American Journal of Education and Information Technology
2021; 5(1): 21-26
http://www.sciencepublishinggroup.com/j/ajeit
doi: 10.11648/j.ajeit.20210501.14

Study of Association Rule Between College Students' Learning Behavior and Academic Records Based on Data Mining

Zhong Li, Ying Li, Haiyang Li, Keke Sun

School of Emergency Management, Institute of Disaster Prevention, Langfang, China

Email address:
Lizhong@cidp.edu.cn (Zhong Li)

To cite this article:
Zhong Li, Ying Li, Haiyang Li, Keke Sun. Study of Association Rule Between College Students' Learning Behavior and Academic Records Based on Data Mining. *American Journal of Education and Information Technology*. Vol. 5, No. 1, 2021, pp. 21-26.
doi: 10.11648/j.ajeit.20210501.14

Received: March 6, 2021; **Accepted**: March 17, 2021; **Published**: March 26, 2021

Abstract: The association rule between college students' daily behavior and school records has been the focus of education. Firstly, this paper summarizes the previous research results on this kind of problem, and studies many factors that affect college students' performance. Secondly, as an example of Institute of Disaster Prevention in China, the data of school records, online time and library time were extracted in this paper. The association between school records and the average daily online time, the average daily network flow and the time staying in library are discussed qualitatively via statistical analysis. It is pointed out that there is a negative correlation between daytime online time and academic records, and a positive correlation between library stay time and academic records. Then, using K-means clustering mining algorithm to analyze the online time and academic records, the results show that the excellent students spend less time online than the poor students, especially in the daytime. And using Apriori association analysis mining algorithm to study the relationship between the length of stay in Library and academic records. The minimum support and minimum credibility are set at 60%, and three strong association rules are obtained, that is, the students with good academic records stay in library for the longest time, the students with general academic records take the second place, and the students with poor academic records stay in library for the shortest time, which is completely consistent with the actual situation. This shows that the results of statistical method and data mining algorithm are consistent, that is, students who study well spend less time on Internet (shorter in the day) and more time in library than those with average records. The conclusion can help teachers to guide students to improve their achievement, so that students can better complete their studies, which has important guiding significance.

Keywords: Data Mining, Clustering Algorithm, Apriori Algorithm, Online Time, Library Time

1. Introduction

Computers are widely used in the field of Education with the development of information age [1]. The education management is improved using various teaching achievement systems, enriching the classroom teaching, and various types of data of students show a rapid increase trend [2], providing rich data sources. How to mine valuable data from these sources to facilitate student management and promote learning is always the goal of educators [3]. Researchers have done a lot for the association rules between various factors and results with the development of big data in recent years, instead of pursuing a causal relationship [4, 5].

As we all know, many factors may affect academic performance, such as daily habits, family conditions, psychological problems, and social relations etc. For example, Gu Jinchi et al. evaluate the parents' education background, travel time, study time, number of failures, family relations, absences, etc. for the students as factors affecting performance using multiple regression and decision tree methods via social surveys, so as to analyze and predict the students' performance [6], pointing out that the number of absence and the number of failures have a negative effect on academic performance. Wu Tangyan took female college

students in Beijing as the objects and discussed the impact of 42 indicators such as the personal factor, school factor and family factor on academic performance [7], pointing out that the personal factor and the school factor have a significant impact on academic performance, but the paper ignored the impact of students' daily behavior on their learning. H Yao et al. analyzed the relationship between student behavior, effort, sleep habits, etc. and academic performance, established a multitask prediction framework to predict students' performance, and pointed out the impact of different factors on their achievements [8]. Zhou Qing et al. analyzed the association between the times of breakfast and scores based on the campus card data in order to predict the performance, and it is believed that students who eat breakfast often have better achievements [9]. Although multi factors that affect students' performance are considered in the above methods, it fails to consider students' comprehensive daily behaviors as influencing factors, which have important impact on achievements.

