Traction of Students' Curriculum Information Based on Association Rule Optimization

Zhe Zhou[1, a], Yong Ouyang[1], Yawen Zeng[1]

[1]Hubei University of Technology, Wuhan 430064, China

[a]1755486005@qq.com

Abstract. In colleges and universities, the data on student' s performance are numerous. However, these data are often not well utilized and only served as a query. In order to make better use of these data, this paper will improve the traditional association rules and dig into the course performance information. Through the analysis of the association rules of students' different grades, the correlation between the courses and their traction are sought. The test results show that the optimized data mining algorithm has a good mining effect on the data, which can highlight the importance of different courses and make the data better reflect the links between curriculum and other courses. In this way, it is convenient for the teaching management department to better arrange the order of courses for students, so as to provide support and help for the teaching activities of students and teachers.

Keywords: Apriori algorithm, credit weighting, correlation analysis.

1. Introduction

With the advent of the era of big data, many fields are inseparable from data, such as commodity sales data analysis [1], Portuguese language prediction [2], etc., which is more so in the field of teaching management [3, 4, 5] . In recent years, the number of students has increased day by day, but these data about course performance information often exist in the database, serving as a query. In order to make more effective use of data, extract deeper meaning, find out the effective information in the course performance data, and assist the teaching management department, it is necessary to use the data mining algorithm to conduct in-depth mining of these data [6, 7] .

However, in the traditional data mining methods, there are many problems. First of all, the amount of data used is relatively small, among which only hundreds of data are used for analysis [8].In addition, the scarcity of data often leads to errors and the conclusions are also very contingent, which is not enough to support the experimental conclusion needed. Secondly, these researchers tend to pay too much attention to find the link between the scores in the course performance information and ignore the importance of the course itself [9], which fails to find the gap between important courses and non-essential courses. So it will cause a certain degree of inaccuracy.

Therefore, this paper improves the method of traditional association rules, weights the importance parameters of the course itself in various ways [10], and conducts a comprehensive research and analysis on the student's grades, so as to better find the connection among different courses, make the most of it and promote the development of teaching activities.

2. Data Preprocessing

2.1 Data Cleaning.

Fortunately, this subject has obtained statistics on all course performance information of our school in the past ten years, including basic courses, professional courses, elective courses, etc.

The huge amount of course grade information data brings a lot of trouble to data preprocessing, and the data itself has many problems. For example, some value is missing, some courses have the same name, and there are multiple names in one course, so it needs to be cleaned accordingly.

In order to make the research data more relevant, different record tables are merged and data of the same profession is selected for association. For the confidentiality of the data and the convenience of

Copyright © 2019, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

writing, the student's name is replaced by N, and the course name is replaced by a letter (Probability Theory and Mathematical Statistics (A), Advanced Mathematics-1 (B), Advanced Mathematics-2 (C), Discrete Mathematics (D), Data Structure (E), Introduction to Computer Science (F), Database Principles and Applications (G), Digital Logic (H), Computer Composition Principles (I), Advanced Language Programming (J), Software Engineering Course Design (K), Algorithm Analysis and Design (L), Java Programming (M), Introduction to Software Engineering (N)).

After integrating the original form and removing the cumbersome features, the duplicated data is merged in a table and unify the course names of different names in the same course. In this paper, some missing data are directly replaced by the average value. While for the part with more missing values, the data set will be deleted, so as to avoid the influence of missing values on the observation results. The pretreatment results are shown in Table 1 (red data is missing).

Table 1 Post-filled course transcripts

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
N1	80	63	87	60	75	90	77	65	65	80	87	64	66	77
N2	88	83	73	84	85	73	81	91	91	73	85	71	93	90
N3	69	80	66	74	76	87	79	77	77	72	86	74	77	72
N4	64	80	64	35	90	76	78	62	62	71	85	63	60	77
N5	73	70	86	61	89	89	82	69	69	94	75	77	80	72
N6	82	82	77	72	79	84	86	90	90	82	92	74	78	77
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

2.2 Data Reconstruction.

Graphical processing of the number of course results after the data cleaning is completed, the results are often presented similar to the situation shown in Fig. 1. The amount of data on the left side is more concentrated. The amount of data in the red part on the right side is relatively sparse.

