Toward Understanding the User Behavior in Sports University Library Using Hierarchical Clustering

Yu-Chia Hsu, Yung-Che Li, and Yung-Hsuan Lin

***Abstract*—The library plays an important role in higher education. the emerging electronic media and digital contents bring libraries to encounter digital innovation and changes in leadership styles. However, analysis of the behavior of book borrowing is still a way to understand the demand of users, so as to provide sufficient resources and develop customized services. The purpose of this paper is to analyze the book borrowing data of the library in order to identify the user typologies. Numerous data records were collected from a sports university library in Taiwan. Each borrowing history record contains book detail and classification number. The characteristics of user behavior were described based on these data after cleaning, aggregation, and transforming. The hierarchical clustering techniques were applied to obtain the user typologies with similar behaviors. Five clusters, the general casual reader, athletes, Art and Literature lovers, course learner, and knowledge seeker, were obtained to represent the classic user typologies of a sports university.**

***Index Terms*—Big data, data mining, behavioral patterns.**

I. Introduction

One of the main mission of a university library is to provide reference materials and study space for students and researchers. Depends on the various subjects that student majored, the degree of resource demand is different. Students of liberal arts and humanities may need to read a large number of historic and literature books. Business school's student may require a space of learning commons in library to conduct case studies. In contract, students majored in sports and excises may spend a lot of time on training in the field, and less relay on reading theoretical books. Therefore, in order to satisfied the different users expectation, the developing characteristics and strategy directions of library in a sports university is dissimilar between arts university, military university, and comprehensive university. Generally, the collection of library is focus on the sports and exercise subject for the university specialized in sports. Most portion of budget are used to buy the sports related books. However, we don't know whether this collection development policy can coincide meet the reader's demands. Consequently, how to allocate the resource and budget have been an issue for library research in recent decades [1]. Some researchers propose to analyze the uses' behavior to identify the user typologies for user engagement and service development in the future [2].

In the past years, issues such as the needs and satisfaction of library readers have been mostly studied through questionnaires or interviews. However, in recent years, with the advancement of information computing technology, the user's behavior patterns, such as the path of searching for data, the demand for electronic media, and the behavior of borrowing books, are all recorded in the database. The data of these behavioral records has become very large over time, and adopting alternative emerging research methods to analyze and interpret big data is necessary.

The rise of big data technology has brought many changes to the library. How to use these big data analysis tools, use emerging information visualization tools, and provide new ways to view the data and mine the hidden information to enhance the library service have gradually become an important issue [3]-[5]. Cano *et al*. [6] use clustering algorithms and association algorithms to obtain the relevant knowledge for library management from the daily lend, return and query data. The characterization of the user is identified as a certain profile by the clustering algorithms. The relationship between the user's behavior about what topics, place or daytime is provided by the association algorithms. Jiang *et al*. [7] use clickstream data to investigate users' information behavior. The results provide an understanding of the OPAC users' actual needs, behavior, preferences, and habits in order to make informed decisions concerning the development of the OPAC systems. Hájek & Stejskal [8] adopt a method of bibliomining to analysis the library user behavior. The conceptual framework for bibliomining includes five basic elements: operation of the library, bibliographic records, bibliometric data, library services and demographic structure of users. These data were used for identifying the typical reader's behavior using the k-mean clustering algorithm. Finally, 11 clusters were obtained based on 36 attributes of behavior. The clustering analysis algorithms also have used for library collection data like book use rate and collection proportion in the digital library [9].

In order to understand the reader's borrowing behavior, this study adopts a hierarchical clustering method to figure out similar readers based on the classification and quantity of the borrowed books.

II. Methodologies for Library Users' Behavior Analysis

A. The Research Subject and Data

A university of sport in Taiwan is took as the research subject in this study. The university of sport is a small scale university including three colleges, the college of sport performance, the college of sport education, and the college of sport industry. The total number of students is about three thousand. The students appear different characteristics

Manuscript received December 20, 2019; revised March 4, 2020.

