Research on the University Fall Enrollment Strategies Based on Apriori Algorithm

Xia Gao¹, Shucheng Liu¹ and Fangqin Xu¹
¹College of Information Technology, Shanghai Jian Qiao University, Shanghai, Shanghai, 201306, China
¹41080@gench.edu.cn ¹Corresponding author’s e-mail: sc_liu@outlook.com

Abstract. Using the historical enrollment data of colleges and universities to mine, discover the relationship between the potential data, and use the results of data mining to provide reasonable suggestions for enrollment policies, which is of great significance to college enrollment. This paper aims at the allocation of enrollment colleges and the allocation of college students in the process of autumn enrollment in colleges and universities, through the Apriori algorithm in data mining, mining historical enrollment data for 3 consecutive years after data cleaning. Based on the frequent itemsets obtained from the experimental results, a reasonable allocation of college places and student college allocation plan was made.

1. Introduction
With the continuous development of information technology, people's focus on information acquisition has gradually shifted from the amount of data to the amount of information. Therefore, the data mining technology based on the extraction and analysis of effective data has ushered in rapid development. With the development of higher education, the traditional experience-based college admissions decision-making has been unable to meet the high efficiency of admissions. This article uses association rule mining in data mining to analyze and mine Shanghai Jian Qiao University’s fall enrollment data over the years to discover the association between the admission college and the region and scores, and to provide data support and suggestions for the school’s fall enrollment decision in the future.

2. Association rule mining concept

2.1. Basic concepts
Association rule mining is used to discover associations between itemsets in large amounts of data.

If there are associations between two or more items, the attributes of one item can be predicted based on the values of other attributes.

2.1.1. Items and itemsets
The indivisible minimum unit information in the database is called an item, and the set of items is called an item set.

2.1.2. Support of association rules
The support degree of the association rule is the ratio of the number of transactions containing both X and Y in the transaction set to the number of all transactions, which reflects the frequency of the items contained in X and Y appearing in the transaction set at the same time.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
\[ Support(X \Rightarrow Y) = support(X \cup Y) = P(X|Y) \] (1)

2.1.3. Confidence in association rules
The confidence of the association rule is the ratio of the number of transactions containing both X and Y to the number of transactions containing X in the transaction set, which is denoted as confidence \( (X \Rightarrow Y) \), which reflects the conditional probability of Y appearing in the transaction containing X.

\[ Confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} = P(Y | X) \] (2)

2.1.4. Minimum support and minimum confidence
The minimum support threshold (min_sup) and the minimum confidence threshold (min_conf) are specified by the user in order to meet certain requirements. Among them, min_sup describes the minimum importance of the association rules, and min_conf specifies the minimum reliability that the association rules must meet.

2.1.5. Frequent itemsets
The support of item set U on data set T is the proportion of transactions containing U in T, i.e:

\[ Support(U) = \frac{||\{t \in T | \exists \subseteq U\}||}{||T||} \] (3)

For item set I, all item sets in the transaction database T that meet the minimum support specified by the user are called frequent item sets.

2.2. Apriori algorithm
Apriori algorithm is the most famous association rule discovery method. The basic idea is to find the frequent itemsets through multiple scans of the database’s support for computer item sets, thereby generating association rules. The Apriori algorithm performs multiple scans of the data set.[1] The first scan obtains a set of frequent 1-item sets \( L_1 \). The k-th scan first uses the result of the kl-th scan \( L_{k-1} \) to generate a set of candidate k-item sets \( C_k \), and then determines the elements of \( C_k \) during the scan support, and finally calculate the set of frequent k-item sets \( L_k \) at the end of each scan. The algorithm ends when the set of candidate k-item sets \( C_k \) is empty[2].

3. Association rule mining

3.1. Data sets and data cleaning
The data used in this article comes from the real autumn recruitment data of Shanghai Jian Qiao University in 17, 18 and 19 years, and the original data is in EXCEL format.

