Research and Practice of Association Rules Method to Solve the Problem of Allocating Places for Spring Major in Colleges and Universities——Taking Shanghai Jianqiao University as an Example

Lei Zhang1*, Yuqian Zheng1
1Admissions Office, Shanghai Jian Qiao University, Shanghai,201306, China
*Corresponding author’s e-mail: zhanglei@gench.edu.cn

Abstract. Aiming at the problem of the allocation of enrollment majors and the allocation of majors of students applying for admission in the spring recruitment of colleges and universities, through the FP-Growth algorithm in data mining, the relationship between the attributes, majors and reports of students applying for in-depth mining will be able to describe a class of students. The attributes and assignments of majors and reports are used as an item set, and its frequent patterns are tapped. Then, according to the set support degree, a set that can reflect the characteristics of students is selected, and a model is built based on this result. At the same time, this paper improves the evaluation criteria of the usual school registration situation. Through the calculation of variance, the overall professional distribution and registration rate are evaluated. Experimental results show that the model has good performance and can reasonably recommend students' professional assignment.

1. Introduction
In enrollment work, improving the rationality of enrollment decisions is the most important part. With the development of science and technology, data mining technology is widely used in the process of finding potential rules and useful information from the huge amount of data, and serves all walks of life. This article will use the association rule method in data mining technology to analyze and mine the spring enrollment data of Shanghai Jian Qiao University over the years, find the tendency of different students in the process of choosing majors, and provide support and suggestions for the school's spring enrollment decision in the future.

2. Association Theory
Association rules are mainly to study the occurrence of a certain event and the occurrence of other events, the rules and relationships between the two, then use digital quantification to analyze the influence between the two, and finally select the minimum credibility and minimum Association rules for support rates. In addition, when the relevant personnel start the mining work, they first need to find the high-frequency project groups from the database, and then select the association rules from these groups.

2.1. Basic concepts of association rules
Some terms are involved in the association rules. The objects analyzed in the association rules are called “items”, and the set of “items” is called “item sets”. The number of occurrences of an item set is called the frequency of the item set, and the frequency of occurrence is called the degree of support. The
formula is shown in formula (1). When the support of a certain item set is greater than a certain threshold set by a person, it is called frequent item set.

\[
\text{support}(X) = \frac{\text{count}(X)}{|D|} \times 100\%
\]

(1)

The association rules are divided into two parts. First, the frequent itemsets used are found through the set threshold, and then the association rules are found from the frequent itemsets. Here, confidence needs to be introduced as a metric value. The calculation method of confidence is as shown in formula (2).

\[
\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}
\]

(2)

For a frequent item set X, take a non-empty true subset Y, assuming \(S = X-Y\), if \(\text{confidence}(S \rightarrow Y) \geq \text{minconf}\), then \(S \rightarrow Y\) is an association rule, where \(\text{minconf}\) is the artificially set minimum confidence.

2.2. Apriori Algorithm
The Apriori algorithm is one of the most representative algorithms in the association rule algorithm[1]. Through other association rule mining algorithms in the “generate-test” line, through continuous iteration of the production candidate set, combined with the minimum scale technology and a priori principles, the Set of outgoing candidates. However, the Apriori algorithm has its drawbacks. Every time a candidate set is generated, the data set needs to be scanned. In addition, the algorithm will form a large number of candidate sets, which brings very large I/O overhead and time and space occupation.

2.3. FP-Growth Algorithm
FP-Growth (Frequent Pattern Growth)[2] is an association analysis algorithm proposed by Han Jiawei and others in 2000. The FP-Growth algorithm is mainly divided into two steps: FP-tree construction and recursive mining FP-tree. The FP-tree construction uses two data scans[3], starting with the item with the smallest support count in the Head table, adopting a divide-and-conquer strategy, determining the conditional pattern base through all prefix paths, constructing the FP-tree, and generating frequent items[4]. The FP-growth algorithm compresses the transaction information in the database into the FP-tree for storage, reducing the huge I/O overhead caused by scanning the database, and improving efficiency in both space and time.

