US-China Education Review A, May 2026, Vol. 16, No. 5, 295-302
doi: 10.17265/2161-623X/2026.05.004

Research on Automatic Mining Method of Behavior Rule Based on Apriori Algorithm

ZHU Aiqun
Shenzhen City Polytechnic, Shenzhen, China
LU Jin*
Shenzhen Polytechnic University, Shenzhen, China

Behaviour rule mining extracts valuable patterns from large amounts of behavioural data, which is crucial for analysing user behaviour, monitoring systems, and detecting security threats. Traditional manual or statistical methods often fail to reveal complex, hidden associations. This study therefore proposes an automated behaviour rule mining method based on an improved Apriori algorithm. This method adapts data preprocessing and feature encoding to behavioural characteristics, introduces an adaptive support threshold and incremental updating to enhance efficiency, and automates the generation and filtering of association rules from frequent behavioural sequences. When evaluated using accuracy, recall, and interpretability metrics on public user behaviour data and simulated system logs, the method was found to effectively mine meaningful rules while maintaining high efficiency with large-scale data. This work offers a scalable and interpretable approach to automated behaviour rule mining that supports intelligent analysis and decision-making.

Keywords: Apriori algorithm, behavior mining, frequent patterns, association rules, automated analysis, data mining

Introduction

With the rapid development of information technology and the internet of things, various information systems, intelligent terminals and industrial equipment continue to generate massive behavioral data, including but not limited to user click streams, device operation logs, network access records, sensor signal sequences, etc. (Li & Yang, 2024). These data contain rich behavior patterns and inherent laws, which can be automatically mined to deeply understand behavior motivation, predict future trends and identify abnormal states, thus providing decision-making basis for key applications such as personalized service recommendation, system performance optimization and network security protection (Sun et al., 2024). Traditional behavior analysis methods mostly rely on manually set statistical summary or fixed rules based on experience, which are often inadequate in the face of dynamic evolution of behavior patterns, context dependence and complex associations, and it is difficult to achieve efficient and intelligent law discovery (Diondra, Barb, & Brooke, 2021; Shimp, 2007).

Acknowledgements: This work was funded by the 2025 Key Research Project of Shenzhen Polytechnic University "Research on Key Methods for Analysis and Prediction of Social Behavior of Specific Characters on Multimodal Big Data (6025310008K)", 2024 Higher Education Scientific Research Planning Project of the Chinese Society of Higher Education "Research on the Analysis of Teaching and Learning Deep Interaction Characteristics in Smart Classroom Environment Supported by Multimodal Data (24XH0407)".

ZHU Aiqun, Ph.D., Shenzhen City Polytechnic, Shenzhen, China.
LU Jin* (corresponding author), Master, Shenzhen Polytechnic University, Shenzhen, China.

Data mining techniques, especially association rule mining, provide the possibility to automatically discover meaningful relationships between itemsets from behavioral data (Alborzi & Khanbabaei, 2026). As a classic algorithm in this field, Apriori algorithm generates frequent itemsets and derives association rules by searching layer by layer, which has been widely used in shopping basket analysis, web log mining and other scenarios (Mehedi et al., 2023; Pan, 2024; X. Y. Zhang & J. Zhang, 2023). However, the direct application of apriori algorithm to behavioral law mining still faces significant challenges (Zhang, 2025). Behavior data usually has obvious time sequence and context dependence, so it is necessary to design appropriate data representation and preprocessing mechanism. At the same time, the combination space of behavior items is very large, and the time and space efficiency of the algorithm directly affects its scalability. In addition, the excavated rules must have good interpretability and practical utility in order to truly support subsequent decisions and actions (Belayadi et al., 2022).

