ibres volume 28, issue 1, pages 25-33 (2018)

Investigating Factors Affecting Library Visits by University Students Using Data Mining

Wichai Puarungroj, Pathapong Pongpatrakant, Narong Boonsirisumpun, and Suchada Phromkhot
The Office of Academic Resources and Information Technology, and
Faculty of Science and Technology,
Loei Rajabhat University, Thailand
wichai@lru.ac.th; pathapong@lru.ac.th; narong.boo@lru.ac.th; suchada.phr@lru.ac.th

ABSTRACT

Background. Providing appropriate library services to students is a challenging task for university librarians. The library at Loei Rajabhat University has some concerns about its small number of visitors. The question of "what is known about the situation?" was raised. As an attempt to answer this question, data mining was employed to gain insights into library and student data.
Objectives. This study used two data mining algorithms—Naïve Bayes and C4.5 decision tree induction—to analyze the data. The results of the data mining were intended to be used in promoting undergraduate students to physically visit the library.
Methods. Data include students' library gate entry collected from the library database and student data collected from the university registrar's office.
Results. The data mining yielded interesting results. Senior students were found to use the library less than younger students. There were two faculties whose students come to the library less than 50%. Current GPA was found to be an influential attribute for predicting library visit.
Contributions. The research identified useful student attributes for predicting library visit. The results of the data mining can be used to increase the rate of library use by organizing activities that target those attributes. For example, the library can collaborate with the instructors to organize programs for students with low GPA.

INTRODUCTION

The university library is the place where university faculty and students visit to read and search for academic resources relevant to their study and research (Kim, 2017; Teoh & Tan, 2011). In addition to services such as loan of books and other library materials, interlibrary loan and information desk, some libraries also provide special services such as research support, citation consulting, subject guide, media service, workshop, and thesis format consulting. By doing this, the library plays an important role in student education and retention. Clink (2015) proposed that the library should be part of the university's effort in increasing student retention by addressing certain student needs. This can be done by offering space, collection and personnel to support students towards the successful completion of the degree program.

Published by WKW School of Communication & Information & NTU Libraries,
Nanyang Technological University

Since academic libraries play a crucial role in supporting academic staff research and student education and retention, the libraries need to have strategies for attracting students to use the library services. However, it is difficult to develop a strategy without understanding the current conditions and situations of the library. The Loei Rajabhat University library faced the situation where programmes offered were not attracting much student participation. There were fewer visitors than the library capacity and space could accommodate. Many attempts were made to attract student attention to the library, such as organizing book fairs, arranging reading activities, providing library orientation courses, and other activities. To improve library services, Renaud, Wang, and Ogihara (2015) suggested that the library needs to gain insights from the collected user data to understand library use patterns before making further efforts on library programs. To do this, library user data and other associated data should be analyzed using data mining algorithms (Renaud *et al.*, 2015; Siguenza-Guzman *et al*., 2015). Leetaru (2015) mentioned that data mining provides opportunities for libraries to expedite and understand the useful knowledge existed in the library data.

This study carried out an analysis of library user data using data mining algorithms to discover hidden patterns, especially the user attributes that determined library visit.

LITERATURE REVIEW

This section reviews the research literature regarding the importance of library use and two data mining algorithms that are used in this research, namely Naïve Bayes and a decision tree induction algorithm called C4.5.

Library Use and Relevant Research

It has long been accepted that a university library grows along with the university. The library nowadays has further responsibility in supporting the university's goal related to student completion of his or her degree program (Clink, 2015; Kim, 2017). Many universities find it important to involve the library in their student retention effort (Haddow, 2013; Soria, Fransen, & Nackerud, 2013). Prior research into the relationship between library use and student retention has found a strong correlation between them. Murray, Ireland and Hackathorn (2016), who conducted a study on the association between library services and student retention with 3,757 freshmen and sophomores at a public university in US, found that library services, especially checking items out and using electronic library resources, were positive predictors for student retention. They suggested that apart from using electronic resources, students who physically visited the library were associated with a low drop-out rate.