The authors are with the Department of Sports Information and Communication, National Taiwan University of Sport, Taichung, Taiwan (e-mail: ychsu@ntupes.edu.tw).

DOI: 10.7763/IJCTE.2020.V12.1271

among three colleges. Students in the college of sport performance are mostly athletes, who are armed to win the medal in sports competition. Students in the college of sport education are developed as the teacher or coach for fundamental physical education. Student in the college of sport industry are mostly similar with comprehensive university, who are majored in management, communication, and physiology.

The data in this research is gathered from the database of the library automation system provided by a university of sport in Taiwan. The original database contains many tables and fields to support the library daily operation, including patron file, bibliographic file, circulation file, etc. Users' behavior can be retrieved and synthesized from these files and transforming to book borrowing records, and reader's profile. The users' behavior was analyzed using big data analysis and data mining technique to investigate the characteristics, preference, and specialty from different major students.

Five years' data are retrieved from the database between Jan. 1, 2013 to Dec. 31, 2017. The user's profile, books catalogue, and book circulation log are located in different data table. After data aggregating, totally 36583 records of book borrowing behaviors are collected for investigating.

In order to assure the quality of analysis, the raw data gathered from database needed to be cleaning, aggregation. The library automation system of the university library in the research have been renovate by three versions. Each migration of data from the legacy system to the moderated system may cause data missing and Inconsistent. We delete the irregular data and null value, totally more than 2700 records, and finally obtained 33770 records after data cleaning. The book borrowing logs refer to 2353 readers of 9 departments in 3 colleges. Finally, we aggregate the data from different table, and extract the data with the fields, reader's id, department id, class id, date of borrowing, book's title, and book's classification number.

The composition of the borrowing data for the reader's department is shown in Table I. Table I present that the usage of the readers in the college of sport industry is the highest, and the lowest is the readers in the college of sport performance.

TABLE I: COMPOSITION OF THE BORROWING DATA RECORDS

College	Department	Number of data records (proportion)
Sport Performance	Sport Performance	1090 (3.2 %)
	Ball Sport	1729 (5.1 %)
	Combat Sport	959 (2.8 %)
Sport Industry	Sport Information and Communication	4537 (13.4 %)
	Exercise Health Science	7588 (22.4 %)
	Sport Management	4057 (12.0 %)
	Recreational Sport	3911 (11.5 %)
Sport Education	Dance	2278 (6.7 %)
	Physical Education	7603 (22.5 %)

B. Data Processing for User Typologies Identification

Fig. 1 demonstrate the data processing flow. Data is collected from all over the database of library automation system. After aggregation and cleaning, the book borrowing records are obtained. The book borrowing records are then sorted based on each user and transformed to the individual reader's profile. In the reader's profile, the main class of each borrowed book is identified from the whole classification number, and calculate the total number according to the main class, separately. The classification number of books is followed the New Classification Scheme for Chinese libraries, which is commonly used for Chinese books. The classification scheme contains total ten main classes, and shown in Table II.

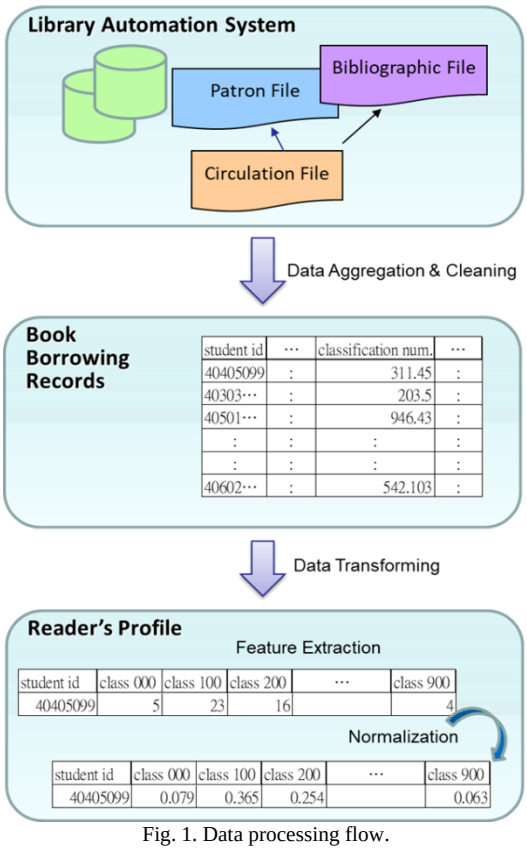

Fig. 1. Data processing flow.