The purpose of this study is to deal with the above challenges, make a series of targeted improvements to the classic Apriori algorithm to make it better adapt to the inherent characteristics of behavior data, and ultimately design a complete, efficient, and highly automated behavior mining process. The specific plan will be carried out systematically from the following levels. Firstly, we will design a structured data preprocessing and feature encoding method for the common characteristics of behavioral data, such as time sequence, high dimensionality and strong context dependence. This method not only includes the conventional cleaning and denoising steps, but also transforms the original behavior event sequence into a standardized transaction representation suitable for association rule mining, while retaining the temporal and logical association information between behaviors as much as possible, laying a high-quality data foundation for subsequent rule mining (Chikhaoui et al., 2014). Secondly, in order to optimize the efficiency and scalability of the algorithm when dealing with large-scale behavioral data, we will introduce an adaptive support threshold setting mechanism and an incremental update strategy (Sreng et al., 2019). The adaptive mechanism is able to dynamically adjust the mining granularity according to the distribution of behavioral items in the dataset, avoiding important but low-frequency patterns from being missed (He et al., 2026). The incremental update strategy allows the system to update efficiently on the basis of the existing mining results when the new data arrives, without the need to re-calculate the global, thus significantly improving the algorithm's ability to process dynamic data streams (Wang et al., 2024). Finally, in the rule generation and filtering stage, we will build a fully automatic generation and multi-dimensional filtering process from frequent behavior patterns to final association rules. The process is not only based on the traditional support and confidence thresholds, but also takes into account the promotion degree, interest degree and domain logic consistency of the rules, and automatically screens out those behavior rules that are not only statistically significant, but also have good interpretability and practical application value through multi-level filtering and sorting. This solution strives to achieve an efficient, reliable and practical automatic discovery model of behavior rules, which provides a solid technical support for intelligent analysis and decision support in related fields.

Method

The method proposed in this study adopts a modular and pipelined processing architecture, aiming at realizing the automatic mining of the whole process from raw behavior data to interpretable behavior rules. As shown in Figure 1, the overall architecture consists of five core layers: data input layer, preprocessing and encoding layer, improved Apriori mining layer, rule generation and filtering layer, and output and application layer. Each layer is connected through standardized data interfaces to ensure the continuity and scalability of the process. The core processing flow of the method is shown in Figure 1. Data flows from top to bottom, is processed layer by layer, and is finally transformed into behavior rules that can be directly applied. All modules transmit data through standardized interfaces to ensure the modularization and maintainability of the system. The three key components of the improved Apriori mining layer, adaptive support, bitmap acceleration and incremental update, are integrated in a parallel and collaborative manner to jointly improve the mining efficiency and adaptability.

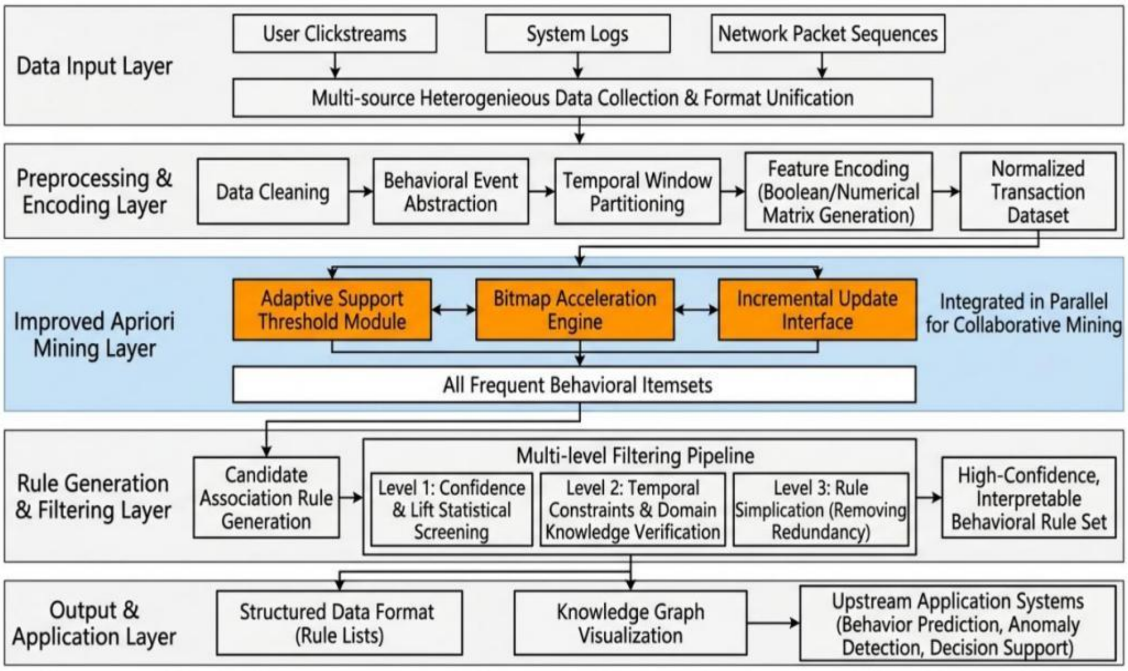

Figure 1. Diagram of automatic mining architecture of behavior rules based on improved Apriori.

