Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 9, No. 2, 2024

Research on the Association Analysis of Online Learning Behaviors Based on the Apriori Algorithm

Zheng Tang, Zhengwei Jiang, Ying Li, Haowei Yuan, Jiayu Han, Chao Chen *

Sichuan University of Science & Engineering, Yibin, Sichuan, China
*** Corresponding author:** Chao Chen (Email: 81799516@qq.com)

Abstract: In the context of rapid development of information technology, online education has become an important driving force for innovation in the field of education. This study aims to deeply analyze the intrinsic connection between online learning behaviors and students' academic achievements. By applying the Apriori algorithm to mine association rules in learning behavior data, it reveals the key behavioral factors affecting academic performance, providing a scientific basis for optimizing online teaching strategies and improving learning outcomes.

Keywords: Online Learning Behaviors; Apriori Algorithm; Association Rule Mining; Teaching Optimization.

1. Introduction

With the continuous evolution and widespread application of information technology, online education has emerged as a prominent topic, leading the trend of innovation in the field of education. Online education, with its convenience, has not only greatly benefited students but also led to profound changes in the education industry. In many situations, due to various constraints, traditional classroom environments often fail to meet students' pursuit of efficient learning. Online education platforms, with their flexibility and personalized teaching models, can fill this gap and precisely meet the diverse learning needs of students. It is worth noting that although online education platforms have accumulated massive amounts of data, covering various dimensions such as students' learning behaviors, activity trajectories, and academic performance, the potential value of this data has not yet been fully realized. Effective data mining and analysis would undoubtedly unlock much meaningful information, further promoting the improvement of educational quality and learning outcomes.

In the field of educational data mining and analysis, scholars have continuously explored more efficient methods to understand students' learning behaviors and their impact on academic performance. Chen Jinyin et al.[1] predicted the academic performance of online learners using a BP neural network, combined with actual entropy analysis, revealing the regularity of online learning behaviors and their positive correlation with offline performance. Lin Pengfei et al.[2] used a deep neural network (DNN) to predict the performance of MOOC learners more accurately than linear regression, identifying potential dropout risks and implementing personalized teaching interventions, significantly improving course completion rates. Bai Jieqiong et al.[3] explored student collaborative learning behavior patterns using lag sequential analysis (LSA), finding that collaborative learning promotes higher-order thinking and knowledge construction, clarifying the key behavioral sequences of effective collaborative learning and providing specific guidance for teachers. Pu Yunwei et al.[4] developed an integrated classification model based on MLP-Bagging, successfully classifying the behaviors of online learners, revealing the characteristics of different learner groups, and improving the classification accuracy rate to 98.72%, demonstrating the practical application value of the model in learning behavior analysis. These studies not only showcase the powerful potential of data-driven educational analysis but also provide educators with strategic bases to optimize the online teaching environment and promote the improvement of student learning outcomes.

2. Apriori Association Analysis Algorithm

2.1. Algorithm Principles

The Apriori algorithm is a classic algorithm used for frequent itemset mining and association rule learning. It is primarily used to find frequent itemsets in a dataset, which are combinations of items that often appear together, and to further generate association rules, which are rules indicating that one event often occurs when another event happens [5]. The core of the Apriori algorithm is based on a fundamental principle: if a set of items is frequent, then all of its subsets must also be frequent.

Support Calculation: Support refers to the frequency at which an itemset appears in all transactions. The Apriori algorithm first calculates the support for all individual items in the dataset and then filters out the frequent single items based on the user-defined minimum support threshold. The formula for calculating support is:

$$\text{Support}(X) = \frac{\text{Count}(X)}{\text{Total Transactions}} \quad (1)$$

Frequent Itemset Generation: Once frequent single items are found, the algorithm combines these items to form pairs and calculates the support for these pairs. Then, based on the minimum support threshold, it filters out the frequent itemsets of two items. This process is repeated iteratively, forming combinations of more items, until no new frequent itemsets can be generated.

