Data Mining Technique to Analysis of Student Library Usage Behavior using Apriori Algorithm

Tewa Promnuchanont
Department of Business Information System,
Faculty of Business Administration and Liberal Arts,
Rajamangala University of Technology Lanna,
Chiang Mai, Thailand

Rujipan Kosarat
Department of Electronic Engineering and
Automation System, Faculty of Engineering,
Rajamangala University of Technology Lanna,
Chiang Mai, Thailand

ABSTRACT
The purpose of this research was to analyze student library access behavior by using association rule data mining. The data about students' access to library services is based on four factors: the number of times the service was accessed, total time of service, number of borrowed books and the cumulative grade point average of students. There are 84,977 data sets used in the experiment. The research method is divided into three steps: pre-processing or data preparation, data selection, data modification and data completion. Then, the data mining process was used to find association rules using an Apriori algorithm to analyze library service behavior that affects student grade point averages. Finally, post-processing was done to take the knowledge base obtained from the data mining process, to test and determine if it is correct or not. The results showed that students who had a high frequency of using the library, spent time in the library, and had a high frequency of borrowing books, had a good average score. Students who never borrowed library books, attended libraries less often, and spent less time in libraries had lower GPAs. It can be concluded that library use behavior affects students' academic performance.

Keywords
Data mining, Association rule, Apriori

1. INTRODUCTION
Higher education plays an important role in the development of human resources in the country. Learning to keep up with technology and expanding knowledge is essential for learning in the 21st century. Therefore, universities need to have learning resources that will help increase the potential and quality of students. With teaching in all faculties in a university it is necessary to rely on books which are the basis of that study and research. As a result, a university's library is an important resource for learning. There are various information resources in a university library such as books, journals, newspapers, databases, and electronic media. Moreover, a university's library also has a system for providing services to library users such as information technology, computers, media as well as searching tools. As a result, university library is filled with a great deal of information available both for university staff and public users.

In the Library of Rajamangala University of Technology Lanna **(**RMUTL**)** Chiang Mai, the Library Automation System **(**OPAC**)** has been set up to offer various services for the library users such as the borrowing of books and for searching for academic journals via the internet network**.** It also stores information related to the services of the library such as the frequency of each borrowed item, types of library users or borrowers, and the period of borrowing or on-loan period. The library staff can analyze this data for making decisions, such as predicting the demand for books in advance and allocating a budget for book purchasing with reference to the needs of users**.** The automatic library system of RMUTL Chiang Mai is the web application system of WALAI AutoLib Ultimate, which gathers details about the various services available from the library in the database**.** As a result, the database inevitably contains a large amount of data. However, at present, this information has rarely been used. As a result, this study aimed to process the information by adopting a Data Mining method. The Big data will be analyzed and extracted as a knowledge base for further use in various fields.

Currently, data mining relies on conceptualization from the CRISP-DM data mining model for the extraction of knowledge or for developing correlation rules between related or relative components. In this article, the knowledge gained from the data CRISP- DM Model has been applied so as to improve the library's information resource lending service. The steps of data mining are shown in Fig 1 [1].

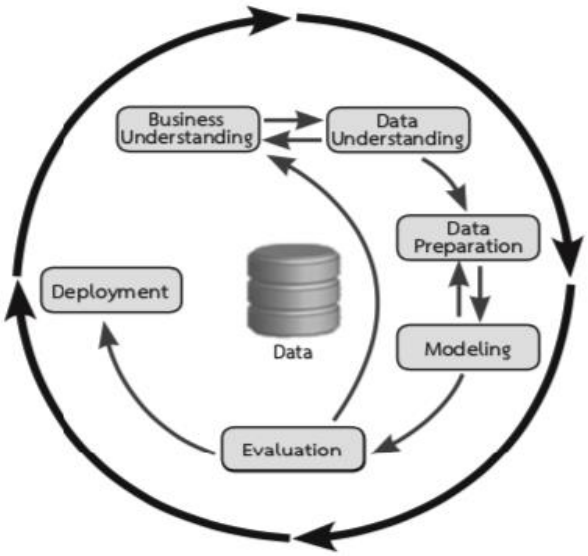

Fig 1: The Data Mining Life Cycle

The procedure for the CRISP- DM Model is as follows: The initial phase focuses on businesses that understand the objectives and requirements of the project from a business perspective. This knowledge is translated into data mining problem definitions and initial project plans designed to achieve objectives. The data understanding process begins purchasing with the collection of initial data and performance activities to familiarize people with the data, identify any data quality issues, discover the first insights in the data or detect