The number of books borrowed by each person is various, some people borrow many books, but others borrow few books. Moreover, the type of books that individuals prefer is different. In order to identify the user typologies, we choose the number of books borrowed in each main class as the features to represent each reader.

TABLE II: NEW CLASSIFICATION SCHEME FOR CHINESE LIBRARIES

Number	Main classes
000	Generalities
100	Philosophy
200	Religion
300	Natural Sciences
400	Applied Sciences
500	Social Sciences
600	History and Geography of China
700	History and Geography of World
800	Linguistics and Literature
900	Arts

In order to compare the similarity of individual borrowing preferences, readers with different borrowing behavior need to have the same expression. Therefore, we present the number of borrowed books in a proportional manner to

achieve normalization. The number of books borrowed by each person in each main class was counted, then divide by the total number of borrowed books to calculate the percentage of individuals in each main class of books. The sum of the values of each main class is one, and the normalization is achieved by sum up all of the components of the vector equal to one.

C. Hierarchical Clustering

In order to group together the reader's behaviors with similar characteristics, the hierarchical clustering is adopted in this study. Like *K*-means clustering, Hierarchical clustering is a type of unsupervised machine learning algorithm.

When use *K*-means clustering, the number of group *k* should be determined first, then the *k* centroids are randomly selected. Each centroid would move to the average position of all items closest to it. Finally, recalculate after moving until all centroids are not finished after moving.

However, a suitable number of group *k* is hardly to determined. Hierarchical clustering need not to pre-set the number of group. It begins by finding the two closest items to form a group, treat the group as an item, and then find the nearest two items to form another group, repeating until all items are in one group. The clustering results are a tree structure diagram (also known as dendrogram) at a glance. The number of clusters can be divisive hierarchical clustering, or agglomerative hierarchical clustering. After splitting and merging through clustering, the optimal cluster number is selected.

In this study, we adopt agglomerative clustering that involves the bottom-up approach. Agglomerative hierarchical clustering begins with layered polymerization at the bottom of the tree structure. In the beginning, we regard each piece of data as a cluster. If we have *n* pieces of data now, we treat the n pieces of data as n clusters, that is, each group contains a piece of information:

The steps of the operation are as follows:

1. Treat each piece of data as a cluster C_i, i =1 to n.
2. Find all the clusters, the closest two clusters C_i, C_j
3. Merging C_i, C_j into a new cluster
4. If the current number of clusters is greater than the number of clusters we expect, repeat steps 2 through 4 until the number of clusters has dropped to the number we requested.

The two clusters Ci and C_j closest to each other are defined by measuring the distance between two clusters. The distance itself can be Euclidean or Manhattan distance. There are different ways to find distance between the clusters, and the results obtained by each method are not the same. The definition of the cluster distances used in this study is known as the Ward's method. Ward's method is a variance-minimizing approach that minimizes the sum of squared differences within all clusters. The distance between two clusters is defined as the sum of the squared distance to the mean of the combined clusters Following is the expression:

$$d(C_i, C_j) = \sum_{a \in C_i \cup C_j} \|a - \mu\| \quad (1)$$

where μ is the mean vector of $C_i \cup C_j$

III. Analysis Results

The hierarchical clustering is adapted for reader typologies analysis in this study. The reader is grouped with similar readers based on the category of the book they borrowing. We perform the analysis in Python programing language via Scikit-Learn, which is a machine learning library. Fig. 2 shows the dendrograms of the hierarchical clustering result. The vertical height of Fig. 2 shows the Euclidean distances between readers. The horizontal axis noted various clusters. The number of cluster is increased when the Euclidean distances is decreased. Explaining reader behaviors is a great challenge when there are too many clusters. Therefore, we set 10 Euclidean distance as the threshold, which is the minimum distance required to a separate cluster. And 5 clusters are obtained.