Association Rule Generation: After finding the frequent itemsets, the algorithm generates association rules. Each association rule consists of an antecedent (if part) and a consequent (then part). To evaluate the effectiveness of these rules, confidence is typically used. Confidence is the probability that the consequent will occur given that the antecedent has occurred, and the formula for calculating

confidence is:

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)} \quad (2)$$

2.2. Algorithm Process

The Apriori algorithm aims to mine frequent co-occurring item combinations, or frequent itemsets, from massive data and further refine strong association rules to provide a basis for understanding implicit relationships in the data and driving decision-making. This process is carried out in stages:

Firstly, the algorithm calculates the support for each individual item in the database, which is the proportion of the item's occurrence across all transactions. Only items with support exceeding a predefined threshold are retained as frequent items. This is achieved by traversing the database and counting the occurrences of each item.

Next, Apriori uses the frequent single itemsets identified in the previous step to build larger candidate frequent itemsets by combining these singles. For example, if both {A} and {B} appear frequently, then {A,B} becomes a candidate. All candidate sets must be validated against actual support, and only those meeting the minimum support criterion are confirmed as truly frequent itemsets. This process of merging and validation iterates, gradually expanding the size of the itemsets, until no new frequent itemsets can be found.

The final stage focuses on the generation and evaluation of association rules. For each frequent itemset, the algorithm explores all possible subset combinations to define the "antecedent" and "consequent" of the rules and calculates the confidence for each rule, which is the probability of the consequent occurring given the antecedent, calculated as the support for the rule {X, Y} divided by the support for {X}. To ensure the practical value of the rules, a minimum confidence threshold is set, and only rules that meet this threshold and the minimum support are adopted.

3. Application of the Apriori Algorithm in Teaching Optimization

3.1. Current Issues and Solutions

In the online teaching environment of higher education, especially for technology-oriented courses like "Database Principles B," quantifying and optimizing student learning outcomes and engagement pose core challenges. These challenges are mainly reflected in the following aspects:

Firstly, the correlation between learning behaviors and academic performance is ambiguous. While it is intuitively believed that active participation in various online learning activities (such as attendance checks, completion of task points, submission of assignments, etc.) helps improve grades, without systematic analysis, the strength and specific impact of this correlation are not clear [6]. Students might invest a lot of time in certain activities with limited impact on their grades, while other activities might have a more significant effect. Secondly, there is underutilization of data mining. Online learning platforms like Chaoxing Xuetong can collect a large amount of student learning behavior data, including but not limited to attendance checks, chapter visits, assignment submissions, etc.[7]. However, the potential value of this data has not been fully utilized, requiring effective data analysis methods to reveal the patterns and trends hidden behind the data. Thirdly, there is a lack of personalized learning paths. Each student has different learning habits and abilities, necessitating the customization of personalized learning paths based on individual learning behaviors and preferences to improve learning efficiency and outcomes.

In this study, we consider combining the analysis of students' comprehensive scores with other dimensions of behavior data in online learning, using the Apriori algorithm to mine the association rules. The other learning behaviors considered include the following seven dimensions of information: attendance checks, chapter visits, completion of task points, submission of assignments, submission of exams, peer review of assignments, and online duration.

3.2. Data Source

The research data for this paper comes from the student learning data recorded by the Chaoxing Xuetong platform, which gathers high-quality course resources from many well-known universities and has a vast learning database, providing rich research material for this study. The "Database Principles B" course, as a compulsory basic course for students majoring in Computer Science, Software Engineering, etc., aims to help students deeply understand the basic principles of database systems, master the basic skills of database design, management, and application, and lay a solid foundation for their future academic research and career development.

This study takes the online behavior data of the "Database Principles B" course in the first semester of the 2022-2023 academic year as the research object. The course adopts an online teaching model and has a total of 94 students participating. Throughout the course learning process, students log into the Chaoxing Xuetong platform for learning, and their online learning behavior data is continuously generated and collected, becoming an important resource for this study. Through in-depth analysis of this data, we can understand the learning situation of students and provide a scientific basis for teaching improvement and course design[8]. At the same time, this data also helps teachers better understand the characteristics and rules of online learning, promoting the continuous development and innovation of online education.