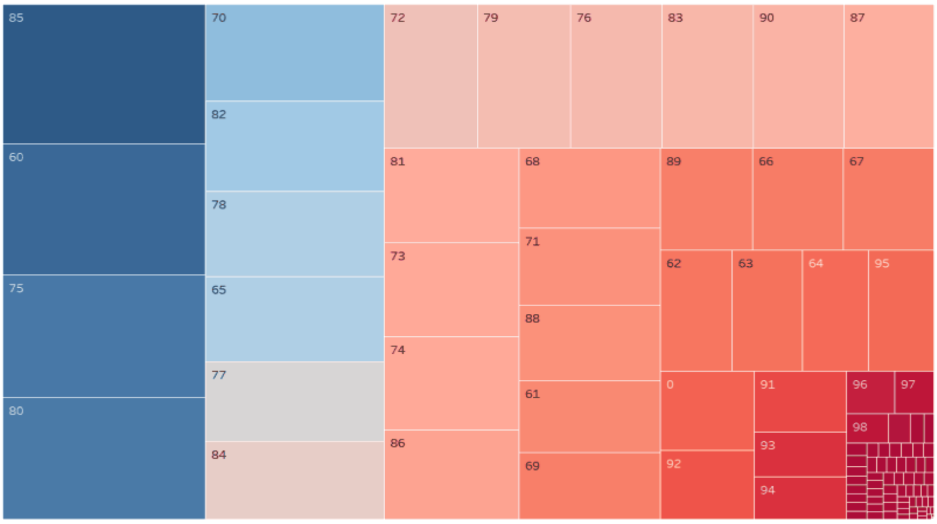


Fig. 1 Score distribution

In order to better complete the mining requirements of the model, this paper will group the parts of the score aggregation. The k-means algorithm is used to divide the students of different courses according to the corresponding scores, so as to facilitate the association rules.

It is not difficult to find that these intervals have obvious defects, that is, the k-means algorithm is susceptible to the extreme value of the score, so the interval distribution of the k-means algorithm is often uneven. Therefore, the result of the k-means algorithm needs to be processed to obtain only one critical value. According to the obtained critical value, the data of each grade is divided into two parts, a value greater than the threshold is set to 1 and anything less than critical value is set to 0, and the data is correlated by the association rule algorithm. According to the division result of the k-means algorithm, the threshold of the course "Probability Theory and Mathematical Statistics" can be set to 73, that is, the portion larger than 73 is set to 1, and the portion smaller than 73 is set to 0.Among them, in order to better distinguish between different courses, different numbers are used instead of different courses. 1 instead of Probability Theory and Mathematical Statistics, 2 and 3 replace Advanced Mathematics-1 and Advanced Mathematics-2, respectively, and so on. Finally, the results as shown in Table 2 were obtained.

Table 2 Data transcripts after pre-processing

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
N1	1	0	3	0	0	6	0	0	0	10	0	0	0	0
N2	1	2	3	4	5	0	0	8	9	0	0	0	13	14
N3	0	2	0	0	0	6	0	8	9	0	0	0	0	0
N4	0	2	0	0	5	0	0	0	0	0	0	0	0	0
N5	0	0	3	0	5	6	7	0	0	10	0	0	13	0
N6	1	2	3	0	0	0	7	8	9	10	11	0	13	14
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

3. Model and Analysis Process

3.1 Model Establishment.

In order to dig out more effective inter-course information, this paper built the Apriori algorithm model firstly. The Apriori algorithm is divided into two stages, as shown in Fig. 2. The first stage is to divide the data into n parts, and find frequent itemsets from each part, and form a candidate set. The second stage is to find global frequent item set from the frequent itemsets.

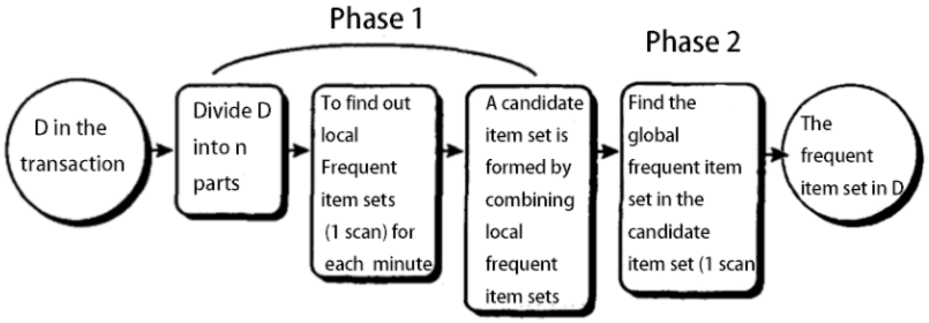

Phase 1
Phase 2
D in the transaction
Divide D into n parts
To find out local Frequent item sets (1 scan) for each minute
A candidate item set is formed by combining local frequent item sets
Find the global frequent item set in the candidate item set (1 scan
The frequent item set in D

