11-24-2021

Use of K-Means Clustering Method for Books Data in Acharya Raghuveer Library, Central University of Himachal Pradesh, Dharamshala, India

Vishal .
Research Scholar, DLIS, Central University of Himachal Pradesh, India, vishucena77@gmail.com

Abhinandan Kumar
Research Scholar, DLIS, Central University of Himachal Pradesh, India, abhi.rs.cuhp@gmail.com

Dr. Pawan Kumar Saini
Assistant Professor, Central University of Himachal Pradesh, India, pawankumarsaini@hpcu.ac.in

Use of K-Means Clustering Method for Books Data in Acharya Raghuveer Library, Central University of Himachal Pradesh, Dharamshala, India

Vishal[1], Kumar, A.[2]and Saini, P.K.[3]*

[1]Research Scholar, DLIS, Central University of Himachal Pradesh, India
[2]Research Scholar, DLIS, Central University of Himachal Pradesh, India
[3]Assistant Professor, DLIS, Central University of Himachal Pradesh, India
*vishucena77@gmail.com
*abhi.rs.cuhp@gmail.com
*pawankumarsaini@hpcu.ac.in

Abstract

Purpose- The study formulates clusters of Library and Information (LIS) discipline books from Acharya Raghuveer Library, Central University of Himachal Pradesh. Cluster formulation aims to identify the utilization rate of books. Furthermore, the study attempts to find the direction for collection management and developing related services for the users in the library.

Design/methodology/approach- Book**s** circulation data set is obtained, formatted and made compatible for further analysis. The Euclidean distance formula is used to find the distance between two points on a plane. Afterwards, the K-means clustering algorithm is implemented to generate the required clusters.

Findings- Three clusters of books namely, Frequently Borrowed books (C1), Rarely Borrowed Books (C2) and Most Frequently Borrowed Books (C3) are derived. Out of a total of 30 book records, 19 books are grouped as often borrowed books, 6 books are grouped as rarely borrowed books and 5 books are grouped as most frequently borrowed books.

Originality/value- The study points out the importance and applicability of data analysis to examine the patterns of resource usages in libraries. Further comprehensive study with data analysis may give directions for collection development and management in libraries.

Keywords: Books Data, Circulation Evaluation, Cluster technique, K-Means Algorithm

1. Introduction

To scrutinize the utilization rate of resources is a key task for libraries. This task is carried out with the aid of book circulation data analysis. Data analysis is the process of examining, filtering, integrating, and modelling data with the objective of identifying usable information, generating conclusions and facilitating decision-making. There are various methods and tools are available to perform data analysis. We have to select data analysis method and tool as per the objective of analysis, type of data set and the specific area of application. K-Means clustering method is used in this study to identify the use and circulation pattern of Library and Information Science (LIS) discipline books in Acharya Raghuveer Library, Central University of Himachal Pradesh.

Clustering is the process of dividing a population or set of data points into groups and assigning them to cluster groups based on similarities. K-Means Clustering is a centroid-based, partitioning clustering and unsupervised learning algorithm. In this algorithm, distance-based measurements are used to determine the similarity between data points. Unlabeled dataset is grouped into K-number of clusters. In this process the value of K should be predetermined and the process is repeated until it does not find the best clusters.

In this study, circulation data set of LIS discipline books in Acharya Raghuveer Library is taken as input in which borrowed books and their borrowing frequency are the variables. Three clusters of LIS books, namely Often Frequently Borrowed (C1), Rarely Borrowed (C2) and Most Frequently Borrowed (C3) is obtained by applying k-means clustering method. The findings of this study may be used in managing the collection of LIS discipline books in the library and users will be better served in terms of their needs and demands. Such an analysis can also be done in the context of another disciplines or entire collection of books in the library. Furthermore, the book recommender system may also be developed.

2. Literature Review

- Veepu Uppal and Gunjan Chindwani (2013) have reinterpreted the findings derived by JianWei Li and Pinghua Chen (2008) in their study as: “How to improve the utilization rate of library resources, how to serve the reader better and how to play more active roles, all have been becoming the concrete task of library in future. The data mining of the books circulation and user needs in library automation system provided effective support for library management.”

- Ping YU (2011) concluded that “The circulation data is the best evidence of the library resources used. It is a mirror can reflect the actual information that readers needed. Therefore, it is a certain reference to grasp reader’s interest, and thus as the basis for strengthening the use of library resources.”