3.3. Data Preprocessing

Data preprocessing is a key step in data analysis projects to ensure data quality, eliminate noise, handle missing values and outliers, and transform raw data into a form suitable for analysis [9]. In this paper, data preprocessing mainly includes the following aspects:

Error Value Handling: The data set contains error values named "#N/A", and the team identified and cleared these 56 erroneous records to ensure data purity and avoid the impact of error values on the analysis results.

Outlier Handling: Within the group identified as students (with a role value of 3), the team identified records with teacher behaviors (such as grading assignments, setting homework, etc.), and judged that these students might be graduate teaching assistants. Considering that these records do not reflect typical student behavior, the team removed such outliers from the data set to focus on the analysis of true student learning behaviors.

Through the above steps, the team completed the cleaning and organization of the raw data, ensuring the accuracy and effectiveness of subsequent statistical analysis and data mining.

3.4. Designation of Discrete Nodes

Before analyzing the data using the Apriori algorithm, it is necessary to discretize the data. The following specifies the discrete interval nodes for discretization.

Due to the different nature of the data, it is divided into two categories:

Category 1: Includes attendance checks, completion of task points, submission of assignments, comprehensive scores, submission of exams, and peer review of assignments. Data in the first category have a maximum value, belonging to finite data. For example, in the completion of task points, there are a total of 44 task points in the class, so the data for all students completing task points will not exceed 44. The meaning of MAX is the total data, such as the MAX value for completing task points in this class is 44, and the rest are similar. To facilitate subsequent discretization, calculate the data nodes at MAX0.6, MAX0.75, and MAX*0.9, as shown in Figure 1:

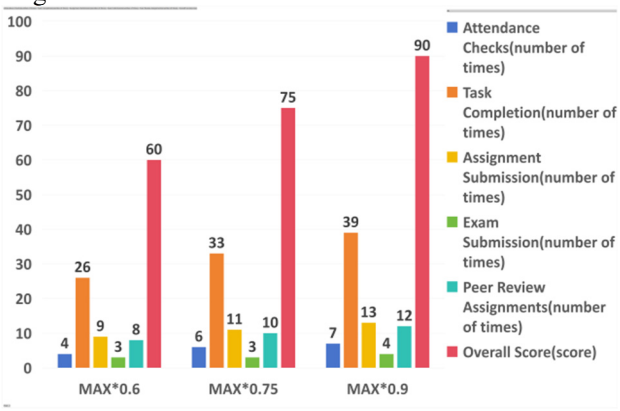

Figure 1. First category data points

The second category: Student chapter visits and online duration. The second category of data does not have a maximum value and is considered unrestricted data. Students can visit chapters without limit and stay online indefinitely, so it is impossible to know the total value of the data. Our team calculated the confidence interval for the online duration of everyone in the class and took the average of the confidence interval as the criterion. The average online duration (AVE) for students is estimated as the average online duration, and the AVE for student chapter visits is the total number of chapters. The data points at AVE1, AVE1.5, and AVE*2 are calculated, as shown in Figure 2:

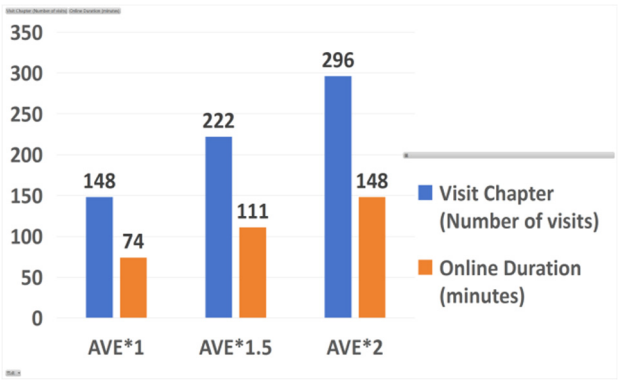

Figure 2. Second category data points

3.5. Data Discretization

The first category of data is divided into discrete intervals according to MAX*0.9, 0.75, and 0.6. X represents the current student behavior data, as shown in Table 1.