interesting fragments so as to make hypotheses about hidden data. The data preparation phase covers all the activities needed to build the final dataset (data that will be introduced into the modelling tools from the original raw data). Data preparation tasks are likely to be performed multiple times rather than in a prescribed sequence. Tasks include selecting tables, records and attributes, cleaning data, building new attributes, and converting data for modeling tools. In the subsequent phase, various modelling techniques are selected and applied, and their parameters are calibrated to optimum values. In general, there are multiple techniques for solving the same type of data mining issue. There are techniques that require specific data formats. Subsequently, at this stage of the project, a researcher will have constructed one or more models that seem to be of high quality from a data analysis perspective. Prior to finalizing the implementation of the model, it is important to further assess the model and review the steps taken to build it to ensure it is meeting business objectives correctly. The creation of the model is generally not the end of the project. Generally, the knowledge acquired will need to be organized and presented so that the client can use it. Depending on the requirements, the deployment phase can be as straightforward as generating a report or as complex as the implementation of a reproducible data mining process. In many cases, the deployment steps will be performed by the user, not the data analyst.

The retrieval of association rules is one of the data retrieval techniques that should be very useful in applications. Association rules are required to assure minimum support and minimum confidence at the same time. Association rule generation consists of two steps: firstly, minimum support is applied to a given set of item elements. Second, using minimum confidence, and frequent item sets, rules are formed [2].

Association Rules will provide information about rules such as: If A then B, where A and B can be particular items, values, words, etc. An association rule is composed of two item sets: Antecedent or Left-Hand Side (LHS) and Consequent or Right-Hand Side (RHS). It describes the relationship between support, confidence, and interestingness. Support and confidence are usually called measures of interest of an association rule. Association rule mining is the process of finding all the association rules with the condition of minimum support and minimum confidence. Initially, the support and confidence values are computed for all the rules and it is then compared with the threshold values to prune with a low value of support or confidence Association rules mining was proposed by Agarwal. Numerous algorithms for the generation of association rules have been presented in time. Some of the popularly known algorithms are Apriori [3], Eclat, and FP-Growth [4], which are used to mine frequent item sets. Mining takes advantage of infrequent data and very low support and confidence values. Still, it always produces an enormous amount of rules [5].

This paper uses association rules to analysis of library service usage behavior that affects student grade point averages by data mining using the association rule method using the Apriori algorithm. This work is organized according to the following: Section 2 reviews some work directly related to this research, and Section 3 outlines the methodology for this work. Section 4 introduces and discusses the results and Section 5 concludes the paper.

2. RELATED WORKS

Most of the time, when it comes to libraries, people tend to think of libraries in schools or universities because libraries are the core of a drive toward knowledge**.** Especially in today's era, the knowledge of science, technology, and information media has changed rapidly and progressed unprecedentedly **[**6**].** At present, libraries are not only located in educational institutions, but are also present in various departments both public and private**.** The important function of a library is to provide loans and return service for information resources**.** Therefore, it is not uncommon for libraries to deal with relevant information, such as data on borrowing and returning services of information resources including books, journals, papers, electronic research, procurements, and information from many service recipients **[**7**].** This research examines the needs of library service users with information that is often recorded as a database in an information system, which is one of the important areas of information science research**.** It allows library administrators to formulate strategies and approaches to optimize the library's use of information resources **[**8**].**

Many research papers focus on data mining in several other methods, such as Association rule discovery [9], classification [10], prediction [11], database clustering [12], segmentation [13], etc.

Zhenyi Zhao, et al., [14] said that the information available today is the advancement of information technology with the increasing complexity of data and time to process which are the main factors found in the data processing. With the rapid development of database technology, there has been such an increasing amount of storage space available that data mining technology has become important. The data mining techniques used for this paper apply to the analysis of data mining techniques using the Apriori algorithm for large data sets.

Yingwei Zhou [15] proposed a method that uses the traditional Apriori algorithm for the book management system, to slow down the system, due to frequent database scans and an excessive set of options. Therefore, the guidebook management system based on the improved Apriori Data Mining algorithm is designed by C/S (Client/Server) architecture and B/S (Browsers/servers) architecture, which are combined to open up all available book information for library staff and borrowers. The relevant information about the book borrower can be extracted from the book lending database using the preprocessing sub-module in the system function module. This article is about cleaning, transforming, and merging the data into the correct form and then using the Apriori data mining. algorithm. The results showed that the system was able to recommend book-related information effectively, and the CPU utilization rate was only 6.47% under the condition that 50 concurrent customers were in a good performance.