Fig. 2 Algorithm flow chart

In order to better observe the relationship between the curriculums, the Apriori algorithm needs to be appropriately simplified and modified, and the upper limit of the elements in the project collection in the algorithm is set to 2, that is, the relationship between the two courses is found only by the algorithm. At the same time, minsup=40% and minconf=70%. The association rules obtained at this time are shown in Table 3.

Table 3 Three Scheme comparing

Rule	Confidence
Database Principle and Application->Advanced Mathematics-1	0.74
Advanced Mathematics-1->Database Principle and Application	0.7
Database Principle and Application->Mathematical Logic	0.72
Mathematical Logic->Database Principle and Application	0.77
Database Principle and Application->Java Programming	0.77
Java Programming->Database Principle and Application	0.83
Advanced Mathematics-2->Discrete Mathematics	0.82
Discrete Mathematics->Advanced Mathematics-1	0.72
Advanced Mathematics-1->Discrete Mathematics	0.7

3.2 Model Optimization.

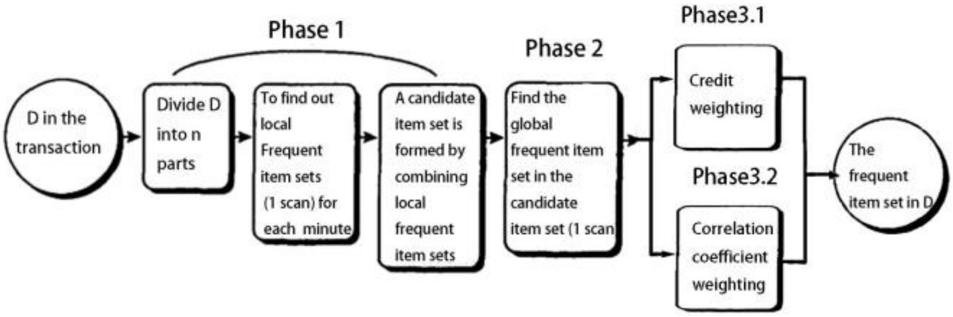
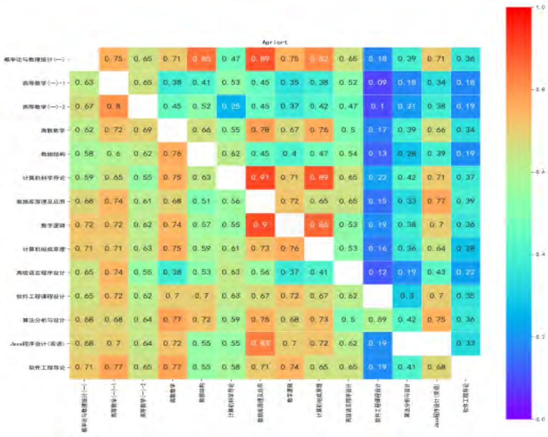

Phase 1
Phase 2
Phase3.1
Phase3.2
D in the transaction
Divide D into n parts
To find out local Frequent item sets (1 scan) for each minute
A candidate item set is formed by combining local frequent item sets
Find the global frequent item set in the candidate item set (1 scan
Credit weighting
Correlation coefficient weighting
The frequent item set in D

Fig. 4 Improved algorithm flow chart

In the research environment of this paper, the original Apriori algorithm only correlates through the scores, and ignores the importance of the curriculum itself and the student's performance, so it will cause a certain degree of deviation. Student achievement can be weighted by the following methods.

(1) Weighted by credits. Important courses tend to have higher odds ratios, such as Advanced Mathematics, Object-oriented Programming, etc., while some course credits are relatively small. Therefore, it is necessary to weight the credits to see the importance of the course. Fig. 5 is an improved correlation matrix diagram. From the figure, it can be seen that after the credits are weighted, the correlation between some low-credit and high-credit courses decreases, and the relationship between some courses is more closely. Therefore, adding credits weights can more effectively show the relationship between courses, which is more objective than traditional methods.