It is indicated in studies that good behaviors of students are the external manifestations of their thinking, morality and civilization, and also a prerequisite for recognization by the society, which can characterize a person's toughness, perseverance, strength and highly-motivated traits [10]. Relevant studies have shown that the longer college students staying in library, the better their performance [11]. So we can say that, college students' access to the library and the stay time are important factors affecting their performance. In this paper, the data such as students' online information, access to the library and academic records etc. were extracted from the campus student database of Institute of Disaster Prevention to establish a model of Apriori association rule for college students' daily behavior and academic performance via association analysis of data mining, so as to analyze the influence of various behaviors and habits on achievements. We hope that it can help educators formulate teaching plans and student management systems with a target, so as to improve academic performance and promote college student talents.

2. Extraction and Preprocessing of Academic Records Data

Considering the completeness and validity of the data, and that college students are often in the stage of graduation practice and design (thesis) in the last year, only the scores of students at certain grade in the first six semesters was extracted in this paper, including online time, access to the library and academic records of each subject, filtering out unnecessary data, and sorting by student IDs and semesters.

Students' academic records are divided into scores and grades. The grades are quantified using a certain method, and the comprehensive score are obtained by formula (1). Assuming that a student's score of i-th subject in a semester is denoted as S_i, and the credit is denoted as Q_i, the student's comprehensive score S for this semester is obtained using the following formula:

$$S=\sum_i (S_i * Q_i) / \sum_i Q_i \tag{1}$$

3. Association Between Internet Habits and Academic Performance

The online behaviors of college students are common. They can access WIFI network everywhere in the campus, with advantage of convenience, fast access and low fees. College students are all interested in network. Students are encouraged to search for information on Internet and learn advanced technology at home and abroad. However, some students will play games and spend time in gossip on Internet, resulting in irregular study/rest and poor performance.

3.1. Association Analysis of Internet Habits and School Records

The students' online data is extracted. The online time (in Min.) and network flow (in MB) in the day and at night are calculated respectively by the semester, time, and flow, which are integrated with scores. The results are shown in Table 1.

Table 1. *Integrated data of surfing habits and scores (partial).*

Student ID	Average score	Total online time, Min.	Total network flow, MB	Total online time in the day, Min.	Total online time at night, Min.
145014133	90.9	126621	276931	61981	64640
135013330	90.3	113730	151637	63101	50629
145023106	90.1	289177	465528	152581	136596
145023210	90.1	137127	317277	79545	57582
125022114	89.9	114467	289571	57752	56715
145021115	89.7	224367	338402	96224	128143
145012112	89.7	18184	58120	10371	7813
145031203	89.7	284141	324924	137357	146784
125041402	89.6	35963	21359	28281	7682
135012129	89.5	99302	125909	69048	30254

This is a multi-dimensional data table. We can use a covariance matrix to observe the association between variables. The covariance is calculated by formula (2).

$$COV\ (X,\ Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1} \tag{2}$$

The covariance matrix is obtained via calculation (Table 2).

Table 2. *Covariance matrix.*

	Average score	Total online time	Total online time in the day	Total online time at night	Total network flow
Average score	1	-0.15	-0.21	-0.05	-0.23
Total online time	-0.15	1	0.95	0.92	0.80
Total online time in the day	-0.21	0.95	1	0.76	0.79
Total online time at night	-0.05	0.93	0.76	1	0.69
Total network flow	-0.23	0.80	0.79	0.69	1

It can be found from Table 2 that online time and network flow are negatively correlated with scores, but the negative association index of online time at night is relatively small, and that in the day is relatively large, indicating that the online time at night has less impact on performance than that in the day. The reason may be that students' courses are generally set in the day. If the students spend too much time online in the day, they will not listen carefully, resulting in declined performance.

To observe the association between the variables more clearly, a heatmap is drawn based on the covariance matrix (Figure 1).

It can be seen from the figure that the darker the purple part, the greater the negative association index, and the lighter the yellow part, the greater the positive association index.

3.2. Cluster Analysis of Online Habits and Scores

Taking 75 points as the demarcation of scores, it is defined that the comprehensive score higher than 75 is excellent, and that no higher than 75 is general, then we can use K-means algorithm for cluster analysis.

K-means algorithm is an efficient clustering analysis algorithm, which will classify *n* sample points into K classes, so that each point belongs to the class corresponding to the closest centroid (the mean of all sample points in a class) [12], where K represents the number of classes in the clustering algorithm, and *means* represents the mean algorithm.

Calculation steps of K-means algorithm:

(1) Take K initial centroids: select K points randomly from the data as the center of the initial cluster.