The original data set was processed for the following aspects:

- Eliminate information that has no practical meaning, does not affect the mining results and affects efficiency, such as name, ID number, home address, etc., and finally selects gender, birthplace, department, language score, math score, English score, and registration Case 7 valid fields.[3]
- Divide Chinese, mathematics, and foreign language performance into five grades: A (150-135), B (134-120), C (119-105), D (104-95), and E (94-0). The data was quantified.
- Eliminate incomplete and erroneous data to ensure validity.
- Divided into 2 data sets with registered and unregistered as marks.

The result after the data cleaning takes Table 1 as an example, a total of about 13,000 valid transactions are obtained, of which about 12,000 have been reported and about 1,000 are unreported.[4]
3.2. Experimental process and result analysis
Based on the magnitude of the data, set min_sup to 5 and perform data mining on the unreported data set to obtain 20 frequent item sets. Set min_sup to 20, and perform data mining on the reported data set to obtain 80 frequent item sets. The results are as follows.
The experimental results show that the number of student registrations is highly correlated with the college. In the three years of 2017-2019, students in Guizhou Province and Guangxi Zhuang Autonomous Region were the most frequently reported in business schools, and these two provinces reported frequent itemsets also has connections with business schools. One of the reasons is the number of students enrolled, and the second is the issue of school recognition. Therefore, you can consider doing more publicity for admissions in these two provinces. It can also be seen that there is only one province of Yunnan in the unreported frequent item set of the letter school. The unreported situation of the School of Journalism and Communication in the Guangxi Zhuang Autonomous Region is relatively frequent, and it does not appear in the registered frequent item set, so it can appropriately reduce the enrollment of the local college. It is noted in the table that in all cases where the situation is not reported frequently within 3 years, the students' math scores are E grade, that is, fail, and the language is D grade or above, that is, pass, in the future admission, you can combine candidates Students should be considered comprehensively in terms of student origin and mathematics and Chinese language performance. Students who pass the Chinese language but fail the mathematics should be admitted carefully, and the number of liberal arts students should be appropriately reduced.

Among the frequent items that have been reported, business schools and information technology schools account for a relatively high proportion of colleges. One of the reasons is that the employment prospects of finance and the Internet are better, and the number of admissions can be increased appropriately. And it can also be seen that in the frequent event concentration in Shanghai, the School of Art and Design and the Jewelry College are only related to Shanghai and have high recognition. Therefore, when enrolling, you can consider transferring students from Shanghai to these two colleges.

4. Conclusion
A large number of college admissions historical data have huge potential value, and their effective use can help the admissions decision-making. This article uses the Apriori algorithm in data mining to mine association rules for the three-year fall enrollment data of Shanghai Jianqiao College, visually displays the results, and makes recommendations for the allocation of college places and student colleges. The experiment is based on a real data set that guarantees the validity and consistency of the data. The results have a high degree of credibility and can support future admissions policies.

However, the classic Apriori algorithm used in this paper is not efficient. In the process of increasing the minimum support, the support of the item set will decrease.

In future research, we can continue to increase individual items, increase the amount of data, and perform more detailed cleaning of the data set, so as to mine more valuable information and provide support for future admissions decisions.

References
[1] Xu, JJ., Zhang, GH. (2020) Application and Practice of Data Mining Algorithms Based on Apriori. Computer Technology and Development, 30: 206–210.
[2] Liu, H. (2019) Association Analysis Based on Apriori Algorithm. China Computer & Communication, 19: 132-136.
[3] Gao, X., Xu, FQ., Zhu, ZM. (2019) The Application of Improved FP-Growth Algorithm in Disease Complications. In: Advanced Science and Industry Research Center. Proceedings of 2019 International Conference on Computational Modeling, Simulation and Optimization(CMSO 2019).Beijing.pp.118-122.
[4] Li, ZY. (2009) Deep Analysis of Apriori Algorithm in Association Rules. Modern Computer, 307: 17-19.
[5] Xu, FQ., Lu, HF. (2017) The application of FP-growth algorithm based on distributed intelligence in wisdom medical treatment. International Journal of Pattern Recognition and Artificial Intelligence, 31:4.