- V. Krishnamurthy and Dr. R. Balasubramani (2014) have highlighted the importance of data mining in libraries in their study- “Data Mining techniques when applied to library transactional databases helps to explore interesting patterns of data that help the librarians to take decisions on usage of books patterns, the kind of books and magazines that has to be purchased. The data mining techniques in library also helps to identify the hidden knowledge that is stored in the databases, which helps the librarians and organization to take decisions based on their own data for providing better services and effective utilization of resources.”

- In a study entitled Comprehensive review on Clustering Techniques and its application on High Dimensional Data, Afroj Alam, Mohd Muqeem and Sultan Ahmad (2021) points out that “Partitioning clustering algorithm is easy to understand, implement and scalable because it will take less to execute compared to another algorithm. It works good for Euclidian distance data.”

All the aforesaid review points give us direction that the study of book circulation record by appropriate data mining method is fruitful for evaluating the usefulness of library resource collection.

3. Objective

The objective of this study is to evaluate the use and circulation pattern of LIS books in the following way:

- To identify the borrowing frequency of books in the form of clusters (C1, C2, and C3).
- To determine the specific cluster of books as most usable, often usable and rarely usable.

4. Methodology

The study involves following steps to collect and analyze data:

Step-1: Data of borrowed book of LIS for a definite periodis obtained from library automation software.

Step-2: Data file format is changed as per the application tool for analysis.

Step-3: Data are arranged into sequence to make them compatible for k-means clustering method.

Step-4: Three K (cluster) is decided. These clusters are Often Frequently Borrowed(C1), Rarely borrowed(C2) and Most Frequently Borrowed(C3).

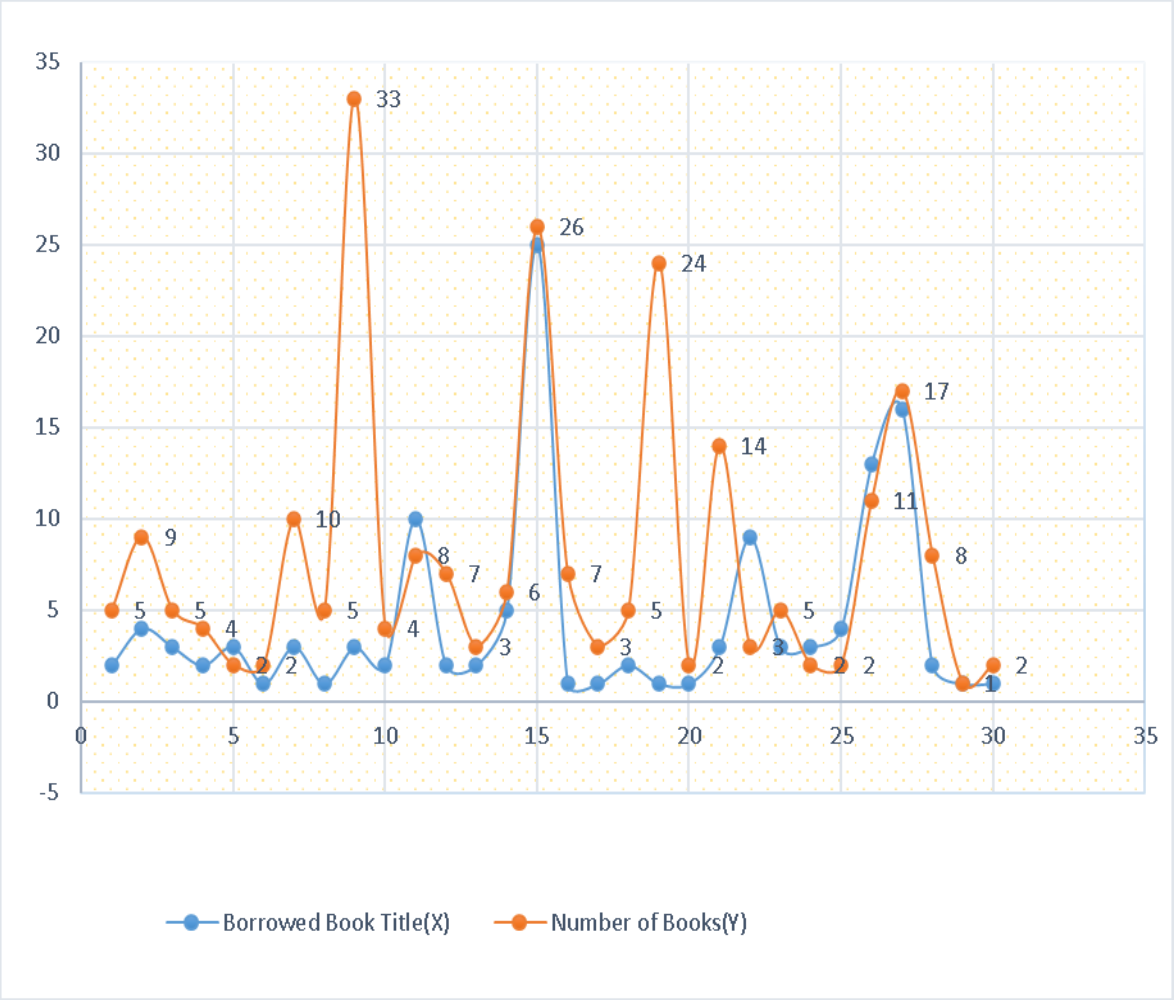
Step-5: Random K points or initial centroids is selected.