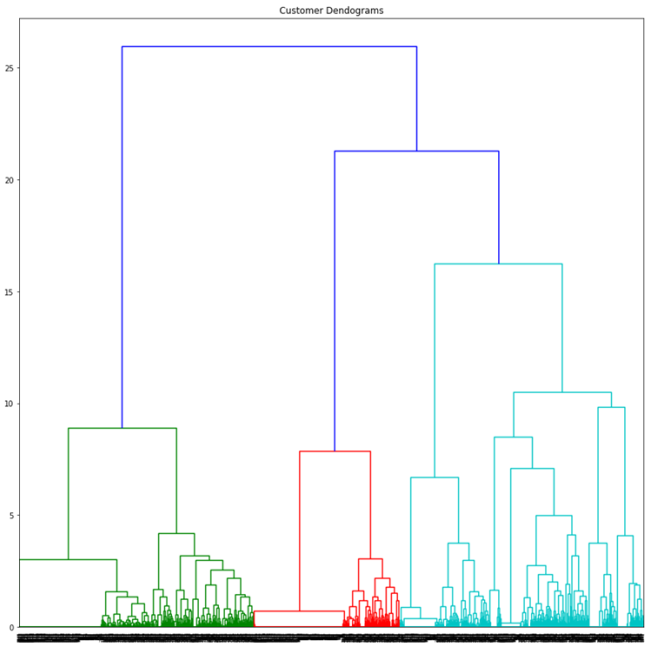

Fig. 2. Dendrograms of the hierarchical clustering result.

Totally, 2353 readers are grouped 5 clusters, denoted by A, B, C, D, and E. The components of department in each cluster are listed in Table III. The largest cluster is the cluster B, which is 885 readers and 38% of total readers. Cluster A is the smallest with 207 readers and 9% of total. The variation of each department in cluster A is slight, and the number of reader in each cluster is close to the average of all departments. But the variation of each department in cluster B and cluster C is significant.

The behaviors for the readers in each cluster are analyzed by comparing the reader's college and department. From the perspective of the number of readers in each cluster, apart from the Department of Recreational Sport in the College of Sport Industry and the Department of Physical Education in the College of Sport Education, most readers in the rest department in the two colleges are identified as in cluster B. The number is almost half of the total number of readers in the department, which is the highest compared to other clusters and significantly higher than in other clusters. The readers in the Department of Recreational Sport are distributed evenly among the cluster B, C, and D, which is between 19% and 25% of the total readers in the department. The numbers of readers in the Department of Physical

Education identified to cluster B and cluster C are the two highest, which is obviously higher than other clusters, accounting for 34% and 26% of total readers in the department. For the College of Sport Performance, the proportion of the total number of readers in the three departments is the highest in cluster C.

TABLE III: COMPOSITION OF THE CLUSTER BY THE NUMBER OF PEOPLE IN THE DEPARTMENT

Department	Cluster Code					
	A	B	C	D	E	*Total*
	College of Sport Performance					
Sport Performance	20	26	77	5	32	*160*
Ball Sport	16	60	144	21	21	*262*
Combat Sport	15	44	43	13	22	*137*
	College of Sport Industry					
Sport Information and Communication	22	130	23	72	12	*259*
Exercise Health Science	35	121	33	8	89	*286*
Sport Management	22	163	21	28	41	*275*
Recreational Sport	20	77	62	87	56	*302*
	College of Sport Education					
Dance	19	119	36	46	20	*240*
Physical Education	38	145	112	57	80	*432*
Total	*207*	*885*	*551*	*337*	*373*	*2353*
Proportion	*9 %*	*38 %*	*23 %*	*14 %*	*16 %*	*100%*