Table 1. Discretization Rules for the First Category of Data

Code Name (First Category)	Discrete Interval
Code 1	X>=MAX*0.9
Code 2	MAX*0.75<=X<MAX*0.9
Code 3	MAX*0.6<=X<MAX*0.75
Code 4	X<MAX*0.6

The second category of data is divided into discrete intervals according to AVE1, 1.5, and 2. The meaning is as follows: if a student's chapter visits equal the average value of the confidence interval (total number of chapters), it represents that the student has visited all the chapters; AVE1.5 indicates that after visiting all the chapters, the student revisited some challenging sections; AVE*2 represents that after visiting all the chapters, the student revisited all the chapters again before the exam to consolidate knowledge, and the online duration is analogous. The discretization intervals for the second category of data are shown in Table 2.

Table 2. Discretization Rules for the Second Category of Data

Code Name (Second Category)	Discrete Interval
Code 1	X>=AVE*2
Code 2	AVE*1.5<=X<AVE*2
Code 3	AVE*1<=X<AVE*1.5
Code 4	X<AVE*1

3.6. Encoding

Table 3. Behavior Encoding Table

Behavior Dimension	Encoding 1	Encoding 2	Encoding 3	Encoding 4
Attendance Checks	T1-High attendance check-ins	T2-More attendance check-ins	T3-Average attendance check-ins	T4-Few attendance check-ins
Exam Submission	E1-High exam submission rate	E2-More exam submission rate	E3-Average exam submission rate	E4-Low exam submission rate
Visit Chapter	C1-High visit frequency	C2-More visit frequency	C3-Average visit frequency	C4-Few visit frequency
Peer Review Assignments	P1-High number of peer reviews	P2-More peer reviews	P3-Average number of peer reviews	P4-Few peer reviews
Task Completion	J1-High completion rate	J2-More completion rate	J3-Average completion rate	J4-Low completion rate
Online Duration	O1-Long online duration	O2-More online duration	O3-Average online duration	O4-Short online duration
Assignment Submission	W1-High submission rate	W2-More submission rate	W3-Average submission rate	W4-Low submission rate
Overall Score	S1-Excellent	S2-Good	S3-Medium	S4-Poor

Based on the aforementioned encoding rules, the encoding process is conducted. Taking participation in attendance

checks as an example, the encodings 1, 2, 3, and 4 are T1, T2, T3, and T4, respectively, with a descending trend in the level of participation [10]. As shown in Table 3.

Table 3 is an encoding table based on the rules defined in Tables 1 and 2. Each code has a specific meaning associated with it. For example, T1 corresponds to Code 1 in Table 1 (X >= MAX*0.9), representing a high number of attendance checks. C1 corresponds to Code 1 in Table 2 (X >= AVE*2), indicating a high frequency of chapter visits, and so on. Through this encoding process, the above-mentioned behavioral data have been successfully discretized.

After discretization, the Apriori algorithm can be used for analysis. Setting the support threshold to 0.4 and the confidence threshold to 0.8, six association rules are derived, as shown in Table 4.

Table 4. Apriori Algorithm Analysis Results

Serial Number	Association Rule	Support Percentage	Confidence Percentage
1	T1→S1	58.33%	100.00%
2	P1, E1, T1→S1	42.71%	98.23%
3	P2,E1→T1	56.25%	96.30%
4	W3→E1	47.92%	95.65%
5	P2→T1	57.29%	94.55%
6	C4→J4	48.96%	93.62%

Association rule 1 (T1→S1) states that students who regularly attend attendance (T1) are likely to receive an excellent overall score (S1). With 58.33% support and 100% confidence, this rule shows a strong positive correlation between attendance and academic performance, meaning that regular attendance is a key factor in achieving excellent academic performance.

Association Rule 2 (P1, E1, T1→S1): When students extensively review peer-assessed assignments (P1), have a high exam submission rate (E1), and frequently participate in attendance checks (T1), they have a higher probability of achieving an excellent overall score (S1).The support of 42.71% and confidence of 98.23% suggest that this combination of behaviors is a strong predictor of outstanding academic performance.