(2) Classify each point into the corresponding class: Classify each point into the closest based on least Euclidean distance criteria.

(3) Recalculate the centroid: Recalculate the centroid of each class using the means method.

(4) Iterative calculation of centroid: repeat the second and the third steps for iterative calculation.

(5) Clustering completed: the clustering center no longer changes.

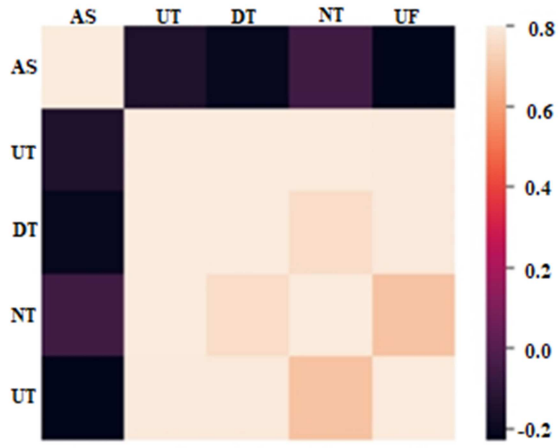

Figure 1. *Heatmap of covariance matrix.*

*(*Notes: AS-Average score, UT-useTime, DT-dayTime, NT-nightTime, UF-useFlow*)*

The scores are classified into two groups in this paper: Excellent and General, so K=2 here. To see the difference between the clusters more clearly, a histogram is adopted to analyze the online data of both groups after clustering. Figure 2 shows the cluster histogram corresponding to Excellent group. It can be seen that the peak of average score is on the right of 75 points. Figure 3 shows the cluster histogram corresponding to General group. It can be seen that the peak of average score is on the left of 75 points.

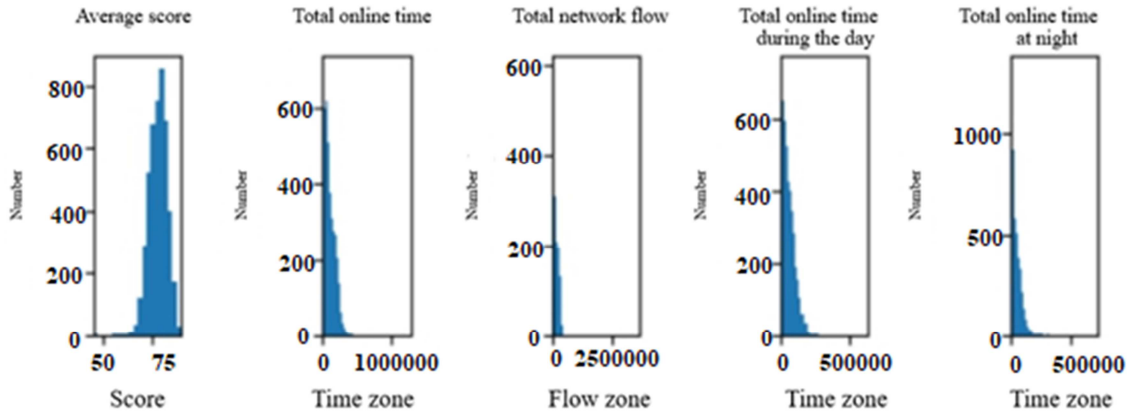

Figure 2. *Cluster analysis of online habits and scores - Excellent group.*

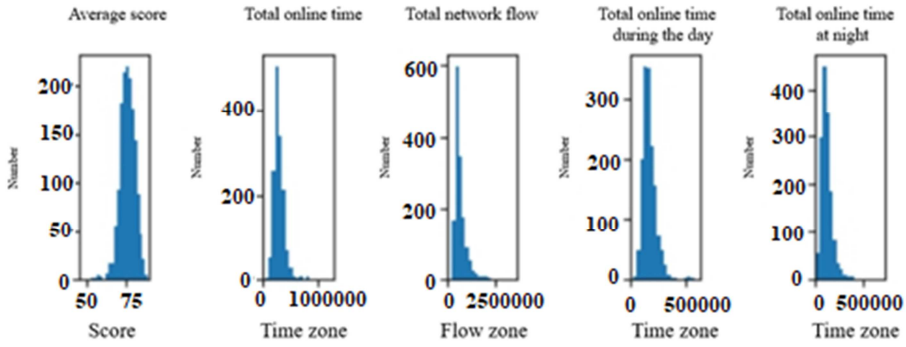

Figure 3. *Cluster analysis of online habits and scores - General group.*

Comparing the two groups, it can be seen that the online time and network flow of Excellent group are lower than those of General group, and the online time and network flow of Excellent group in the day are less than those of General group, indicating that students' habit of surfing the internet have a certain impact on performance. Therefore, students should be guided to reduce the time spent on the Internet in the day as possible and focus on classroom learning.