TABLE IV: NUMBER OF BOOKS BORROWING FOR EACH CLUSTER

Main Classes of Books	Cluster Code					
	A	B	C	D	E	*total*
000	22	58	29	2	46	157
100	500	668	124	123	528	1943
200	13	42	4	19	27	105
300	490	551	108	56	401	1606
400	149	1598	336	223	2261	4567
500	321	1096	2325	381	912	5035
600	17	180	33	18	273	521
700	49	736	145	143	906	1979
800	205	10171	499	1088	733	12696
900	187	855	75	3808	236	5161
total	1953	15955	3678	5861	6323	33770

The behaviors for the readers in each cluster are further analyzed by comparison the books categories they borrowed. Table IV show the statistic results. Except for cluster A is difficult to observe which categories of book borrowings are more prominent, the other four clusters can clearly find that one of the books categories has a much higher borrowing frequency than other categories. It can be seen that the readers in cluster B borrow more linguistics and literature books (main class 800), about 6.3 times the second highest main class in the same cluster. Similarly, the readers in cluster C most preferred social science (main class 500), in cluster D most preferred arts (main class 900), and in cluster E most preferred applied science (main class 400). The characteristic of books borrowing behavior is significant in each cluster. The highest frequency of book borrowing main class are 6.3, 4.7, 3.5, and 2.5 times the second highest main class in the same cluster for cluster B, C, D, and E, respectively.

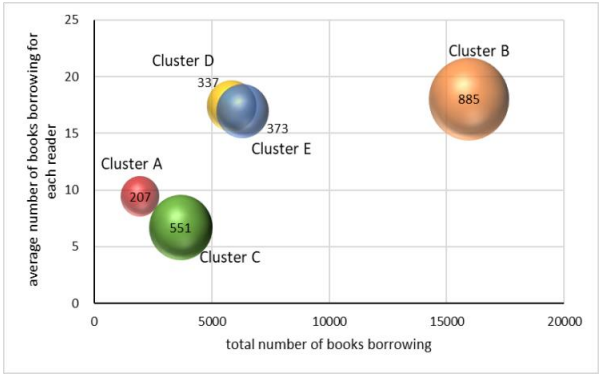

Fig. 3. Comparison of the number of reader and book borrowing in each cluster.

Fig. 3 shows the number of people in each cluster with the behavior of borrowing as a bubble chart for comparison. In Fig. 3, the x-axis is the total number of borrowed books in the cluster, and the y-axis is the average number of borrowed books per person in the cluster. The circle (i.e. bubble) in the figure represents the cluster, and its size implies the number of people in the cluster.

The location of the cluster in the chart can identify the degree of demand for library resources. The lower left corner of the chart indicates that lightweight users with fewer library resource demands. The clusters located in this corner are A and C, and the average number of borrowed books is less than 10. The upper right corner indicates heavy users who need more library resources. These clusters are D, E, and B, and the average number of borrowed books of them are between 17 and 18, which is significantly higher than that of light users.

Based on the results of the cluster analysis and the above discussion, we try to outline the typical type of readers of the University of Sport. There are five typical types described as follows.

The users in Cluster B are the most common type, representing a group of non-athletes, with the largest number of users. They use the library most frequently, and the reading preference is the main class 800 (Linguistics and Literature). It also implies most users like to read the literature and history collections. Such collections are usually novels, essays, literary works, etc. A large number of literary works and the fast reading speed may lead to the highest books borrowed rate. The behavior of such users are speculated to be casual reading

Cluster C is a typical representative of athletes. Overall, they borrow the least amount of books, but the types of borrowing books are especially focused on the main class 500 (Social Science). This main class coverage sports teaching, training method books and the majority of sports-related books. How to improve competitive performance is inferred as the main reading purposes for this type of user.

Cluster D, which is also a heavy user, has a particular reading preference for the main class 900 (Art) and 800 (Linguistics and Literature). It is notable that the 900 main class also contains many comics, which are also mostly for casual reading purposes.