Eni Heni Hermaliani et al., [16] proposed a basic method for computation of association analysis with the apriori algorithm. It is used to process the best-selling products and help in deciding which products will be published in the sales inventory. The results of this study provide a model to compute a high frequency model analysis. The formation of the association rules and the established model resulting from the retail sale of fruit products with the greatest support and confidence in the best-selling fruit products.

Phatthaphong Pongphatthanakan et al., [17] proposed a method that uses the C5.0 algorithm, Neural Network, and CART to analyze student library usage factors. It does this by using information about accessing services through automatic gates with 9 fundamental factors: access date, time, gender, faculty, year, province, date of birth, blood type and scores. A total of 79,953 samples were used in the experiment. After that, the C5.0 algorithm, Neural Network, and CART were processed to study and compare the efficiency of data extraction. The results

showed that the C5.0 algorithm yielded 97.78% accuracy.

Sandhya Harikumar and Divya Usha Dilipkumar [21] offer a variant of the Apriori algorithm using the QR decomposition concept to reduce dimensions thus reducing the complexity of the traditional Apriori algorithm.

3. METHODOLOGY

The main program of the proposed algorithm is shown in Fig 2.

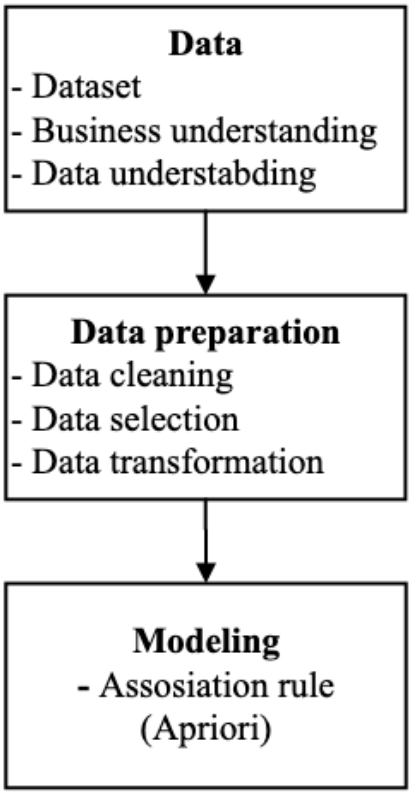

Fig 2: Main program of the proposed algorithm

3.1 Dataset

The data set used in this study was obtained from the Library of RMUTL Chiang Mai, Thailand. There is a total of 84,977 samples of data for borrowed books from libraries between 2019 and 2021.

3.2 Data Preparation

The steps of the data mining process are as follows:

3.2.1 Data cleaning

The dataset is raw data. Therefore, the data needs to be cleaned by parsing, correcting, standardizing, and duplicate elimination.

3.2.2 Data selection

The data is generated by the various reports on library book borrowing in RMUTL Chiang Mai. The initial data contains the borrowing of books from students or academic staff from 2019 to 2021 with 15 listed attributes, which include member codes, member's name, faculty, department, barcode number, book title, author, borrowing date, return date, and borrowed days. The data includes different types of values, either chain values or numeric values. The target is shown as an analytical report. The analytical report was grouped into three categories (good, average and poor). The data was then processed to generate rules.

3.2.3 Data transformation

This research focuses on analyzing the library access behavior of undergraduate students by using a correlation rule data mining technique using the Apriori algorithm. This research uses the data set of access to the library services of RMUTL Chiang Mai with data attributes as shown in Table 1.

Table 1. Data attribute

No.	Attribute
1	Student_id
2	Gender
3	Service_date
4	Period
5	Faculty
6	Program
7	Service_count
8	Borrow_count
9	All_time
10	GPA

The attributes in Table 1 consist of student ID, gender, date of use, period, faculty, field of study, number of times used. number of times borrowed Total time spent using the service and GPA.

3.3 Association rule

The association rules using the Apriori algorithm discussed in section 3 were applied to generate rules as shown in Fig 3.

L_1= {frequent items};
for (k= 2; L_{k-1} !=Ø; k++) do begin
C_k= candidates generated from L_{k-1}
for each transaction t in database do
The count that is enclosed in t of all candidates in C_k is to be incremented
L_k = candidates in C_k with min_sup
end
return $\cup_k L_k$

Fig 3: The association rules using the Apriori algorithm

The first passage of the algorithm simply counts element occurrences to determine the major 1-item sets. A successive pass k contains two phases: Initially, the large item sets L_{k-1} found in the $(k-1)^{th}$ pass are used to generate the candidate item sets C_k. Then, the database is scanned and the support of candidates in C_k is counted. It is necessary to determine the candidates in C_k for quick counting that are contained in a given transaction t.