3. Enrollment Association Rules Mining Based on FP-Growth

3.1. Data sets and data preprocessing
The student source information that can be obtained during the enrollment process is very complete, but some of the information has different formatting information formats due to the different requirements for each school. For example, the address is not filled in as required, and some information such as the nationality has a system for distinguishing Numbers, etc. At the same time, the information with personal identification information, such as admission ticket number, ID number, name, etc., will not be of any help in mining, and it will also lead to personal information security issues that need to be deleted.

This article uses data cleaning and data integration[5] to perform data processing on the source information. First, the data extracted from the three-year data were selected as gender, ethnicity, political outlook, year of birth, assigned department, assigned major, grade, language scores, math scores, foreign language scores and registration status as the data extracted by the association rules; The professional direction in the course was removed, the major categories were retained, and the scores of Chinese, mathematics, and foreign language were divided into the highest and lowest scores, and the average was divided into six levels; the students who actually registered were marked as registered, but not registered.
Of students are marked as unregistered. After marking the grades, the three-year data is aggregated into a data set.

3.2. Model design
After the data is preprocessed, frequent items are mined on the data set through the FP-Growth algorithm. This is a fast frequent item mining algorithm. After the support level is set, the frequent item set matching the support level will be raised from the data set.

Figure 1 Algorithm flowchart

After the frequent itemsets are extracted, in order to guide the enrollment work based on past experience, a professional assignment model for recruiting frequent itemsets based on the FP-Growth algorithm is designed. The flow chart is shown in Figure 1. After importing new enrollment information, according to the characteristics of the student, the model can automatically adapt to the major that meets its characteristics, and proposes experience-based suggestions for the student's major assignment. The workflow of the model is shown in the figure. After the new data has produced its results, it will form a new frequent item set through FP-Growth to provide support for its subsequent data set, so this is a constantly improving model.

3.3. Evaluation standard
This article mainly studies the use of association rule algorithm to solve the problem of division of majors in the enrollment process, and the main measure of evaluating this algorithm is the registration rate of each major, comparing the actual number of students registered $N_{ci}$ and the actual number of admission letters issued $N_{total}$. The calculation method of the registration rate is as shown in formula (2).

$$P = \frac{N_{ci}}{N_{total}} \times 100\%$$

(3)

The measurement of the registration rate needs to be combined with the whole school registration rate $P_{total}$ to evaluate the professional registration rate $P_t$, and compare the professional registration level in each professional. To measure the distribution of professional places in the school, the method of calculating discrete values is used, and the calculation method is as shown in formula (4).

$$F = \sum_{i=1}^{n} [(P_i - P_{total}) \times 100]^2$$

(4)

4. Experimental results and analysis
The algorithm in the experiment is written in Python voice, and the storage of the recommended rules is in the form of an Excel table. The hardware configuration of the experiment is 16G memory, 1T hard disk, Intel(R) Core(TM) i5-6500 CPU @ 3.2GHz, operation The system is Windows 10. The experimental program is developed using the Python programming language, and its operation diagram is shown in Figure 2.
Figure 2 Running Process

The experimental results show that the registration rate has the highest correlation with majors. Among all the majors opened in 2017-2019, the registration rate of network engineering majors has maintained 100% year after year, and the registration rate of computer science and technology has also remained at 100% in 2018 and 2019. The mechanical design and manufacturing and automation majors, as the major with the largest number of enrolled students for three consecutive years, perform better than other majors. In addition, there is a correlation between the registration rate and the student's score. The results show that among the students who register, most of the Chinese and mathematics scores are good or average, but the foreign language scores are poor; In the fourth and fifth grades, the sixth-grade student registration rate is 100%, and the unregistered students in the first, second, and third grades account for nearly 60% of the total unregistered students.