The user behavior of cluster E is speculated to be for

course learning. Most of these users borrow professional books. The first three main classes of preference are 400 (Applied Science), 500 (Social Science), and 700 (History and Geography of World).

Cluster A can be regarded as a reader for pursuing advanced knowledge. They preferred the main class 100 (Philosophy) and 300 (Natural Science), including the psychology, computer science, and physiology.

IV. Conclusion and Suggestion

The university library has the functions of collecting, organizing and using various resources to support teaching and research. Students use the university library resource for the purpose of learning. This study analyzed the utilization of library resources to understand the reader's behaviors. The statistical analysis results of borrowing records show that the most frequently borrowed books categories are literature and linguistics. Although the research object is a sports library, most of the books are related to sports. But the books that are most often borrowed are not books of sports. The book landing ranking of the sport performance college show the topic related to sports psychology and sport instruction is most popular, which is more representative of the characteristics of sports. Therefore, we can infer that athletes use the library's purpose to assist in training and improve athletic performance and build a good player mentality. The students of the rest colleges rely on the function of the library to be similar to the comprehensive university. Textbooks and casual books are most often borrowed. The leisure function of the university library is very important for students. Students use the library to read literature as a priority, followed by books related to professional courses.

In the future, the direction of school purchase should be based on language and literature books to increase students' willingness to use the library.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Yu-Chia Hsu conceived of the presented idea, designed the computational framework, interpreted the results, and wrote the manuscript with input from all authors. Yung-Che Li and Yung-Hsuan Lin performed calculations and analyzed the data.

References

[1] T. Sihachack and L. Yu, "Analysis of user's behavior on borrowed book record in National Central Library University of Laos," *Transactions on Machine Learning and Artificial Intelligence*, vol. 4, no. 5, 2016.

[2] L. Easton, S. Adam, T. Durnan, and L. McLeod, "Identifying and classifying user typologies within a United Kingdom hospital library setting: A case study," *Evidence Based Library and Information Practice*, vol. 11, no. 4, pp. 14–30, 2016.

[3] C. Wang, S. Xu, L. Chen, and X. Chen, "Exposing library data with big data technology: A review," in *Proc. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016, pp. 1–6.

[4] P. Hájek and J. Stejskal, "Library user behavior analysis–use in economics and management," *WSEAS Transactions on Business and Economics*, vol. 11, no. 1, pp. 107–116, 2014.

[5] Y. Cheng and Q. Liu, "Process and application of data mining in the university library," in *Proc. 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, 2019, pp. 123–127.

[6] L. Cano, E. Hein, M. Rada-Orellana, and C. Ortega, "A case study of library data management: A new method to analyze borrowing behavior," in *Proc. Annual International Symposium on Information Management and Big Data*, 2018, pp. 112–120.

[7] T. Jiang, Y. Chi, and H. Gao, "A clickstream data analysis of Chinese academic library OPAC users' information behavior," *Library & Information Science Research*, vol. 39, no. 3, pp. 213–223, 2017.

[8] Y. Ge, "Application of clustering analysis algorithm in digital library," in *Proc. International Conference on Education, Management, Commerce and Society (EMCS-15)*, 2015.

[9] K. Ahmad, Z. JianMing, and M. Rafi, "An analysis of academic librarians competencies and skills for implementation of big data analytics in libraries: A correlational study," *Data Technologies and Applications*, vol. 53, no. 2, pp. 201–216, 2019.

Yu-Chia Hsu is an associate professor of Sports Information and Communication Department, National Taiwan University of Sport in Taichung, Taiwan. He has also served as the director of the Office of Library and Information Services. His research interest is computational intelligence and sports information management, especially applied in sports analyzing and forecasting.

Yung-Che Li received his bachelor's degree from National Taiwan University of Sport, Taiwan. He majored in sport information and communication and interested in computer science. He is currently a master student of Computer Science Department, Tung-Hai University, Taiwan. His research interests include user behavior and data analysis.

Yung-Hsuan Lin was born in Taichung, Taiwan. He is a student in Taiwan University of Sport since 2015 and still learning data mining and database management.