4. Association Analysis of Library Time and Scores

Relevant studies have shown that the longer college students staying in library, the better their achievement [9]. Therefore, college students' access to the library and the time staying in library are important factors affecting their performance. At present, an access control system has been set at entry/exit of libraries in most colleges, and the time when students enter and exit the library is recorded. The data such as the number of entries and the time staying in library can be obtained by calculation.

4.1. Statistical Analysis of Time Staying in Library and Scores

For convenient research, the average daily time staying in library in each semester will be taken as an index that affects academic performance.

The library entry data of college students at certain grade is extracted to calculate the average daily time staying in library (in Min.) for the semester by statistics, and a Hash chart is drawn based on the student's comprehensive score for the semester (Figure 4).

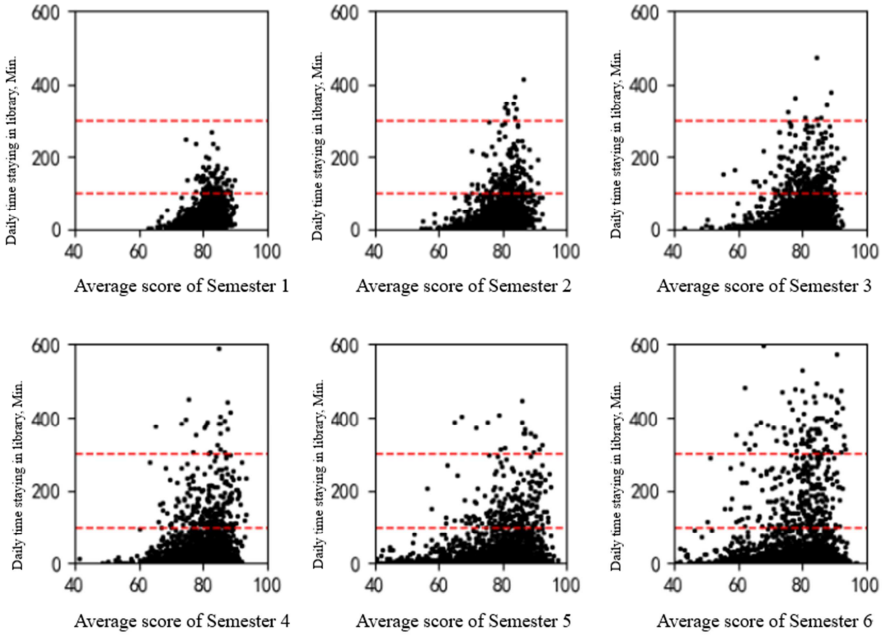

Figure 4. *Hash chart of time staying in library of students at Grade 2015.*

Figure 4 shows a statistical hash chart of the scores and the average daily time staying in library of students at Grade 2015 of Institute of Disaster Prevention in the first six semesters. The abscissa represents the average score and the ordinate represents the average daily time staying in library. It can be seen from the figure that, students with excellent performance spend more time in library every day in general. In addition, students at higher grades stay longer in library, which is completely consistent with the actual situation.

4.2. Association Analysis Between Time Staying in Library and Grades

Apriori algorithm is a common method for association rule mining [13]. It is required to set the minimum support and the minimum confidence for the algorithm. Large support indicates that there are enough data samples, and large confidence indicates that the proportion of data that meets the requirements is high enough. The frequent itemsets greater than the minimum support are generated with this algorithm, and then the association criteria greater than the minimum confidence is generated in the frequent itemsets [14]. The frequent itemset refer to itemset T containing item A, whose support is greater than or equal to the minimum support $minS$, that is:

$$S_{A\to T} = \frac{|T(A)|}{|T|} \geq minS \tag{3}$$

The frequent itemset containing k items is called frequent k-itemset, denoted as L_k, and the item whose support is greater than or equal to the minimum support will belong to frequent k-itemset.