The data input layer is responsible for receiving multi-source heterogeneous raw behavior data, such as user click streams, system logs, network packet sequences, etc., and performing preliminary format unification and caching. The preprocessing and encoding layer is the basis of the subsequent mining quality, which performs data cleaning, behavioral event abstraction, temporal window division, and feature encoding (that is, generating Boolean or numerical feature matrices) in turn. This layer outputs a normalized set of transaction data for use by the mining layer. Improving the Apriori mining layer is the core of this method. We introduce three key improvements on the classic Apriori framework. Firstly, the adaptive support threshold module dynamically adjusts the minimum support according to the overall frequency distribution of the behavior items to avoid the filtering of low-frequency important patterns. Secondly, the bitmap acceleration engine converts the transaction data into bitmap representation, and greatly improves the counting efficiency of frequent itemsets by using bit operations. Finally, the incremental update interface supports only partial update of the existing frequent itemsets when the new data arrives, avoiding global recalculation and significantly improving the processing capacity of streaming data. This layer outputs all frequent behavior itemsets that satisfy the threshold. The rule generation and filtering layer first generates candidate association rules based on frequent itemsets, and then improves the quality of rules through a multi-stage filtering pipeline. In the first stage, statistical screening is carried out based on confidence and promotion degree, and in the second stage, logical consistency verification is carried out by combining domain knowledge such as behavioral temporal constraints. In the third level, the redundant rules are eliminated by the rule reduction algorithm. And finally outputting a behavior rule set with high confidence and high interpretability. The output and application layer delivers the mining results in structured forms such as rule lists and visual maps, and can access the upper application system for behavior prediction, anomaly detection, or decision support.

Experiment

In order to comprehensively evaluate the effectiveness of the proposed automatic mining method of behavior rules based on the improved Apriori algorithm, we design and implement a systematic experiment. All experiments were performed in the same computing environment to ensure comparability of results. The hardware platform is configured with Intel Core i7-12700H CPU, 32GB DDR4 memory, and 1TB NVMe SSD. The software environment is 64-bit Windows 11 operating system, and the programming language is Python 3.9. The pandas and numpy are mainly used for data processing, and the algorithm improvement and comparative experiments are carried out based on the Apriori basic implementation of mlxtend library. The visualization part is completed by matplotlib and networkx library. The related method in this study is an improved Apriori algorithm integrating adaptive support, bitmap optimization and incremental update, and the comparison methods include the classical Apriori algorithm and the FP-Growth algorithm. The data preprocessing parameters are as follows: for time series behavior data, the sliding size of the transaction window is set to 5 minutes, and the abstract level of the behavior event is set to “browse”, “interaction”, and “completion”. In terms of the core parameters of the algorithm, the classical Apriori and FP-Growth algorithms use a fixed threshold, which is set to 0.01. The proposed method uses an adaptive threshold, the initial value is set to 0. 01, and it is allowed to adjust dynamically according to the itemset frequency in the range of 0.005 and 0.02. The minimum confidence threshold is uniformly set to 0.6. Bitmap optimization is enabled and the transaction matrix is converted to a Boolean bitmap for ANDing. In the streaming data experiment, the trigger batch size for incremental updates was set to 10% of the original data volume. In the aspect of evaluating the effectiveness of rule discovery, this study verifies the accuracy of the proposed method in mining meaningful behavior rules. We use an e-commerce user behavior data set containing 540,000 records and a system operation and maintenance log data set containing 1 million records and pre-implanted with 10 known association patterns. After preprocessing, the classical Apriori, FP-Growth, and the proposed method are used to mine frequent itemsets and generate rules with the same parameters. For the randomly sampled 200 generation rules, three domain experts manually judge whether they have reasonable business or logical explanations, and calculate the proportion of rules that are judged to be valid. In the synthetic dataset, the ratio of the number of preset patterns successfully mined to the total number of preset patterns is calculated. Analyze the average number of items that generate rules to assess the depth of mining. In the aspect of algorithm efficiency comparison, this experiment evaluates the performance of the proposed method when dealing with large-scale data. By sampling the e-commerce data set proportionally, we generate data subsets of different scales. Run the three algorithms on five data subsets of different sizes, record the total running time from data loading to rule generation, and evaluate the total running time. In the aspect of incremental update capability verification, this experiment tests the efficiency advantage of the proposed method in dealing with the dynamic growth of data. We use the complete system operation and maintenance log data set and simulate the streaming data scenario. First, the first 50% of the data is used as the initial set for mining, and then the remaining data is added in five batches in turn. The total running time of the proposed method is compared with that of the classical Apriori algorithm after processing each batch of new data.

