Open and Trustworthy Digital Societies: 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Virtual Event, December 1–3, 2021, Proceedings 23 (pp. 34-42). Springer International Publishing.

[11] Lytras, M. D., Daniela, L., & Visvizi, A. (Eds.). (2018). Enhancing knowledge discovery and innovation in

the digital era. IGI Global.

[12] Sun, Z., & Stranieri, A. (2020). A knowledge discovery in the digital age. PNG UoT BAIS, 5(1), 1-11.

[13] Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. Social Science Research, 110, 102817.

[14] Bandaru, S., Ng, A. H., & Deb, K. (2017). Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey. Expert Systems with Applications, 70, 139-159.

[15] Liang, C. H. E. N. (2018). Research on Knowledge Discovery Service for Digital Library Collection Resources Based on Semantic Association. Journal of Library and Information Sciences in Agriculture, 30(3), 38.

[16] Azevedo, A. (2019). Data mining and knowledge discovery in databases. In Advanced methodologies and technologies in network architecture, mobile computing, and data analytics (pp. 502-514). IGI Global.

[17] Liu, J. (2021). Service Innovation of University Library in the Big Data Era. International Journal of Frontiers in Sociology, 2021, 3.0 (1.0).

[18] Chen, C. (2021). University library service based on big data. International Journal of Frontiers in Sociology, 3(1), 95-103.

[19] Sun, C. L., Shi, D., & Guo, D. M. (2014). Analysis on the library digitization construction under the big data era of university. Applied Mechanics and Materials, 687, 2656-2659.

[20] Zhao, L., Xing, L., & Liu, J. (2020). Strategies of information service based on intelligent library. In Frontier Computing: Theory, Technologies and Applications (FC 2019) 8 (pp. 1382-1386). Springer Singapore.

[21] Roy, B. K., Mukhopadhyay, P., & Biswas, A. (2022). Discovery layer in library retrieval: VuFind as an open source service for academic libraries in developing countries. Journal of Information Science Theory and Practice, 10(4), 3-22.

[22] Yu, J. (2021, December). Construction of Smart Library Based on the Big Data Mining and Knowledge Discovery. In 2021 Smart City Challenges & Outcomes for Urban Transformation (SCOUT) (pp. 57-61). IEEE.

[23] Eero Hyvönen. (2025). Serendipitous knowledge discovery on the Web of Wisdom based on searching and explaining interesting relations in knowledge graphs. Journal of Web Semantics100852-100852.

[24] Yue Yuan,Chengcheng Song,Kejun Zeng,Liying Gao,Yu Huang & Yixing Chen. (2025). An occupant-centric control case study based on internet of things and data mining for an office space. Journal of Building Engineering111925-111925.

[25] Samia A. Elseginy. (2025). Exploring binding mechanisms of omicron spike protein with dolutegravir and etravirine by molecular dynamics simulation, principal component analysis, and free binding energy calculations. Journal of Biomolecular Structure and Dynamics(4),2059-2072.

[26] Yong Jun Im,Tai Woo Chang,Sangun Park & Hyun Soo Kim. (2025). Object detection models and active learning for improvement of e-waste collection management systems in Korea. Waste Management379-389.

[27] Santosh Pattar,Veena Badiger & Yash Kangralkar. (2025). Context-aware IoT search engine through fuzzy clustering: Search space restructuring and query resolution mechanisms. Internet of Things101494-101494.