Association Rule 3 (P2, E1→T1): Even if students only moderately review peer-assessed assignments (P2),but have a high exam submission rate (E1), they still have a high probability of frequently participating in attendance checks (T1). With a support of 56.25% and a confidence of 96.30%, it indicates that students may actively engage in class even if they do not review the most assignments.

Association Rule 4 (W3→E1): Students with an average homework submission rate (W3) have a higher probability of having a high exam submission rate (E1). The support of 47.92% and confidence of 95.65% suggest that students may perform well in exams even with a regular homework submission rate.

Association Rule 5 (P2→T1): Even if students only moderately review peer-assessed assignments (P2), they still have a high probability of frequently participating in attendance checks (T1). With a support of 57.29% and a confidence of 94.55%, it indicates that students may actively engage in class despite not reviewing the most assignments.

Association Rule 6 (C4→J4): Students who infrequently visit chapters (C4) are more likely to have a low completion rate for task points (J4). The support of 48.96% and confidence of 93.62% suggest that students who do not regularly access course materials may also struggle with completing course tasks.

4. Conclusion and Recommendations

Based on the association rules derived from the algorithmic analysis, it is evident that there is a significant positive correlation between high classroom engagement, positive examination attitudes (such as high exam submission rates), and academic performance. This correlation is observed not only in the influence of individual factors but is even more pronounced when multiple factors are considered together. Furthermore, it can be noted that there is a close relationship between the number of peer-reviewed assignments and the frequency of classroom attendance. Even a moderate level of peer review activity can significantly increase the enthusiasm for classroom participation, suggesting the important role of interactive learning in promoting student engagement.

However, this paper also identifies potential shortcomings, such as the low frequency of chapter visits being associated with a similarly low completion rate of task points . This may reveal issues in students' self-directed learning processes or indicate that insufficient exposure to course materials leads to a decline in learning outcomes[11]. Based on the above conclusions, the following recommendations can be made:

4.1. Strengthening Supervision of Attendance and Examinations

Teachers should adopt various measures to encourage students to actively participate in attendance checks and examinations. For students who are not actively participating in attendance, appropriate reminders or guidance can be provided [12]. Teachers can regularly check and provide feedback on students' attendance and exam submissions to timely understand their learning status and offer necessary help and support.

4.2. Encouraging Peer Review of Assignments

Peer review of assignments is an effective learning method that can enhance student engagement. Teachers should encourage students to review and evaluate each other's assignments after submission. To ensure the quality of reviews, teachers can provide standards and guidelines for peer review and review the students' review results for audit and feedback. Additionally, teachers can establish a reward mechanism for peer review, such as extra points or rewards for students who review carefully and evaluate accurately. This way, students can not only learn different problem-solving approaches from their peers' work but also improve their logical thinking and objective analytical abilities.

4.3. Optimizing Course Content and Access Methods

Teachers should pay attention to students who visit chapters less frequently and take appropriate measures to improve their engagement and learning outcomes. Teachers can adjust the teaching content to be more aligned with students' interests and needs [13]. For example, incorporating more case studies, real-world application scenarios, and interactive learning materials can increase student interest and

engagement. Moreover, teachers can optimize the way courses are accessed by providing online discussion forums, study groups, etc., to facilitate easier access and participation in the learning of course content.

4.4. Designing Personalized Learning Paths

Based on students' learning behavior data, developers of learning platforms can design personalized learning paths to help students of different levels learn more effectively [14]. Analyzing students' learning preferences and strengths can lead to the recommendation of suitable learning resources and paths. For instance, for students who are not proficient in a particular knowledge area, the learning platform can recommend related learning videos, practice questions, and explanatory articles to help them consolidate and improve. The learning platform can also dynamically adjust learning paths based on students' learning progress and performance, ensuring that students can learn at their own pace.

