Application of FCM Clustering Algorithm in Digital Library Management System

Yuqing Shi

Special Issue
AI-Driven Cybersecurity Solutions for Next Generation-5G/6G Enabled Cyber Physical Systems
Edited by
Dr. Muhammad Adil, Prof. Dr. Muhammad Khurram Khan and Dr. Ahmed Farouk

https://doi.org/10.3390/electronics11233916
Application of FCM Clustering Algorithm in Digital Library Management System

Yuqing Shi

Library, Hohai University, Nanjing 210098, China; shiyuqing@hhu.edu.cn

Abstract: Traditional library systems are gradually being replaced by digital libraries. Digital libraries are developing from simple database-based storage and retrieval to knowledge-based implementation. The fuzzy C-means (FCM) clustering algorithm is an example of data collection and data processing technology. It evaluates and draws conclusions based on mathematics, large data, and other technologies. In order to better improve the digital library management system, this paper applied FCM clustering algorithm to the digital library management system. Based on the in-depth study of the FCM clustering algorithm, this paper built a digital library management system. The clustering algorithm was used to cluster library borrowing records and reader information. It provided technical support and suggestions on library collection construction and book purchase and promoted book management to form a good splitting cycle. The experimental results extracted during the evaluation phase demonstrated that the overall error rate of the suggested FCM clustering algorithm for information clustering is 3.66%, which is better than the existing comparative models. This shows that applying the FCM clustering algorithm to a digital library management system has some practical significance.

Keywords: library management system; digital library; FCM clustering algorithm; cross-estimation of mistake rate

1. Introduction

Libraries have a long history as cultural centers for providing information and knowledge. Libraries also carry out digital reform with the changes in society. As part of the digitization of library resources, digital library management systems are essential for systematically storing resources and ultimately providing services to online readers. In order to meet the information needs of teachers and students quickly, effectively, and completely, this paper uses the FCM clustering algorithm to study the library management system.

The purpose of digital libraries is to provide users with the information they need. At present, there are many studies on digital libraries. Muhammad examined the download behavior of readers of two well-known journals in education and found that by using cumulative download counts over a few months, downloads could be predicted more accurately [1]. MandaliaH described experience in developing digital libraries in local languages and innovated methods for designing relevant local content so as to improve the availability of localization [2]. Gooding P outlined the design and implementation of digital library systems. The purpose was to define the stages and steps required to design a digital library system so that digital library developers could use these steps as a guide for designing digital libraries in the most efficient way [3]. Das surveyed 1220 research papers on digital libraries published by BRIC countries between 2000 and 2019 and found that more and more research was done on digital libraries in BRIC countries [4]. Aditya A provided an in-depth analysis of the quality of service of the Jakarta Digital Library of the Jakarta Provincial Library and Archives Office. The survey results showed that the service quality of the DKI Jakarta Provincial Library and Archives Bureau’s digital library was not entirely considered good, and there was still room for improvement [5]. However, the
current research on digital library management system still needs to be improved, and new algorithms need to be introduced to study it.

The FCM clustering algorithm has been paid increasing attention because of its ability to analyze big data. Yang J developed a clear algorithm for C-means classification based on histogram (AFCMH). The experimental results showed that the computational time of this method was significantly shorter than that of the traditional segmentation algorithm based on mean shift [6]. Xiao M recommended an improved C algorithm that was not aligned with the spatial algorithm to solve the problem of C-means fusion [7]. Kiki M presented a parallel FCM cluster integration algorithm based on MapReduce. The experimental results showed that the parallel FCM clustering integration algorithm had good performance, fast speed, and scalability [8]. Wang H Y proposed a new FCM clustering algorithm validity function to verify the validity of clustering results. The results showed that the proposed validity function could more accurately obtain the optimal number of clusters for the dataset, which has a strong adaptability [9]. Huang D proposed a parallel fuzzy clustering method based on cut set to solve the complex and inefficient problem of fuzzy number classification processing. The theoretical analysis and application showed that this method had good classification accuracy and efficiency for fuzzy data clustering [10]. In order to provide better services to readers, this paper uses the FCM clustering algorithm to improve the digital library management system.

This paper first discusses the database creation process and the business process of the data management system of the book management system based on the FCM clustering algorithm. It also uses the FCM clustering algorithm to retrieve useful data from it, which provides support and evaluation for data integration, production, and analysis. This study plays a positive role in timely understanding the information needs of readers, intelligent allocation of library resources, improving the efficiency of resource utilization, book quality, book management, and so on.