Soria, Fransen, and Nackerud (2013) investigated the impact of library use on undergraduate student retention by conducting research with 5,368 freshmen at the University of Minnesota-Twin Cities. The data were collected from various sources, including library use such as library database access, website login, borrowing and returning books, using computer workstations, attending library workshops, and in-person interaction. The results of data analysis indicated that students who were library users showed higher competency in their studies than those who were non-library users. Moreover, the library users showed a higher returning rate after the end of the first semester compared to non-library users. Likewise, Haddow (2013) analyzed undergraduate student data and library use data to find an association between library use and student retention. Her research revealed that students with book check-out or use of electronic resource records had higher retention rates than those without. She also found that senior students had higher drop-out rate than younger students.

Kuh and Gonyea (2015) also stressed the importance of the library in enhancing students' performance and their ability to deal with assiduous tasks.

Data mining research in library has also been conducted. Renaud *et al*., (2015) examined library and university data in order to make sense of library use patterns, student behavior, and the correlation between student achievement and library use. They analyzed student library use as reflected in the percentage of book check-out activity, and found that humanities students (especially students from English department) used the printed collection more than students in other faculties. They also found that library use had a positive relationship with senior students' performance, but less likely to have a correlation with that of freshmen. In order to discover factors that influence library use, classification data mining algorithms can be applied (Siguenza-Guzman *et al*., 2015). The classification algorithms have been employed for extracting important information from the dataset and have been used in various fields, especially education (Romero & Ventura, 2010). Data mining has become common for discovering patterns in students' data and predicting their performance. Examples of classification algorithms are decision trees (e.g., ID3, C4.5), Support Vector Machine (SVM), Naïve Bayes, and Neural Networks (Romero & Ventura, 2010; Siguenza-Guzman *et al*., 2015). These algorithms have been tested widely by prior research for classifying important attributes and estimating the values of the targeted variables (e.g., Bravo & Ortigosa, 2009; Puarungroj *et al*., 2017). The selection of the most suitable algorithm for a given dataset can be determined by its predictive accuracy.

C4.5

A C4.5 is one of decision tree algorithms, which has been accepted as a promising and reliable data mining algorithm (Ahlemeyer-Stubbe & Coleman, 2014). It is an improved version of the ID3 algorithm, which was widely employed in data mining. C4.5 is a predictive model that arranges a given dataset into a decision tree (Kohavi & Quinlan, 1999). The tree is constructed by determining the splitting points of data attributes, which is done by measuring data attributes' purity (Rokach & Maimon, 2010). The calculation of purity has to be carried out in order to reduce uncertainty in selecting proper attributes as nodes of the tree. C4.5 uses information gain for measuring purity (Han, Kamber & Pei, 2012). After the tree has been constructed, it can overfit the dataset. Therefore, the C4.5 is needed to be pruned back until it can be predictive in general (Han, Kamber & Pei, 2012).

Naïve Bayes

Naïve Bayes is a powerful classification algorithm for predictive modeling (Han, Kamber & Pei, 2012; Larose & Larose, 2015). It is a probabilistic classifier based on Bayes' theorem. The classification takes the assumption that the attributes, which are used to predict the possible value of the class label (the output attribute), are independent and uncorrelated (Allahyari *et al*, 2017; Larose & Larose, 2015). The term *naïve* represents this independent assumption. The predicting value of the class label is carried out by using Bayes' theorem to calculate its probabilities based on prior knowledge of conditions. This algorithm is effective for classifying data in different fields. For example, it has been applied widely to predict student performance in educational research (Romero & Ventura, 2010) and it is also applied for extracting information from structured (e.g., RDBMS data), semi-structured (e.g., XML and JSON), and unstructured text resources (e.g., word documents) in a text mining research field (Allahyari *et al*., 2017).

Table 1. Attribute values conversion in the dataset

Attribute	Values	Converted Values
Year of Study	Freshman – 1st	1st
	Sophomore – 2nd	2nd
	Junior – 3rd	3rd
	Senior – 4th	4th
Sex	Female	F
	Male	M
Faculty	Education	ED
	Science and Technology	SC
	Humanities and Social Science	HU
	Management Science	MN
	Industrial Technology	IN
Major	There were 68 majors	Three digit codes assigned
Current GPA	Used real values	-
Visit Library	Never visit	0
	Visit	1

RESEARCH METHOD

There are four key steps in the data mining process:

1. *Data Selection*. Library gate entry data were collected from the library database, which recorded students who used their library card to enter the library. The data collection was conducted based on the ethical guidelines for educational research (British Educational Research Association, 2011). The data were anonymized and confidentially used for research as a whole without identifying anyone individually. The data were selected in one academic year starting from 1st August 2016 to 31st May 2017, which comprised 150,387 records. Furthermore, student data were collected from the university registrar's office database and integrated with the library gate entry data. The student data were limited to students who entered the university between 2013 and 2017. There were 13,352 records of students in the university.
2. *Data Pre-processing*. Data preparation was carried out in two steps. Firstly, library gate entry data were cleaned by deleting some records that were not relevant to the topic, such as records of external users, postgraduate students, faculty members, and staff. The exit gate records were also deleted. Records of students with GPA of 0.00 were also deleted, as it meant that these students would drop-out without completing any semester of study. There were 1,943 records of students with GPA 0.00. The student data were finally reduced to 11,409 records. There were 5 attributes pulled from the student database including year of study, sex, faculty, major, and current GPA as shown in Table 1. Secondly, both library gate entry data and student data were integrated. The data were then converted to a format appropriated for the Weka data mining software.
3. *Data Mining.* This step employed two data mining algorithms, namely Naïve Bayes and C4.5. The predictive models were built and the predictive accuracy of the two algorithms were estimated and compared. The data analysis and visualization were done using the Weka data mining software.

Table 2. Attributes in the dataset

Attribute	Status
Year of Study	Input
Sex	Input
Faculty	Input
Major	Input
Current GPA	Input
Visit Library	Output

Table 3. Number of students who visited or did not visit the library

	Number of Students	Percentage
Visited the library	6,753	59%
Did not visit the library	4,656	41%

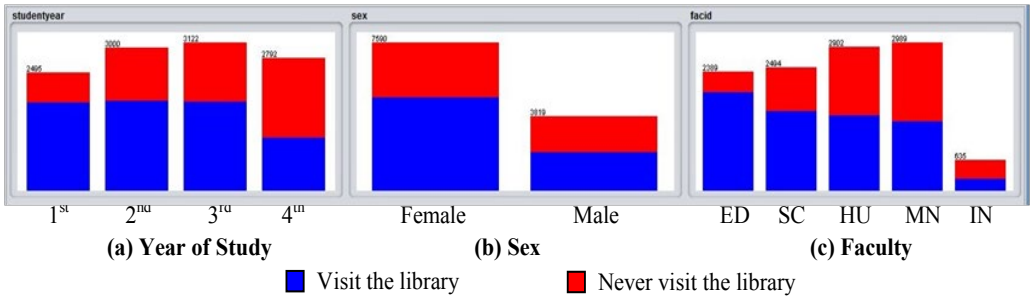

Figure 1. Number of students who visited or did not visit the library classified by year of study, sex, and faculty affiliated

4. *Data Interpretation.* This step attempted to interpret the data mining output. The expected outcome was to identify the attributes of students who have visited the library at least once during the academic year. This can be used to suggest to the librarians what to do to attract the interest of students and persuade them of the benefits of using the library to support their studies until completing their degree programs.

DATA ANALYSIS AND RESULTS

The process of data cleaning and integration resulted in 11,409 data records. After filtering, the most relevant attributes were selected for the analysis. Predictive models (i.e. classifiers) were developed to classify students into two categories—whether they came or did not come to the library—using the attributes listed in Table 2.

From the total of 11,409 students in the dataset as shown in Table 3, 6,753 (59%) visited the library at least once during the academic year, while the rest (41%) never did. The proportion of students who visited/did not visit the library for each attribute category is shown in Figure 1-2.

Figure 1(a) shows that 75% of freshmen, 63% of sophomores and 60% of juniors visited the library. Interestingly, only 40% of seniors visited the library, which seemed too low for students who were preparing for their future jobs. Figure 1(b) shows that more female students used the library than male students: 63% of female students used library compared to 52% of male students. Figure 1(c) shows that students from the Faculty of Education (ED) had the highest rate of library visit (83%), while students from the Faculty of

Figure 2. Number of students who visited or did not visit the library classified by majors of study in each faculty

Science and Technology (SC) came in second (64%). About half of the students (52%) from the Faculty of Humanities and Social Science (HU) came to the library. At the same time, students from the Faculty of Management Science (MN) and the Faculty of Industrial Technology (IN) have low rates of library visit (47% and 39% respectively).