Step-6: K clusters (C1, C2 and C3) is formulated by assigning each data point to its closest.

Step-7: New centroids have been regenerated by calculating variances. As a result, clusters are also reformulated until regenerated centroids get the same value as in the previous generated centroids.

5. Data Collection and Analysis

Only two attribute of borrowed books which are title and borrowing frequency are needed to perform the analysis. So the data set of two variables (X and Y) is taken where variable X represents borrowed books titles and Y represents borrowing frequency. The obtained data have been changed into digits, because the initial data is not in the form of digits or numbers. The data set belongs to the books of LIS discipline only. The detail description of the compatible data set is as shown in the graph-1.



Graph- 1.Complete Data set of Variable X and Y

Random K points or initial centroids are selected as shown in Table-1. In order to measure the distance between objects and means, the Euclidean distance formula (used to find the distance between two points in a plane) is applyed:

$$D = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}.$$

Nearest cluster distance calculation is shown in table-1.1 and obtained cluster members are shown in Table-1.2

Randomly selected centroids or initial data point from data set (X, Y)		
Centroid 1	2	4
Centroid 2	5	6
Centroid 3	9	3

Table-1

Serial No.	Borrowed Book Title (X)	Number of Books (Y)	Distance(D)			Nearest Cluster Distance
			C1(2,4)	C2(5,6)	C3(9,3)	
1	2	5	1.41	3.16	7.28	C1
2	4	9	5.38	3.16	7.81	C2
3	3	5	1.41	2.23	6.32	C1
4	2	4	1.41	3.6	7.07	C1
5	3	2	2.23	20	6.08	C1
6	1	2	2.23	5.65	8.06	C1
7	3	10	6.082	4.47	9.21	C2
8	1	5	1.41	4.12	8.24	C1
9	3	33	28.98	27.07	8.12	C3
10	2	4	1.41	3.6	7.07	C1
11	10	8	8.94	5.38	5.09	C3
12	2	7	3.16	3.06	8.06	C2
13	2	3	1.41	4.24	7.07	C1
14	5	6	3.6	1.41	5	C2
15	25	26	31.82	28.28	28.01	C3
16	1	7	3.16	6.4	8.94	C1
17	1	3	1.41	5	8.06	C1
18	2	5	1.41	3.16	7.28	C1
19	1	24	20.02	18.43	22.47	C2
20	1	2	2.23	5.65	8.06	C1
21	3	14	10.04	8.24	12.52	C2
22	9	3	7.07	5	1.41	C3
23	3	5	1.41	2.23	6.32	C1
24	3	2	2.23	4.47	6.08	C1
25	4	2	4	4.12	5.09	C1
26	13	11	13.03	9.43	8.94	C3
27	16	17	19.1	15.55	15.65	C2
28	2	8	4.12	3.6	8.6	C2
29	1	1	3.16	6.4	8.24	C1
30	1	2	2.23	5.65	8.06	C1

Table- 1.1

Clusters	Cluster Members
Cluster 1	1, 3, 4, 5, 6, 8, 10, 13, 16, 17, 18, 20, 23, 24, 25, 29, 30
Cluster 2	2, 7, 12, 14, 19, 21, 27, 28
Cluster 3	9, 11, 15, 22, 26,

Table-1.2

Movements have been found in clusters data points in the region, so re-generation of new centroids is required. The new centroids are calculated by the following formula:

$$\mathbf{C} = \frac{\sum \mathbf{m}}{\mathbf{n}}$$

In the above formula C is the data centroid, m is a member of data that belongs to a particular centroid and n is the amount of data that is a member of a particular centroid.