The establishment of a new library management system requires a comprehensive, detailed, and thorough study of the current system. This was done specifically by means of direct interviews and ad hoc visits and participation in business practices, working with staff through the most basic work processes. The main way in which information management was originally carried out in the library management and processing center was based on manual processing, at most printing some lists using the computer’s text system, with a high volume of information processing, prone to errors and lacking a systematic and standardized means of information management. The library management system that is now ready to be established implements computerized unified management of the library’s book management, reader management, book lending management, and other daily management work in order to improve efficiency and management levels [11]. This system analyzes the demand from several aspects such as library collection construction, shelf management, collection utilization rate, and personalized construction, and carries out the design of the management system.

2.1. Needs of Library Management Systems

The purpose of the library management system is to make the information management of the library efficient, large-scale, and automated in order to improve the efficiency of work. Library management systems can use clustering algorithms to analyze the types of books that have been borrowed frequently, to increase or decrease the number of different types of books in the process of book purchase, and to help developers understand new books on the shelves [12]. The clustering algorithm flow is shown in Figure 1.
Figure 1. Flow chart of clustering algorithm.

The library management system is divided into five parts: control system, book management, borrowing management, book return management, and information list. The control system contains the system control interface, including user access, adding users, editing users, deleting users, and exiting the system. Book management includes adding books, modifying books, and deleting books. Borrowing management includes borrowing books and changes in borrowing information. Book return management manages book borrowing by returning books and viewing the information on book returns. The information list includes a list of books, a list of borrowings, and a table of user details. The library management system hierarchy module is shown in Figure 2.

Figure 2. Library management system level module.

Currently, libraries divide their collections into different sections according to regulations. Combining similar books together may increase the readability of some but may not interest the reader in other publications. At the same time, if it arranges the shelves according to specific numbers: each type of book should have enough space for new books. If the number reserved is too large, it would waste space. If not enough space is reserved,
it would be difficult for books to be stored. As books continue to be sorted and classified, they quickly get old and damaged.

The "28 effect" exists in many places; it is the same as libraries. There are thousands of books in the library, not all of which will be used by readers. Through data mining technology, books with high and low reading volumes can be found. Popular collections can be increased, and the loss value of books can be removed by donation, so as to improve the utilization rate of books.

For groups with different reading habits, using the same working mode, the number and duration of books borrowed limit the active reading group’s reading interest. Inactive reading groups cannot get accurate guidance and publicity, and their attitudes towards reading cannot be easily changed. To address this situation, data mining can be used to obtain useful information about groups. For example, by analyzing their borrowing data, an active reader can increase the maximum number of books borrowed and reduce reading restrictions. For ordinary readers, these methods find the most popular books to introduce and good books to recommend. In addition, according to the borrowing habits of the readers, including the borrowing behavior of the library readers and the types of books they borrow, each reader must go through the library to create a space for a dedicated display and reading of books, which can attract all readers to read eagerly in the library. In addition, according to the borrowing number of different books and the time characteristics of borrowing readers, many reading activities can be carried out scientifically to improve the effective utilization of books and improve services.

2.2. Overall Design of Digital Library Management System

This system is a network application system using B/S mode. The advantage of B/S mode is that the system is easy to develop, maintain, and upgrade, and the cost of control is low. This feature has the same user interface. It is easy to use and can be used by different users with different login settings. In general, the client does not need to install new software. It can be used on the browser Web page, which makes the system easy to maintain and manage. In addition, the application runs on the server side, and the management, updating, and upgrading of the system are very flexible, reducing the interaction between the server side and the client side. The B/S mode is shown in Figure 3.

![Figure 3. B/S mode.](image-url)

The whole library website system uses a three-tier structure: data access layer, business logic layer, and user interface layer.

The user interface of the device is located directly in the user interface section, and the device has multiple interactive pages. Users can access the system directly through
this layer to perform necessary tasks. Depending on the type of user, different users use different credentials to log on to the system. Taking a university library as an example, the system is divided into three different users: teachers, students and administrators. The permissions are from bottom to top. In addition to reading and browsing all the resources on the website, teachers and students can read and borrow different textbooks, but they have different borrowing powers. Administrators can update login information and manage all administrative issues and settings after login.

The business logic layer implements the business functions of an application, situated between the user and data layers, which is the core component of the entire hierarchical type. The system has several modules that provide user interface functionality and access the database by calling the functionality provided by the data access layer.