The implementation of Apriori algorithm is shown in Figure 5. First, search for and generate a frequent itemset L_1 with length 1 based on the given minimum support, and then generate a frequent itemset L_2 with length 2 using L_1, and so on, until all frequent itemsets are generated [15].

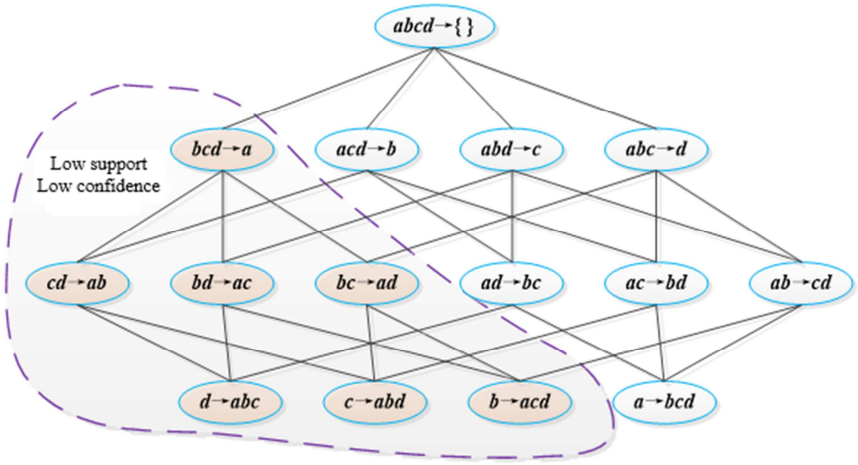

Figure 5. *Diagram of Apriori algorithm.*

The strong association criterion is generated based on the minimum confidence in the frequent itemsets. For each frequent itemset L, calculate the confidence of all non-empty subsets L' of L. If $C_{L'\to(L-L')}$ is greater than the minimum confidence $minC$, namely

$$C_{L'\to(L-L')} = \frac{|T(L)|}{|T(L')|} \geq minC \tag{4}$$

The strong association criterion is generated: $L' \Rightarrow (L - L')$.

It can be seen that the duality of sample data is required for Apriori algorithm. Therefore, it is necessary to dualize the scores and the time staying in library.

The students' comprehensive scores were summarized by class and semester, sorting from high to low to form the ranking for each semester by the class. The top 20% scores are recorded as Grade-A, the last 20% scores are recorded as Grade-C, and the middle 60% scores are recorded as Grade-B, stored in the database by the semester and the student ID. Similarly, the time staying in library is also sorted by the duration, with the longest 20% as Grade-I, the shortest 20% as Grade-III, and the middle 60% as Grade-II.

First, the association between the student's time staying in library in the first six semesters and the comprehensive scores is analyzed using Apriori algorithm. Set the minimum support and the minimum confidence to be 0.6, and get the strong association rules for each semester using Apriori algorithm. Then the strong association rules of the six semesters are combined based on the average of same kind, to obtain the comprehensive strong association rules between the college students' time staying in library and scores in the first three years. The results are shown in Table 3.

Table 3. *Strong association rules between library time and scores.*

Grade	Grade of library time	Support	Confidence
A	I	0.62	0.77
B	II	0.76	0.84
C	III	0.67	0.71

It can be seen from Table 3 that students with excellent records stayed in library for the longest time, followed by students with general records, and students with poorer records have the shortest time staying in library, which is consistent with the actual situation. Therefore, students should be guided to go to library in daily education and stay as long as possible to get better achievements.

5. Conclusion

The influence of college students' learning behaviors on academic performance has always been one of the issues that education experts are interested in, especially in the current

information age. All kinds of information about students in school can be recorded, providing rich sources for analysis and study. In this paper, the impact of students' online habits and time staying in library on academic performance is analyzed using data mining methods. The calculation results show that students with excellent performance spend less time online than those with general performance, especially in the day, the time difference is more obvious. In terms of the time staying in library, students with excellent performance spend more time in library than those with general performance. Therefore, the educator shall actively guide students to spend more time in library and less time on Internet, and not to surf the Internet as possible, especially in the day. The students shall spend precious time on studying and searching information, and master the time in college to get great achievements.