After the process of calculation, new centroids have been obtained as shown in Table-2. On the basis of new centroids, nearest cluster distance calculation is shown in table-2.1 and obtained cluster members are shown in Table-2.2

Re-generated new centroids		
Centroid 1	1.94	3.47
Centroid 2	4.5	11.8
Centroid 3	11.8	16.2

Table-2

Serial No.	**Borrowed Book Title (X)**	**Number of Books (Y)**	**Distance(D)**			**Nearest Cluster Distance**
			C1(1.94,3.47)	**C2(4.5,11.8)**	**C3(11.8,16.2)**	
1	2	5	1.53	7.28	14.88	C1
2	4	9	5.9	2.84	10.61	C2
3	3	5	1.86	6.96	14.24	C1
4	2	4	0.53	8.19	15.64	C1
5	3	2	1.81	9.91	16.7	C1
6	1	2	1.74	10.4	17.84	C1
7	3	10	6.61	2.34	10.76	C2
8	1	5	1.79	7.64	15.55	C1
9	3	33	29.54	21.25	18.96	C3
10	2	4	0.53	8.19	15.64	C1
11	10	8	9.24	6.68	8.39	C2
12	2	7	3.53	5.41	13.44	C1
13	2	3	0.47	9.14	16.44	C1
14	5	6	3.97	5.82	12.25	C1
15	25	26	32.23	24.94	16.44	C3
16	1	7	3.65	5.94	14.18	C1
17	1	3	1.05	9.47	17.05	C1
18	2	5	1.53	7.24	14.88	C1
19	1	24	20.55	12.69	13.32	C2
20	1	2	1.74	10.4	17.84	C1
21	3	14	10.58	2.66	9.07	C2
22	9	3	7.07	9.88	13.49	C1
23	3	5	1.86	6.96	14.24	C1
24	3	2	1.81	9.91	16.7	C1
25	4	2	2.53	9.81	16.2	C1
26	13	11	13.38	8.53	5.33	C3
27	16	17	19.51	12.62	4.27	C3
28	2	8	4.53	4.54	12.77	C1
29	1	1	2.64	11.35	18.64	C1
30	1	2	1.74	10.4	17.84	C1

Table- 2.1

Clusters	Cluster Members
Cluster 1	1, 3,4, 5, 6, 8, 10, 12, 13, 14, 16, 17, 18, 20, 22, 23, 24, 25, 28, 29, 30
Cluster 2	2, 7, 11, 19, 21
Cluster 3	9, 15,26, 27

Table- 2.2

Again, movements have been found in clusters data points in the region, so re-generation of new centroids is essential. Calculated values of new centroids, nearest cluster distance calculation and obtained cluster members are shown in Table-3, Table-3.1 and Table-3.2 respectively.

Re-generated new centroids		
Centroid 1	2.42	3.95
Centroid 2	4.2	13
Centroid 3	14.25	21.75

Table-3

Serial No.	**Borrowed Book Title (X)**	**Number of Books (Y)**	**Distance(D)**			**Nearest Cluster Distance**
			C1(2.42,3.95)	**C2(4.2,13)**	**C3(14.25,21.75)**	
1	2	5	1.13	8.29	20.75	C1
2	4	9	5.29	4	16.35	C2
3	3	5	1.19	8.08	20.17	C1
4	2	4	0.42	9.26	21.56	C1
5	3	2	2.03	11.06	22.72	C1
6	1	2	2.41	11.45	23.78	C1
7	3	10	6.07	3.23	16.26	C2
8	1	5	1.76	8.61	21.35	C1
9	3	33	20.03	15.9	29.05	C2
10	2	4	0.42	9.26	21.56	C1
11	10	8	14.39	7.65	8.59	C3
12	2	7	3.07	6.39	19.17	C1
13	2	3	1.03	10.23	22.39	C1
14	5	6	21.60	6.69	3.29	C3
15	25	26	31.56	24.52	11.55	C3
16	1	7	3.36	6.8	19.82	C1
17	1	3	1.7	10.49	22.95	C1
18	2	5	1.13	8.29	20.75	C1
19	1	24	20.1	11.45	13.43	C2
20	1	2	2.41	11.45	23.78	C1
21	3	14	10.06	1.56	13.66	C2
22	3	9	11	6.64	19.47	C2
23	3	5	1.19	8.08	20.17	C1
24	3	2	2.03	11.06	22.72	C1
25	4	2	2.5	11	22.25	C1
26	13	11	12.71	10.82	9.02	C3
27	16	17	18.83	12.45	5.06	C3
28	2	8	4.07	5.46	18.41	C1
29	1	1	3.27	12.41	24.61	C1
30	1	2	2.41	11.45	23.78	C1