The data layer provides access to external databases, which is a lower part of the overall layer-based system. The data access layer provides services in the business logic layer, including user information, book content details, and so on [14]. User information includes basic information for temporary users and administrators. Web site content data mainly include book ordering information, borrowing information, and other information.

The library system has two components: the client system and the server. The library management system structure is shown in Figure 4 [15].

![Figure 4. Library management system structure.](image)

The main functions of the client system are the registration of user information, the login of users to the library system, the browsing of books and their basic information, the book inquiry function, the borrowing of necessary books, and the inquiry of library staff about pre-purchase situations. The main functions on the server side are to manage book resources, manage user information, manage book bookings and after-purchase books, and recommend smart books.

The business model of the library system has 23 business links. Users only need to browse books online at anytime and anywhere [16]. These 23 business links are mail application, mail form, application form, successful registration of client system, adding user name password, user information library, user login, successful login, book browsing, book borrowing, history information database, storage of books, adding to shopping list, submitting, adding new book information to new shopping library, submitting new book information, extracting related information, submitting recently purchased book information, submit sample results and user names, read user names and e-mail addresses, store user data, assign information to mail servers, and send various subscription information to mail readers. The library business process is shown in Figure 5.
3. Algorithms of Digital Library Management System

Cluster analysis is an important feature of data mining. Integration is one of the most common methods in the data mining industry used to detect a set of unknown objects in a database. Through the clustering process, people who do not meet the same criteria are grouped into different groups, each of which is called a class.

When providing a dataset $V\{v_i | i = 1, 2, \ldots n\}$ representing data objects, the data are grouped according to the similarity between data objects $v_i$ and is satisfied:

$$C_j \parallel j = \{1, 2, \ldots k\}, \quad C_j \subseteq V \quad (1)$$

$$C_j \cap C_i = \emptyset , \quad \cup_{i=1}^{k} C_i = V \quad (2)$$

This process is called clustering, and $C_i(i = 1, 2, \ldots n)$ is called clustering (class).

3.1. Cluster Analysis Principles

The input to cluster analysis is represented by a set of criteria for $(x, s)$ or $(x, d)$, where $X$ represents a set of samples, and $s$ and $D$ are similar or inconsistent values [17] between the combined samples, respectively. The output of a clustering system is a set of $x = (G_1, G_2, \ldots , G_k)s$, and $G_k(k = 1, 2, \ldots N)$ is a subset of $x$, as follows:

$$G_1 \cup G_2 \cup \ldots \cup G_k = X \quad (3)$$

$$G_j \cap G_i = \emptyset , \quad i = j \quad (4)$$

The member $G_1, G_2, \ldots , G_k$ in $X$ is called a class, and each class has certain characteristics. In clustering, the results of cluster analysis are classes and their attribute descriptions.

3.2. Cluster Analysis Criteria

3.2.1. Measures of Similarity

Generally, clustering algorithms use spatial distance as a measure to calculate the difference between two samples. The inequality table is defined by $d(x, y)$, and the difference is often referred to as distance.
A. Manhattan Distance

\[ d(i, j) = \left| x_{i1} - y_{j1} \right| + \left| x_{i2} - y_{j2} \right| + \ldots + \left| x_{im} - y_{jm} \right| \] (5)

In formula (5), the distance from data object i to data object j is \( d(i, j) \); \((x_{i1}, x_{i2}, \ldots x_{im}), (x_{j1}, x_{j2}, \ldots x_{jm})\) are the characteristic \( m \) of data object i and data object j, respectively.

B. Minkowski Distance

It is supposed that \( x, y \) is the corresponding part and \( N \) is the dimension. Minkowski distance measurements for \( x \) and \( y \) are in the following forms:

\[ d(x, y) = \left[ \sum_{i=1}^{N} |x_i - y_i|^r \right]^{1/r} \] (6)

when \( r = 1 \), the formula is as follows:

\[ d(x, y) = \sum_{i=1}^{N} |x_i - y_i| \] (7)

when \( r = 2 \), the formula is as follows:

\[ d(x, y) = \left[ \sum_{i=1}^{N} |x_i - y_i|^2 \right]^{1/2} \] (8)

C. Euclidean Distance [20]

\[ d(i, j) = \sqrt{|x_{i1} - y_{j1}|^2 + |x_{i2} - y_{j2}|^2 + \ldots + |x_{im} - y_{jm}|^2} \] (9)

D. Quadratic Distance

\[ d(x, y) = \left( (x - y)^T A (x - y) \right)^{1/2} \] (10)

where \( A \) is a nonnegative definite matrix [21].

\[ d(x, y) = \left[ \sum_{i=1}^{n} a_{ii} |x_i - y_i|^2 \right]^{1/2} \] (11)

E. Minkowski Distance

\[ d(i, j) = \left( |x_{i1} - y_{j1}|^q + |x_{i2} - y_{j2}|^q + \ldots + |x_{im} - y_{jm}|^q \right)^{1/q} \] (12)

where \( q \) is a positive integer [22].