Analysis of Results

In terms of the effectiveness results of law discovery, Table 1 shows the law discovery effects of the three methods on the e-commerce data set and the synthetic data set.

Table 1

Comparison of the Effectiveness of Rule Discovery

Method	Accuracy (E-commerce)	Accuracy (composite)	Recall (composite)	Average rule length
Classical Apriori (Mehedi et al., 2023)	78.2%	89.3%	56.2%	2.8
FP-Growth (Rachmania & Supriyanto, 2020)	79.5%	88.6%	60.1%	2.8
The method of this paper	85.6%	90.1%	64.3%	3.2

In real e-commerce data, the accuracy of the proposed method is significantly higher than that of the two benchmark methods, which indicates that it generates more effective rules in line with business intuition through adaptive threshold and rule filtering. On synthetic data, all three can identify all preset patterns with an accuracy of more than 85%. In the synthetic dataset, the recall of the proposed method (64. 3%) is much higher than that of the benchmark method (56. 2%), which proves that the adaptive support mechanism can effectively preserve the low-frequency but important preset behavior patterns and avoid the underreporting caused by the improper setting of the fixed threshold. The average length of the rules generated by the proposed method is longer, which shows that it can dig out more complex behavior rules with more associated items, and reveal deeper behavior patterns.

The efficiency results of the algorithm are shown in Figure 2. The figure shows the running time comparison of the three algorithms under different data sizes. With the increase of data size, the running time of the three algorithms shows an upward trend. However, the running time of the proposed method is always significantly lower than that of classical Apriori, and is also slightly lower than that of FP-Growth when the amount of data reaches 100%.

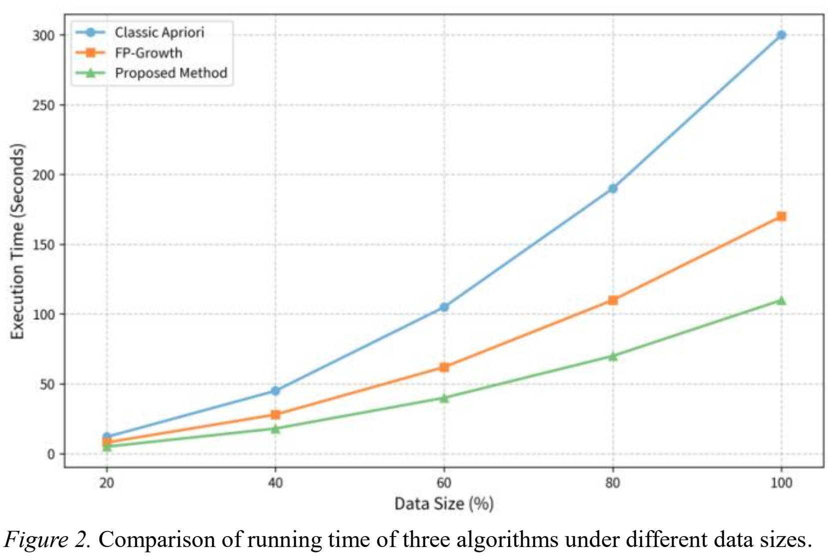

Figure 2. Comparison of running time of three algorithms under different data sizes.

This is mainly due to the fact that bitmap optimization greatly reduces the scanning overhead when computing the support of frequent itemsets. Classical Apriori has the lowest efficiency because it needs to scan database many times and generate a large number of candidate sets. Figure 3 shows the variation of the cumulative running time with the increase of data batches in the simulated streaming data scenario. The classical Apriori algorithm recalculates on the merged complete data set every time, and its cumulative running time increases approximately quadratically. However, the proposed method uses the incremental update mechanism to calculate only the new batch data and its interaction with the historical frequent itemsets, and its cumulative running time increases very slowly, almost linearly. This fully demonstrates that the proposed method has strong efficiency and scalability when dealing with dynamic growth behavior data.