4.5. Continuous Data Monitoring and Analysis

Teachers and learning platforms should continuously monitor students' learning behavior data to promptly identify issues and adjust teaching methods [15]. Teachers can regularly review students' learning behavior data to understand their learning situations and progress, and quickly identify problems and difficulties in learning. Learning platforms can analyze students' learning needs and preferences based on learning behavior data, continuously optimizing and improving the functions and services of the learning platform.

Acknowledgments

This paper was funded by the China University Student Innovation and Entrepreneurship Training Program, Project Number: S202310622062; and it funded by the Teaching Reform Research Project of Sichuan University of Science & Engineering "Exploration and Practice of Top-notch Talent Training Model Integrating Discipline Competition and Academic Training", Project Number: JG-24024.

References

[1] Chen Jinyin, Fang Hang, Lin Xiang, et al. "Personalized Learning Recommendations Based on Online Learning Behavior Analysis" [J]. Computer Science, 2018, 45(S2): 422-426+452.

[2] Lin Pengfei, He Xiuqing, Chen Tiantian, et al. "Prediction and Intervention of MOOC Learner Dropout Under the Perspective of Deep Learning" J]. Computer Engineering and Applications, 2019, 55(22): 258-264.

[3] Bai Jieqiong, Zhou Jing. "Research on Student Cooperative Learning Behavior Patterns Based on Lag Sequential Analysis" [J]. Educational Theory and Practice, 2024, 44(10): 52-58.

[4] Pu Yunwei, Jiang Ying, Tian Chunjin, et al. Online learning behavior analysis based on MLP-Bagging ensemble classification model [J/OL]. Journal of yunnan university (natural science edition), 1-10 [2024-07-30]. HTTP: / /http:// kns.cnki.net/kcms/detail/53.1045.n.20240103.1614.006.html.

[5] Yan Haiwei, Zhang Qingliang, Lin Chunhua, et al. "Student Course Performance Correlation Analysis Based on the Apriori Algorithm" [J]. Computer Programming and Skills, 2023, (11): 13-15.

[6] Niu Qiuyue, Zhao Yingying. "Analysis of College Students' Online Learning Behavior Characteristics Integrating Multi-Source Data" [J]. Science and Technology Information, 2023, 21(13): 169-173.

[7] Lu Linyuan, Liu Ganhong, Wang Runhua. "Optimization of Online Teaching Design Based on Online Learning Behavior Analysis" [J]. Modern Information Technology, 2024, 8(06): 173-177.

[8] Zhang Jing, Luo Xin, Zhang Wanxin, et al. "Research on Learning Behavior Analysis Based on Learning Tong Platform" [J]. Information and Computer (Theoretical Edition), 2024, 36(01): 245-247.

[9] Lu Pengfei. "Student Online Learning Behavior Association Feature Mining Model Based on Improved Apriori Algorithm" [J]. Software, 2024, 45(01): 98-100.

[10] Wu Yuefen, Gong Zicheng, Yuan Jiang, et al. "Online Learning Behavior Analysis Based on MOOC" [J]. Journal of Hunan Institute of Science and Technology (Natural Science Edition), 2023, 36(02): 26-31.

[11] Suo Qi, Zuo Pei. "Evaluation and Difference Study of Learners' Learning Effectiveness in Online Education" [J]. Journal of Anhui Open University, 2024, (01): 38-46.

[12] Li Xinya. "Research on Personalized Learning Path Planning Based on Online Learning Behavior" [D]. Nanjing University of Posts and Telecommunications, 2023.

[13] Liu Lingyan, Zhao Bo. "Construction of Learners' Online Learning Ability Model: An Empirical Study" [J]. Chinese Adult Education, 2023, (16): 3-9.

[14] Zhang Shulian, Qiao Haiying, Tang Hamely. "Online Learning Behavior and Its Influencing Factors Based on MPOC" [J]. Journal of Hebei Normal University (Education Science Edition), 2022, 24(06): 121-127.

[15] Du Xingli. "Student Behavior Feature Analysis and Application Research Based on Data Mining" [D]. Southwest University of Science and Technology, 2023.