The workflow is to optimize the data in a format that is usable with the association rules by reducing the data size. With Big data it takes a lot of time to find patterns. Therefore, the data should be reduced to manageable proportions and the resulting format should remain the same. If this is done it will take less time to find patterns with interesting attributes and to find data patterns: The number of visits to the library (Service_count) was divided into high library attendance (SH), average library attendance (SM) and less than average library attendance (SL). The number of times a book was borrowed (Borrow_count) was divided into borrowed above average (BH), borrowed equal to average (BM), borrowed less than average (BL), and never borrowed (BN). The total time spent in the library (All_Time), the cumulative grade point average (GPA). The total time in the library (All_Time) is divided into total time

above average (AH), total time equal to mean (AM), and total time less than average (AL). Cumulative GPA (GPA) is divided into high GPA 3.00 - 4.00 (GG) and Low GPA 0.00 - 2.99 (GL) shown in table 2.

Table 2. Data attribute

Itemset	The attribute name
SH	High library attendance
SM	Mediocre library attendance
SL	Low library attendance
BH	Borrowed above average
BM	Borrowed equal to average
BL	Borrowed less than average
BN	Never borrowed
AH	Total time above average
AM	Total time equal average
AL	Total time less than average
GG	High GPA 3.00 - 4.00
GL	Low GPA 0.00 - 2.99

In this paper, the association rule using the Apriori algorithm using WEKA can be measured as follows: Support is the percentage of the amount of data with corresponding members as a rule per number of Total data. Confidence is the percentage of the amount of data that conforms to the law for the total number of left- side rule- based members. The associative rule is created using an Apriori algorithm, set to the following parameters: Class association rule is true, minimum support is 0.1, minimum confidence is 0.5, and rule is 10.

4. RESUTLS

The dataset used in the present study was obtained from The Library of Rajamangala University of Technology Lanna, Chiang Mai, Thailand. There is a total of 84,977 samples of data on library book borrowing from 2019 to 2021.

An analysis of library service usage behavior that affects average student grade points by the use of data mining using the association rule method, that uses the Apriori algorithm, for all 10 rules as shown in table 3.

Table 3. The result of using the Apriori algorithm for association rules.

No.	Attribute	Confidence (%)
1	BH, GG	92.0
2	SH, AH, GG	83.0
3	AH, GG	78.0
4	SH, GG	75.0
5	NB, GL	55.0
6	NB, SL, GL	54.0
7	NB, AL	54.0
8	SL, AM	52.0
9	SL, GL	51.0
10	NB, SL, AL	50.0

Form Table 3 shows the result of using the Apriori algorithm for association rules.

1) Students who borrowed high numbers of library books (BH) are students with good grades (GG) at the confidence level of 92%.

2) Students who use library services often (SH) when they are in libraries (AH) are students with good grades (GG) at 83% confidence level.

3) Students who spend a lot of time in libraries (AH) are students with good grades (GG) at 78% confidence level.

4) Students with high levels of access to library services (SH) are students with good grades (GG) at the confidence level of 75%.

5) Students who have never borrowed library books (NB) are those with low grades (GL) at the 55% confidence level.

6) Students who have never borrowed library books (NB) and have low levels of access to library services (SL) are students with low grades (GL) at 54% confidence level.

7) Students who have never borrowed a library book (NB) or spent any time in the library (AL) are low-performing students at 54% confidence level.

8) Students with low access to library services (SL) and low amounts of time spent in the library (AM) are low-performing students at a 52% confidence level.

9) Students with low access to library services (SL) were students with low grades (GL) at a 51% confidence level.

10) Students who have never borrowed a library book (NB), have low levels of access to the library (SL), and have spent little time in the library (AL) are low-performing students at a 60% confidence level.

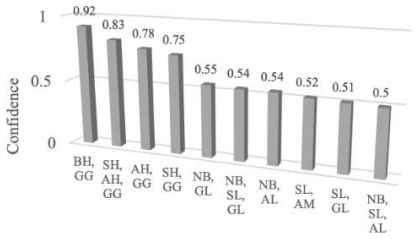

Fig 4: The result of using the Apriori algorithm

The experimental results show that most students who use library services, spend a lot of time in the library and borrow library resources have good grades. For students who have less access to library services in a variety of areas, their academic achievement is low grades. But not always, the two cases mentioned may depend on each person's character, environment, learning resources, and a high IQ, or EQ of that person, causing the evaluation of the experiment to be inaccurate.