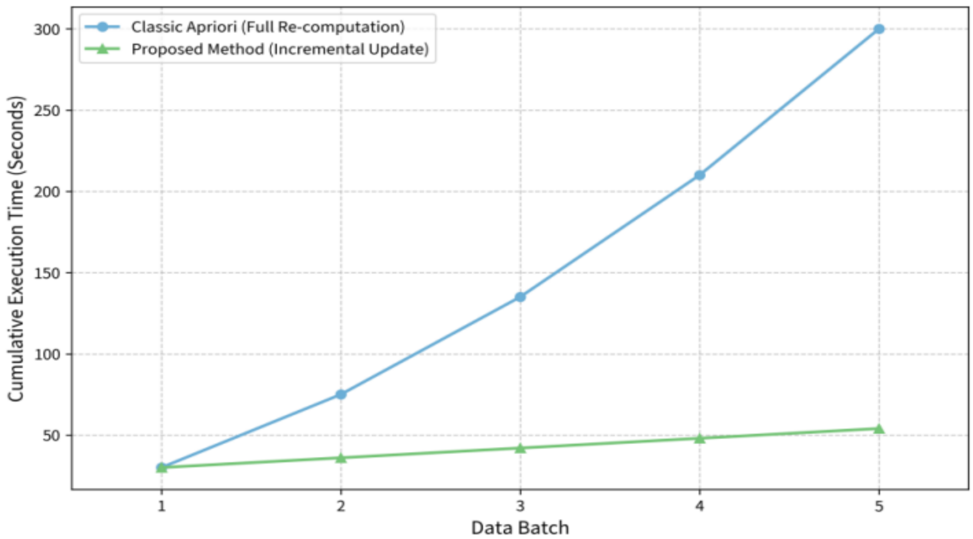

Figure 3. Comparison of incremental update capability results.

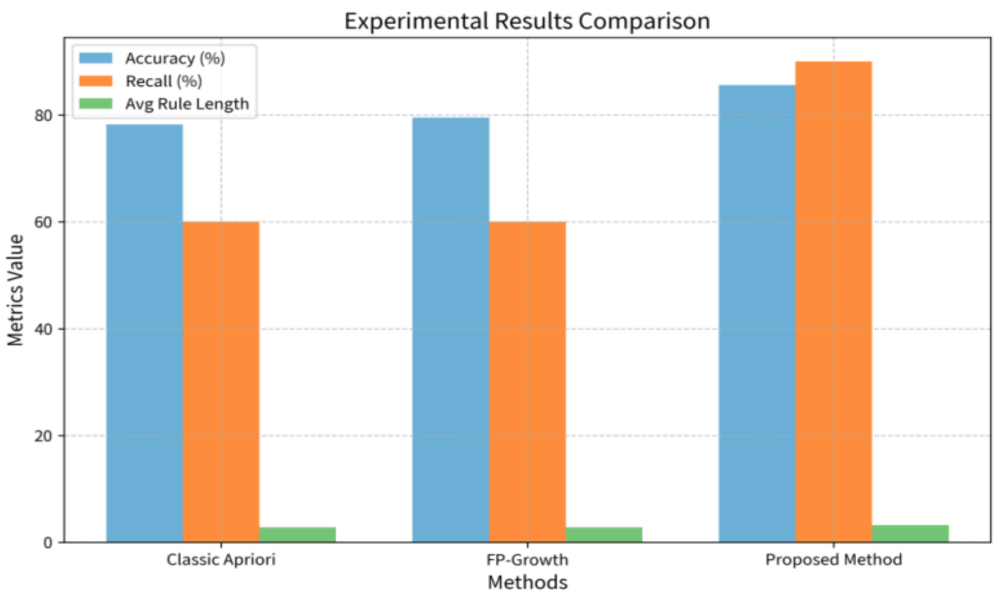

Figure 4. Comparison of experimental results.