Table-3.1

Clusters	Cluster Members
Cluster 1	1, 3, 4, 5, 6, 8, 10, 12, 13, 16, 17, 18, 20, 23, 24, 25, 28, 29, 30
Cluster 2	2, 7, 9, 19, 21, 22
Cluster 3	11, 14, 15, 26, 27

Table- 3.2

Still, movements have been found in clusters data points in the region. Accordingly, re-generation of new centroids is necessary. Thus, the calculation has been repeated. New centroids, nearest cluster distance calculation and obtained cluster members are shown in Table-4, Table-4.1 and Table-4.2 respectively.

Re-generated new centroids		
Centroid 1	2.42	3.95
Centroid 2	5.66	12.66
Centroid 3	14.66	25.33

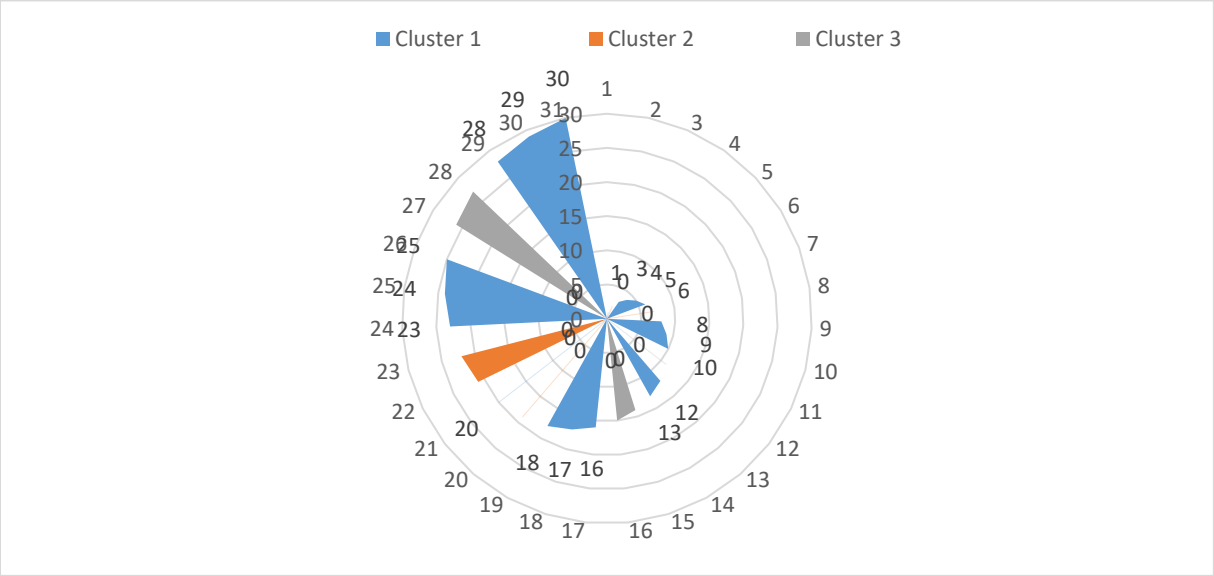
Table- 4

Serial No.	Borrowed Book Title (X)	Number of Books (Y)	Distance(D)			Nearest Cluster Distance
			C1(2.42,3.95)	C2(5.66,12.66)	C3(14.66,25.33)	
1	2	5	1.13	8.48	23.94	C1
2	4	9	5.29	4.01	19.5	C2
3	3	5	1.19	8.1	23.43	C1
4	2	4	0.42	9.4	24.8	C1
5	3	2	2.03	10.98	26.08	C1
6	1	2	2.41	11.63	27.03	C1
7	3	10	6.07	3.76	19.26	C2
8	1	5	1.76	8.96	24.49	C1
9	3	33	20.51	13.95	29.55	C2
10	2	4	0.42	9.4	24.8	C1
11	10	8	8.59	17.94	6.36	C3
12	2	7	3.07	6.74	22.27	C1
13	2	3	1.03	10.33	25.66	C1
14	5	6	18.26	7.04	3.29	C3
15	25	26	31.56	23.49	10.36	C3
16	1	7	3.36	7.33	22.86	C1
17	1	3	1.7	10.72	26.17	C1
18	2	5	1.13	8.48	23.94	C1
19	1	24	20.1	12.26	13.72	C2
20	1	2	2.41	11.63	27.03	C1
21	3	14	10.06	2.97	16.25	C2
22	3	9	10.22	6.64	23.03	C2
23	3	5	1.19	8.1	23.43	C1
24	3	2	2.03	10.98	26.08	C1
25	4	2	2.5	10.78	25.65	C1
26	13	11	12.71	14.42	7.52	C3
27	16	17	18.83	11.21	8.43	C3
28	2	8	4.07	5.92	21.46	C1
29	1	1	3.27	12.55	27.9	C1
30	1	2	2.41	11.63	27.03	C1

Table- 4.1

Clusters	Cluster Members	Remarks
Cluster 1	1, 3, 4, 5, 6, 8, 10, 12, 13, 16, 17, 18, 20, 23, 24, 25, 28, 29, 30	Often frequently
Cluster 2	2, 7, 9, 19, 21, 22	Rarely
Cluster 3	11, 14, 15, 26, 27	Most frequently

Table- 4.2



Graph- 2. Graphical Representation of Clusters 1, Cluster 2 and Cluster 3

Now, compare each cluster members from Tables 3.2 and Table 4.2. There is no movement of cluster data points, it indicates that desired cluster has been formulated. Now if new centroids are regenerated, then we will get the same result as the previous generated centroids i.e., which is recorded in Table-4.

6. Conclusion