F. Cosine Distance

\[ d(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} \] (13)

3.2.2. Cluster Criterion Function

A. Error Square and Criterion Function E

Given a dataset \( X = \{x_1, x_2, \ldots, x_n\} \) containing \( n \) objects, it is divided into separate partitions \( K \) based on the size of similarity, each partition being a cluster containing \( n_1, n_2, \ldots, n_k \) elements. To evaluate the results of clustering, a squared error criterion function [23] is used:

\[ E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \overline{x_i}||^2 \] (14)

\[ x_i = \frac{1}{n_i} \sum_{x \in C_i} x \] (15)

B. Weighted Average Square Distance and Criterion J
The square distance $J$ of the average weight is defined as [24]:

$$J = \sum_{j=1}^{c} p_j s_j^*$$  \(16\)

where $P$ is the sum of the variables and $s_j^*$ is the orthogonal mean between the samples in the class.

$$s_j^* = \frac{2}{n_j(n_j - 1)} \sum_{x \in X_j} \sum_{x' \in X_j} \|x - x'\|^2$$  \(17\)

$$p_j = \frac{n_j}{n}, \ j = 1, \ldots, c$$  \(18\)

Substituting formulas (17) and (18) into formula (16) gives:

$$J = \frac{1}{n} \sum_{j=1}^{c} n_j s_j^*$$  \(19\)

C. Class Spacing and Criteria $F$

To describe the distance ratio between categories of aggregation results, the difference $F_2$ between the distance between categories $F_1$ and the average weight of the categories can be used, which is defined as [25]:

$$F_1 = \sum_{j=1}^{c} (m_j - m)^T (m_j - m)$$  \(20\)

$$F_2 = \sum_{j=1}^{c} p_j (m_j - m)^T (m_j - m)$$  \(21\)

In the formulas, in the mean vector sample $m_j$, $m$ is the average vector of all samples, and $p_j$ is the probability.

The distance between data processing for different components defines the degree of separation between different types of data. It can be seen that the larger the distance value, the better the separation effect of various grouping results, and the higher the quality of grouping.

4. Application Experiment of the FCM Clustering Algorithm in Digital Library Management System

This paper takes a university digital library as the research object, uses the FCM clustering algorithm to cluster its book borrowing information and reader information, and verifies the accuracy of the clustering results.

4.1. Clustering of Book Borrowing

In the library system, there is a large amount of book borrowing data, and the frequency of book use can be obtained by cluster analysis of the data of these books. This paper chooses last year’s digital library borrowing information to clear up the non-cluster analysis data. Finally, three aspects of data—book number, borrowing date, and return date—are selected to compose the initial data. This paper uses the FCM clustering algorithm to classify books into three categories according to their heat, and shows the results of clustering the database, as shown in Table 1 and Figure 6.

**Table 1. Library classification clustering data.**

<table>
<thead>
<tr>
<th>Classification Borrowing Number</th>
<th>Category</th>
<th>Average Annual Number of Loans</th>
<th>Average Number of Loan Days Per Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very popular books</td>
<td>4269</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>More popular books</td>
<td>2358</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Unpopular books</td>
<td>956</td>
<td>13</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6. Library lending data graph.

From the data, the average number of books borrowed per year is between 0 and 6000, indicating that the usage of different types of books varies greatly. The data in the second and fourth quarters are significantly higher than those in the first and third quarters, which may be caused by final exams in the second and fourth quarters. For borrowing days, books returned within 1–3 days do not meet the reader’s true borrowing needs, so the length of borrowing should be set to 4–60 days. For popular books, the number of days of borrowing is much greater than the number for unpopular books, indicating that the popular book type is more in line with the needs of the reader. The analysis of the above results shows the readers’ interest and preference for books. Among them, the number of books in the most popular categories is very large, which can increase the number of such books, timely supplement the purchase, and check the integrity of the books to ensure that the readers can borrow them when they need them. In addition, in terms of book placement, one might consider placing such books in a prominent place in the library for readers to borrow. This increases the rationality of book purchase and allocation of book purchase funds, and does not cause blindness in book purchases.