(2) Pre-calculate the weight of each candidate by studying the relevance of the course. The correlation coefficient is a statistical indicator used to reflect the closeness of the correlation between variables. By weighting the correlation coefficients, the choice of candidates can be optimized, and the relevant links between courses can be better analyzed. Correlation is divided into positive correlation and negative correlation, so there are some disciplines with good grades, but the other disciplines have relatively poor performance, which is more reliable.



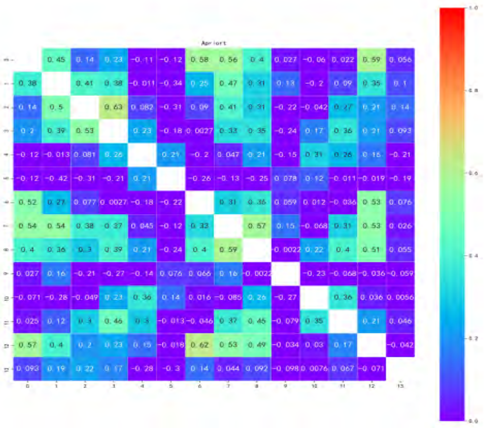


Fig. 5 Improved 1 correlation matrix

Fig. 6 Improved 2 correlation matrix

Probability Theory, Discrete Mathematics, Advanced Mathematics are some basic courses. Learning these courses has a very important positive effect on the learning of other courses, such as Digital Logic and Algorithm Analysis and Design.

Although some courses do not see the connection on the bright side, the credits are also low, but the quality of these courses affects the study of other courses, so it is not possible to relax the study of those subjects with lower credits.

Compared with the two different optimization methods, the first credits are weighted, and the emphasis is on the importance of the curriculum itself. The subjects with high credits are often weighted more. This method is only affected by the importance of the subject, and other factors such as student achievement have less impact on it. The second weighted correlation focuses on the student's performance data, and calculates the correlation of the student's scores after normalization and then weights them. This method is easily influenced by the extreme values of students' performance and it is biased.

4. Conclusions and Prospects

This paper improves the traditional Apriori algorithm by two different optimization methods. Compared with the traditional Apriori algorithm, the optimized algorithm based on the importance and relevance of the course will make the obtained subject relevance data more convincing and practical, and better assist the teaching research. However, the data used in this article is only for some majors, the data has a certain bias. How to collect, analyze and research more data in the future and get more convincing results will be the direction we continue to work hard.

Acknowledgments

Thanks to The 4th China College Students' "Internet+" Innovation and Entrepreneurship Competition（project number：201810500046） for giving me this opportunity and the help of my sister and teacher，Yong Ouyang.

References

[1]. Shao Tingting. Optimization of Weighted Apriori Algorithm and Its Application in Commodity Sales Data Analysis[J/OL].China Business Theory, 2019(04):245-247[2019-03-20]. https://doi.org/ 10.19699/j.cnki.issn2096-0298.2019.04.245.

[2]. Wang Yueqing. Analysis and Decision of Portuguese Language Prediction Model Based on Data Mining[J].Communication world,2019,26(02):268-269.

[3]. Joksimović S, Gašević D, Loughin T M, et al. Learning at distance: Effects of interaction traces on academic achievement[J]. Computers & Education, 2015, 87:204-217.

[4]. Fu Xiangyan, Hu Feng. Application of Apriori Algorithm in Student Achievement Management System[J].Computer Fan,2018(12):186.

[5]. R. S. J. D. Baker, Kalina Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, vol. 1, no. 1, pp. 8-16, 2009.

[6]. Liao Xuanzhi. Application of Data Mining in Curriculum Correlation and Achievement Prediction of Medical Colleges and Universities[J].Information and Computer (Theoretical Edition),2019(02):154-156.

[7]. Wang Hairong. Application of Data Mining in Student Score Analysis[J]. Electronic Design Engineering, 2013(04): 54-56.

[8]. Wang Yumei. Application of Apriori Algorithm in Student Score Analysis[J].Digital Communication World,2018(12):177+181.

[9]. Cai Liuping, Xie Hui, Zhang Fuquan, Zhang Longfei. Study on Big Data Mining Method Based on Sparse Representation and Feature Weighting[J].Computer Science,2018,45(11):256-260.