Clustering of LIS discipline books in Acharya Raghuveer Library has been performed in this study by using K-means clustering algorithm. The purpose of clustering is to examine the usability and borrowing pattern of books related to LIS discipline. Group of often frequently borrowed books, rarely borrowed books and most frequently borrowed books is easily obtained by analyzing the circulation data set as per the steps of adopted clustering method. In this study data of 30 books are collected, of which 19 books are grouped as often borrowed books, 6 books are grouped as rarely borrowed books and 5 books are grouped as most frequently borrowed books. The findings of the study are crucial for managing the collection of books and developing policy for collection development in the library.

This study is confined to small scale dealing with books of single discipline, whereas the library has a huge collection of books related to various disciplines. Therefore, a comprehensive study is necessary in this context. On the basis of which the facility of book recommendation system may also be given to the readers.

7. References

1. Alam, A., Muqeem, M., & Ahmad, S. (2021). Comprehensive review on Clustering Techniques and its application on High Dimensional Data. *International Journal of Computer Science & Network Security*, *21*(6), 237-244.
2. Huang, C. M., Kang, S. H., Chang, C. C., & Lu, S. H. (2015). Apply Data Mining Techniques to Library Circulation Records and Usage Patterns Analysis.
3. Krishnamurthy, V., & Balasubramani, R. (2014, March). An association rule mining approach for libraries to analyse user interest. In *2014 International Conference on Intelligent Computing Applications* (pp. 122-125). IEEE.
4. Li, J., & Chen, P. (2008, December). The application of association rule in library system. In *2008 IEEE International Symposium on Knowledge Acquisition and Modeling Workshop* (pp. 248-251). IEEE.
5. Rai, P., & Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, *7*(12), 1-5.
6. Suresh, R., Anand, I., Vianesh, B., & Mohammed, H. R. (2018, March). Study of clustering algorithms for library management system. In *2018 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)* (pp. 221-224). IEEE.
7. Uppal, V., & Chindwani, G. (2013). An empirical study of application of data mining techniques in library system. *International Journal of Computer Applications*, *74*(11).
8. Wang, C., Xu, S., Chen, L., & Chen, X. (2016, June). Exposing library data with big data technology: A review. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)* (pp. 1-6). IEEE.
9. Yu, P. (2011, May). Data mining in library reader management. In *2011 International Conference on Network Computing and Information Security* (Vol. 2, pp. 54-57). IEEE.
10. Zhou, Y. (2020). Design and implementation of book recommendation management system based on improved Apriori algorithm. *Intelligent Information Management*, *12*(3), 75-87.