4.2. Clustering of Reader Information

In this paper, an aggregation algorithm is used to collect the borrowing information of readers and classify the reader groups according to their needs, so as to better serve the readers. This paper selects the reader information table and the reading history information table in order to analyze the historical information of the books borrowed last year and sort the reader groups according to the different results of the books borrowed. The results of reader information clustering are shown in Table 2 and Figure 7.
Table 2. Cluster data table of reader information.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Category</th>
<th>Average Annual Reading</th>
<th>Total Number of Borrowing Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular readers</td>
<td></td>
<td>59</td>
<td>179</td>
</tr>
<tr>
<td>Average readership</td>
<td></td>
<td>34</td>
<td>115</td>
</tr>
<tr>
<td>Infrequent readers</td>
<td></td>
<td>22</td>
<td>58</td>
</tr>
</tbody>
</table>

![Regular readers](indicative_icon) [Average readership](indicative_icon) [Infrequent readers](indicative_icon)

Figure 7. Clustering data graph of reader information.

As can be seen from Table 2 and Figure 7, readers can be classified into three categories according to their amount of reading: those who read frequently, those who read generally, and those who do not read regularly. Additionally, those who read frequently differ greatly; they have different requirements for reading. In true library services, most libraries follow the same rules for users. In fact, borrowing times are often different between readers. For readers who like to read, because of the large demand for borrowing, the library can consider changing the number of books borrowed and circulated, rather than treating readers with different needs in the same way, to better serve the readers. On the other hand, for those who read frequently, their borrowing interest can be further evaluated by providing them with effective publicity services.

4.3. Assessment of Cluster Analysis Results

In order to evaluate the accuracy of cluster analysis results, this paper uses the error rate cross-estimation method [26] to verify the performance of cluster analysis. The basic principle of the cross-estimation method is to group each group of data and find out each sample group. This paper compares the classification results to get the percentage of error data. The smaller the percentage, the better the clustering ability of the clustering algorithm. It uses the above methods to calculate the number of false positives for each type of data in the FCM clustering algorithm used in this paper. Figure 8 and Table 3 are obtained.
The error rate of the system is obtained by dividing the number of errors of each type of data by the number of data in the algorithm. From the data, the overall error rate of the suggested FCM clustering algorithm for information clustering is 3.66%, which is better than the existing comparative models. Therefore, the use of the FCM clustering algorithm in the digital library management system works well.

5. Discussion on the Application of the FCM Clustering Algorithm in Digital Library Management Systems

This paper innovates the digital library management system through the FCM clustering algorithm and verifies the accuracy of this algorithm in the digital library management system through experiments. This paper mainly addresses the following two aspects:

(1) It improves the structure of library collection and the utilization of resources. Data aggregation consists of three datasets: basic information records of readers, reading information, and returning records of readers. Readers are classified and collected according to the information provided by the library.
(2) Based on the analysis of the reader’s behavior, it identifies the interest, reading behavior, and reading trend of a specific group or individual and predicts the reader’s future reading behavior. According to the reading habits of the readers, it can use a clustering algorithm to divide the readers into groups with different characteristics and personalize the services according to the characteristics of different groups.

Prospects for future work:
(1) Research on various clustering algorithms are applied. Clustering algorithms can be used in many fields, but because of the lack of familiarity with the various fields and the limited access to data, the application of research has not been fully developed.
(2) Improvements have been made to address the shortcomings of the algorithms. Due to the shortage of time, the idea of improving the FCM algorithm is not mature. In future work, improvements are made to address the shortcomings of the FCM algorithm.

6. Conclusions
Taking a digital library as an example, this paper analyzed the lending data and reader information of various types of books, and used FCM clustering analysis to cluster and organize the data. The quality of numerous books and the reading habits of readers were assessed, and the effectiveness of the whole FCM clustering process was verified through the error rate cross-estimation method, The study showed that the application of the FCM clustering algorithm to the library management system can effectively reduce the overall error rate of the management system and provide theoretical support for the library collection layout and book acquisition.

Funding: This work was supported by the Big Data Research Project of Jiangsu Library in 2021 (Project No.2021JSTD022). The result is one of the results of “2021 Big Data Research Project of Jiangsu Library”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References