Figure 2 shows the number of students classified by 68 majors of study. Figure 2(a) indicates that in the Faculty of Education, most of the study majors with a higher number of enrollment had high rates of library visit. In particular, almost all the students in Thai Studies (99%), Physics (97%), and Chemistry (96%) visited the library. Figure 2(b) depicts library visit rates by students from the Faculty of Science and Technology. Most of the students in Chemistry (83%) visited the library, whereas only a small number of Agricultural students (39%) went to the library. Figure 2(c) shows the library visit rates for students from the Faculty of Humanities and Social Science. The highest use was with English major students. Interestingly, students from Social Development showed no sign of library visit. In Figure 2(d), some majors in the Faculty of Management Science show rates of library visit near to 70%: Finance, and Logistics. However, there are some majors that had very low rates of library visit: General Management (0%) and Management (9%). Figure 2(e) shows library

Table 4. Predictive Accuracy of Data Mining

Algorithm	Accuracy
Naïve Bayes	76%
C4.5	80%

Table 5. Attribute Importance for Prediction

Attribute	Importance
Current GPA	58%
Major	21%
Year of Study	16%
Faculty	5%

visit rates for students from Industrial Technology: most of the majors had library visit rates of less than 50%, except for Civil Technology (59%).

Data mining algorithms were employed to analyze the dataset using the Weka data mining software. The prediction accuracy of Naïve Bayes and C4.5 is shown in Table 4. It can be seen that C4.5 with the accuracy level of 80% performed better than that of Naïve Bayes (76%). Therefore, C4.5 was chosen for analyzing the dataset.

The dataset was then analyzed using C4.5 to evaluate the attributes that have predictive power towards the students' library visit. Table 5 shows the attribute importance that had a strong effect on student visiting the library. As shown in Table 5, the result indicates that four attributes: Current GPA, Major, Year of Study, and Faculty were important for predicting students' having visited the library, while other attributes had no effect on this prediction.

CONCLUSION

In order to provide the right services to students, the university library needs to learn about its current situation before moving forward. Some authors have suggested that one of the key roles of academic libraries is to support its university's goal in increasing the rate of student retention (Clink, 2015; Haddow, 2013; Soria, Fransen, & Nackerud, 2013). The library at Loei Rajabhat University faced a question of how to attract its students to visit the library. This question was raised when there were too few visitors than its capacity can accommodate. To answer this question, there was a need for data analysis. To do so, library and student data were analyzed using two data mining algorithms. The data mining yields some interesting results with useful implications. Firstly, fourth-year students visited the library less than younger students. Only 40% of fourth-year students visited the library at least once, compared with 75% of first-year students. Secondly, two faculties, whose students used the library less than 50%, were the Faculty of Management Science (47%) and Faculty of Industrial Technology (39%). Finally, the current GPA was important for predicting students' library visit.

The results carry further implications. The university library should pay more attention to the faculties that had the low rates of library visits including the Faculty of Management Science and the Faculty of Industrial Technology. The survey can be made to examine the needs of the students from the two faculties. Since current GPA has shown its association with students' library visit, the library can provide services and create activities that target students with low GPA. To do so, the library can collaborate and work closely with the instructors who take care of students in the targeted majors and faculties (Kuh & Gonyea,

2015). Some activities or competitions with price can be considered to attract more student interest. Teoh and Tan (2011) highlighted that making students familiar with the resources in the library in their earlier year at university can give them a positive attitude towards library and enhance the chance of future use. It is also important to attract more senior students to the library. Since the statistics show that the percentage of fourth-year students visiting the library is low, some programs in the library can focus on preparation for work after graduation, such as foreign language learning support, résumé writing consultation, and applying to graduate school (Bordonaro, 2006). This research demonstrates a possible approach for analyzing library-related data by using data mining methods that should be worthwhile for other libraries. They are able to replicate this protocol with or without additional input attributes. Some attributes may be useful for investigating further such as hours spent in library and number of items borrowed (Collins & Stone, 2014). The future research can focus on the usage of e-resources such as e-books, e-journals accessed by students and academic scholars in the university. This can help librarians improve e-resource services and acquire the right e-resources as needed by their users.