At the same time, during the experiment, after trying to extract frequent itemsets from 2017 and 2018 data, the model is used to make professional allocation suggestions for some data in 2019. The experimental results show that the model performs well in some majors that have been established for three years. Almost 80% of the allocations in the majors are the same as the real results. At the same time, in these data, the percentage of students registering has reached 99%. This shows that the professional distribution model generated by using the FP-Growth algorithm is feasible.

According to the evaluation criteria, the registration rates for 2017, 2018, and 2019 are shown in Table 1, and the change charts of the registration rates of some majors are shown in Figure 3. It can be seen that the overall registration rate in 2019 has increased compared with the previous two years, in which the registration rate of engineering majors is relatively high, while the liberal arts has declined. In addition, the discrete situation change trend chart is shown in Figure 4. The three-year professional quota allocation metric F is 3.75, 7.66, and 1.46, respectively, and the overall performance in 2019 is better.

Through the experimental results of Figures 3 and 4, the following recommendations are made for the allocation of professional places in the future. First of all, when designing the enrollment plan, the enrollment of engineering students can be appropriately increased, and the enrollment of liberal arts, especially linguistics, can be appropriately reduced. In addition, in the result of frequent item extraction, no frequent items related to the gem major among the registered students were found. At the same time, as the only year of enrollment in 2018, the registration rate was 100%, indicating that the major is more recognized and can be appropriate. Increase the number of people planning. At the same time, more attention can be given to students with upper-middle grades. Among the students who apply for the school, some students belong to students with partial or middle grades. When enrolling, you can fully examine the interval of their scores, combined with their application majors, to make a judgment on whether to admit.
Table 1. 2017-2019 annual report rate

| Row labels                                         | 2017 | 2018 | 2019 |
|---------------------------------------------------|------|------|------|
| Gem and material technology                       | 100% |      |      |
| Engineering Management                            | 97%  | 100% |      |
| Business management                               |      |      | 100% |
| Mechanical design and manufacturing and automation | 98%  | 98%  | 99%  |
| Computer Science and Technology                   | 98%  | 100% | 100% |
| Automotive Service Engineering                    | 100% | 96%  |      |
| Japanese                                          | 100% | 97%  |      |
| Digital Media Technology                          |      |      | 100% |
| Network Engineering                               | 100% | 100% | 100% |
| Journalism                                        |      | 97%  | 100% |
| English                                           | 100% | 98%  |      |

5. Summary
In this paper, the FP-Growth algorithm is used to mine association rules for the processed spring admission data, and the evaluation standard registration rate and discrete registration rate are designed. The experimental results show that this paper is feasible for the data preprocessing method, and the FP-Growth algorithm can be used to realize the association rule mining for the data set. At the same time, because the FP-Growth algorithm is mined based on the source data that does not contain personal information, in the future practical application scenarios, the distribution of majors for a certain group of students has higher credibility, can quickly match student characteristics, and improve the rationality of professional distribution.

However, in the experimental process, it was found that through two years of data modeling, in the process of improving support, the number of candidate sets will drop sharply, and when the third year of data is added, the number will increase. Therefore, in the future research work, it is necessary to continue to increase the attributes that can describe the individual, at the same time increase the amount of data, further improve the quality of the candidate set, and better serve the process of enrollment planning and student professional allocation.

References
[1] Xu, Y., Zhan, R.Z., Tan, G. 2019. An Improved Apriori Algorithm Research in Massive Data Environment. Shenyang, 21-29
[2] Zeng, Y., Yin, S.Q., Liu, J.Y., Zhang, M. 2015. Research of improved FP-growth algorithm in association rules mining. Scientific Programming.
[3] Gao, X., Xu, F Q., Zhu, Z M. (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] Xu, F Q., Lu, H F. (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

[5] Ordonez, C. (2011) Data set preprocessing and transformation in a database system. Intelligent Data Analysis, 15(4):613-631.