There are many factors affecting achievements. In this paper, only online time and the time staying in library are discussed, and some conclusions are drawn. It is required to study taking more factors into consideration in the next step, such as habit of breakfast, bathing, exercise, rest, readings, hobbies and interests, etc., in order to obtain more objective and true association between college students' daily behavior and achievements, as well as more information for reference, which has important significance to guide students to improve their performance and study better.

Acknowledgements

This paper was supported by the Fundamental Research Funds for the Central Universities (No. ZY20180121). Thanks the departments of Institute of Disaster Prevention for related data.

References

[1] SHEN Hang-jie, JU Sheng-gen, SUN Jie-ping. Performance prediction based on fuzzy clustering and support vector regression [J]. Journal of East China Normal University (Natural Science). 2019 (05): 66-73+84.

[2] LI Tiebo. Student Behavior Characteristics Analysis and Prediction Based on Campus Big Data [J]. Journal of Chongqing University of Technology (Natural Science). 2019, 33 (07): 201-206.

[3] CHEN Xi, MEI Guang, ZHANG Jinjin, XU Weisheng. Student grade prediction method based on knowledge graph and collaborative filtering [J]. Journal of Computer Application. 2020. 40 (2): 595-601.

[4] LI Zhong, AN Jianqin, LIU Haijun, SONG Yiyao. Association mining algorithm and its development trend [J]. Intelligent Computer and Applications. 2017. 7 (5): 22-25.

[5] YU Qinyang. Campus Big Data Based Analysis and Visualization of Students' Abnormal Behaviors [D]. Beijing University of Technology. 2019.

[6] GU Jinchi. Analysis and prediction of influencing factors of student academic records based on multiple regression and decision tree model [J]. Management observation. 2019 (25): 156-157.

[7] WU Tangyan. The influence of students, school and family factors on the academic records of female college students [J]. Frontier. 2015 (11): 115-117.

[8] H Yao, D Lian, Y Cao, et al. Predicting Academic records for College Students [J]. Acm Transactions on Intelligent Systems & Technology. 2019. 10 (3): 1-21.

[9] ZHOU Qing, WANG Wei-fang, GE Liang, XIAO Yi-feng, TAI-Dai. Student Performance Prediction based on Campus Card Data and Curriculum Classification [J]. Computer Knowledge and Technology. 2018, 14 (24): 236-239.

[10] HE Hong, LIU Dongbo, ZHANG Bi. Correlation analysis between online learning behavior and student academic records on SPOC platform based on learning style classification [J]. Science and technology information. 2019. 3: 207-210.

[11] YANG Fang. A Research On Students' Borrowing Behavior in Local University Library Based on Data Mining - Take Ordos Institute of Technology as An Example [D]. Inner Mongolia University of science and technology. 2020.

[12] YANG Yi-fan, HE Guo-xian, LI Yong-ding, K-Means Algorithm for Optimizing Initial Cluster Center Selection [J]. Computer Knowledge and Technology. 2021.17 (5): 252-255.

[13] DONG Xuan-meng, GUO Li-wen, DONG Xian-wei, WANG Fu-sheng. Association Mining of Influencing Factors of Coal Spontaneous Combustion Based on Apriori Algorithm [J]. Journal of North China University of Science and Technology (Natural Science Edition). 2021.43 (1): 21-25.

[14] Huang Ke, Bi Chunguang, Wang Jinlong, Guo Hai, Yuan Shuai. Research on application of improved Apriori algorithm based on frequent itemsets in smart greenhouse [J]. Journal of Chinese Agricultural Mechanization, 2020, 41 (9): 182-189.

[15] HUANG Hui, ZENG Qingtao, TANG Mingjie, ZHANG Xiaoliang, Application of Student Achievement Analysis Based on Association Analysis [J]. Journal of Beijing Institute of Graphic Communication. 2021. 29 (2): 130-136.

Biography

Zhong Li (1966 -), the corresponding author, male, PH. D, Professor, main research direction: space physics.

Ying Li (1996 -), female, Master of Engineering, research direction: data processing.

Haiyang Li (1995-), male, Master of Engineering, research direction: information processing technology.

Keke Sun (1994-), male, Master of Engineering, research direction: information processing technology.