5. CONCLUSION

From the findings of the association rules, using an Apriori algorithm to analyze library service behavior that affects students' grade point averages, two correlation rules can be observed: 1) Students with high frequency of using the library, spending time in the library, and high frequency of borrowing books were students with good GPAs. 2) Students who never borrowed library books, had low library access levels and spent little time in the library were students with low GPAs. From the results of this research, it can be concluded that library access behavior affects student performance, so universities should think of new activities to encourage students with low GPAs to use libraries. This encourages students to pursue knowledge for continuous self-improvement.

The extraction of association rules from a large data base is increasingly becoming resource intensive. In addition, the result seen form use of the Apriori algorithm of the association rules and the better Apriori exclusion algorithm have shown that the algorithm is not simple. It also reduces the number of

candidates for frequent entries and has the benefit of quick search speeds. This not only saves processing costs but also improves algorithm efficiency. The results of the improvement of the Apriori mining algorithm show that it is not only simpler, but also more effective than existing techniques.

6. REFERENCES

[1] Wirth, R., and Hipp, J. 2000. CRISP-DM: Towards a Standard Process Model for Data Mining. Practical application of knowledge discovery and data mining.

[2] Arora1, J., Bhalla, N., and Rao, S., 2013. A Review on Association Rule Mining Algorithm. International Journal of Innovative Research in Computer and Communication Engineering.

[3] Sornalakshmi, M., Balamurali1, S., et al., 2020. Hybrid method for mining rules based on enhanced Apriori algorithm with sequential minimal optimization in healthcare industry. Neural Computing and Applications.

[4] Jiawei, H., Jian, P., and Yiwen, Y., 2000. Mining frequent patterns without candidate generation. ACM SIGMOD Record.

[5] Magdalene Delighta Angeline, D., 2013. Association Rule Generation for Student Performance Analysis using Apriori Algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA).

[6] Liu, Y., 2018. Data Mining of University Library Management Based on Improved Collaborative Filtering Association Rules Algorithm. Wireless Pers Commun, Springer Science+Business Media.

[7] Noppon, L., 2017. Modification of Data Mining Technique for Better Understanding of Book-Loan Behaviors. Journal of Rajanagarindra.

[8] Zhu, Z., and WANG, J., 2007. Book Recommendation Service by Improved Association Rule Mining Algorithm. Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong.

[9] Shweta, M., and Garg, K., 2013, Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms. International Journal of Advanced Research in Computer Science and Software Engineering.

[10] Srinivas1, B., Ramesh, G., and Sriramoju, S., 2018. An Overview of Classification Rule and Association Rule Mining. International Journal of Scientific Research in Computer Science, Engineering and Information Technology.

[11] Shao, Y., et al., 2018. A novel software defect prediction based on atomic class-association rule mining. Expert Systems With Applications.

[12] Chuan Tan, S., 2018. Improving Association Rule Mining Using Clustering-based Discretization of Numerical Data. International Conference on Intelligent and Innovative Computing Applications (ICONIC).

[14] Zhao, Z., et al., 2021. An improved association rule mining algorithm for large data. Journal of Intelligent Systems.

[15] Zhou, Y., 2020. Design and Implementation of Book Recommendation Management System Based on Improved Apriori Algorithm. Intelligent Information Management.

[16] Hermaliani, E, H., et al., 2020. Data Mining Technique to Determine the Pattern of Fruits Sales & Supplies Using Apriori Algorithm. Journal of Physics.

[17] Pathapong. P., et al., 2017. Using Data Mining Techniques for Analyzing Factors that Influence Students' Library Use. PULINET Journal Provincial University Library Network.

[18] Lambodar .J., and Narendra K., 2014. A Model For Prediction Of Human Depression Using Apriori Algorithm. International Conference on Information Technology.

[19] Jena, L., and Kamila. N., 2014. A Model For Prediction Of Human Depression Using Apriori Algorithm. 13th International Conference on Information Technology.

[20] De Choudhury, M., Counts, S., and Horvitz, E., 2013. Major life events and behavioral markers in social media: Case of childbirth. In 16th ACM Conference on Computer Supported Cooperative Work (CSCW).

[21] Sandhya, H., and Divya, U., D., 2016. Apriori algorithm for association rule mining in high dimensional data. In 2016 International Conference on Data Science and Engineering (ICDSE)