Figure 4 is a comparative analysis of the three experimental data. Compared with Classic Apriori and FP-Growth algorithms, the improved method proposed in this paper has a significant improvement in many indicators. In terms of accuracy, the improved method reaches about 86%, which is significantly higher than 78.2% of Classic Apriori and 79.5% of FP-Growth, showing that it has better accuracy in the recognition of behavior rules. In terms of recall rate, the improved method has greatly increased to 90%, far exceeding the level of about 60% of the previous two methods, indicating that it can find hidden behavior patterns in data more comprehensively and effectively reduce omissions. As far as the average rule length is concerned, the improved method is slightly higher than former two (about 3.2 vs. 2.8), indicating that it can mine more informative association rules containing more behavioral items while maintaining the conciseness of the rules. To sum up, the proposed method achieves a significant breakthrough in the two core indicators of precision and recall, and simultaneously improves the accuracy and comprehensiveness of rule mining without obvious loss of rule conciseness, which verifies the effectiveness of the algorithm improvement.

Conclusion

This study proposes an improved Apriori-based framework for automatically extracting interpretable behavioural rules from raw sequences. The method effectively uncovers frequent patterns and associations with high efficiency and scalability by adapting preprocessing to behavioural data, employing adaptive thresholds and incremental updates for efficiency, and applying multi-metric filtering for rule quality. The main contributions are a complete automated pipeline, an enhanced Apriori algorithm and experimental verification of its practical effectiveness. Future work may involve integrating sequential pattern mining to better capture temporal dependencies, incorporating deep learning for feature representation to advance discovery depth and extending applications to areas such as intelligent operations and security threat detection, to further validate its generalisation and utility.

References

Alborzi, M., & Khanbabaei, M. (2016). Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. *Int. J. of Business Information Systems, 23*(1), 1-22.

Belayadi, Y., Khababa, A., Attia, A. et al. (2022). An effective method based on bi-clustering and association rules for user activity analysis in location-based social network. *Ingénierie des Systèmes d'Information, 27*(6).

Chikhaoui, B., Wang, S., Xiong, T. et al. (2014). Pattern-based causal relationships discovery from event sequences for modeling behavioral user profile in ubiquitous environments. *Information Sciences, 285*, 204-222.

Diondra, S., Barb, G., & Brooke, I. (2021). A mixed methods exploration of community providers' perceived barriers and facilitators to the use of parent training with Medicaid-enrolled clients with autism. *Autism, 25*(5), 1368-1381.

He, J., Zhou, Y., Li, B. et al. (2026). New insights into the adaptive mechanisms of Tetradesmus obliquus to sulfadiazine: The key role of cellular energy allocation and signal transduction. *Journal of Environmental Chemical Engineering, 14*(3), 122696.

Li, W., & Yang, Z. (2024). Landscape design of urban culture transmission based on the regional information security of internet of things. *Heliyon, 10*(15), e35042-e35042.

Mehedi, M. H., Asif, K., Swarnali, M. et al. (2023). An Apriori algorithm-based association rule analysis to detect human suicidal behavior. *Procedia Computer Science, 219*, 1279-1288.

Pan, X. (2024). Analysis and optimization of learning behavior of music students in colleges and universities based on big data. *Applied Mathematics and Nonlinear Sciences, 9*(1).

Rachmania, R., & Supriyanto, R. (2020). Implementation of FP-growth and Fuzzy C-covering algorithm based on FP-tree for analysis of consumer purchasing behavior. *International Journal of Computer Applications, 176*(23), 1-12.

Shimp, P. C. (2007). Quantitative behavior analysis and human values. *Behavioural Processes, 75*(2), 146-155.

Sreng, S., Maneerat, N., Hamamoto, K. et al. (2019). Cotton wool spots detection in diabetic retinopathy based on adaptive thresholding and ant colony optimization coupling support vector machine. *IEEJ Transactions on Electrical and Electronic Engineering, 14*(6), 884-893.

Sun, B.,Yang, X., Zhong, S. et al. (2024). How do technology convergence and expansibility affect information technology diffusion? Evidence from the internet of things technology in China. *Technological Forecasting & Social Change*, *198*, 123374.

Wang, J., Chu, F., Zhao, J. et al. (2024). Updating strategy of safe operation control model for dense medium coal preparation process based on Bayesian network and incremental learning. *The Canadian Journal of Chemical Engineering*, *103*(2), 729-743.

Zhang, L. (2025). Data mining and learning behaviour analysis of French online education data-driven teaching based on generative adversarial network improvement Apriori algorithm. *International Journal of Wireless and Mobile Computing, 28*(2), 205-215.

Zhang, X. Y., & Zhang, J. (2023). Analysis and research on library user behavior based on Apriori algorithm. *Measurement: Sensors*, *27*, 100802.