REFERENCES

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). *A practical guide to data mining for business and industry*. West Sussex, UK: John Wiley & Sons.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A brief survey of text mining: Classification, clustering and extraction techniques.* Retrieved from https://arxiv.org/pdf/1707.02919.pdf

Bordonaro, K. (2006). Language learning in the library: An exploratory study of ESL students. *The Journal of Academic Librarianship*, *32*(5), 518-526.

Bravo, J., & Ortigosa, A. (2009). Detecting symptoms of low performance using production rules. In Barnes, T., Desmarais M., Romero, C. & Ventura, S. (Eds.), *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, 1-3 Jul 2009, Cordoba, Spain* (pp. 31-40). Retrieved from http://www.educationaldatamining.org/EDM2009/uploads/proceedings/edm-proceedings-2009.pdf

British Educational Research Association. (2011). *Ethical guidelines for educational research*. London: British Educational Research Association.

Clink, K. (2015). The academic library's role in student retention. *PNLA Quarterly, 80*(1), 20-24.

Collins, E., & Stone, G. (2014). Understanding patterns of library use among undergraduate students from different disciplines. *Evidence Based Library and Information Practice 2014*, *9*(3), 51-67.

Haddow, G. (2013). Academic library use and student retention: A quantitative analysis. *Library & Information Science Research*, *35*(2), 127-136.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques (3rd ed.)*. Waltham, Mass.: Morgan Kaufmann.

Kim, J. (2017). User perception and use of the academic library: A correlation analysis. *The Journal of Academic Librarianship*, *43*(3), 209-215.

Kohavi, R., & Quinlan, R. (1999). *Decision tree discovery.* Retrieved from http://ai.stanford.edu/~ronnyk/treesHB.pdf

Kuh, G. D., & Gonyea, R. M. (2015). The role of the academic library in promoting student engagement in learning. *College & Research Libraries*, *76*(3), 358-385.

Larose, D., & Larose, C. (2015). *Data mining and predictive analytics (2nd ed.)*. Hoboken, N.J.: John Wiley & Sons.

Leetaru, K. H. (2015). Mining libraries: Lessons learned from 20 years of massive computing on the world's information. *Information Services & Use*, *35*(1-2), 31-50.

Murray, A., Ireland, A., & Hackathorn, J. (2016). The value of academic libraries: Library services as a predictor of student retention. *College & Research Libraries*, *77*(5), 631-642.

Puarungroj, W., Boonsirisumpun, N., Pongpatrakant, P., & Phromkot, S. (2017) A preliminary implementation of data mining approaches for predicting the results of English exit exam. In *2017 2nd International Conference on Information Technology (INCIT), 2-3 Nov 2017, Nakhonpathom, Thailand* (pp. 89-94). Piscataway, N.J.: IEEE. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8257847

Renaud, J., Britton, S., Wang, D., & Ogihara, M. (2015). Mining library and university data to understand library use patterns, *The Electronic Library, 33*(3), 355-372.

Rokach, L., & Maimon, O. (2010). Classification trees. In Maimon O, & Rokach L. (Eds.), *Data mining and knowledge discovery handbook* (pp. 149-174). New York: Springer.

Romero, C., & Ventura, S. (2010) Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, *40*(6), 601-618.

Romero, C., Ventura, S., Espejo, P.G., & Hervás, C. (2008). Data mining algorithms to classify students. In Baker, R.S.J., Barnes T. and Beak J. E. (Eds.), *Proceedings of 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008* (pp. 8-17). Retrieved from http://www.educationaldatamining.org/EDM2008/uploads/proc/full proceedings.pdf

Siguenza-Guzman, L., Saquicela, V., Avila-Ordonez, E., Vandewalle, J., & Cattrysse, D. (2015). Literature review of data mining applications in academic libraries. *The Journal of Academic Librarianship 41*(4), 499-510.

Soria, K. M., Fansen, J., & Nackerud, S. (2013). Library use and undergraduate student outcomes: New evidence for students' retention and academic success. *portal: Libraries & The Academy, 13*(2), 147-164.

Teoh, Z. M., & Tan, A. K. (2011). Determinants of library use amongst university students. *Malaysian Journal of Library & Information Science, 16*(2), 21-31.

Tsai, C. F., Tsai, C. T., Hung, C. S., & Hwang, P. S. (2011). Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. *Australasian Journal of Educational Technology*, *27*(3), 481-498.