Advances in Intelligent Systems Research, volume 163

8th International Conference on Management, Education and Information (MEICI 2018)

Research and Application of Mining Algorithm Based on Association Rules in Colleges Management System

Shaorong Feng

School of Information Science and Engineering, Xiamen University, 361005 Xiamen, China

shaorong@xmu.edu.cn

Keywords: Association Rule; Apriori; Colleges Management; Frequent Item; Candidate Item

Abstract. This paper aims at the insufficient of Apriori association rule mining algorithm. To avoid the blind search in the process of mining and enhance the search efficiency in the collection of frequent item, we investigate the improvement of Apriori algorithm. Then, we use the algorithm for mining on data collection of the students' scores of the college of computer in a high school, and achieve good performance. The results show the associations of the scores between courses, and the impact of the sequence of courses on scores. These results can be used for reference and instruction for better works including training schemes and educational reform.

Introduction

With the rapid development of China's higher education and the rapid expansion of enrollment, college resources have become increasingly tense. How to make flexible and reasonable academic arrangements and teaching management becomes more important. How to effectively analyze and process the teaching information of all students in colleges, so as to further optimize the teaching resources and improve the quality of teaching, has become a new subject to be studied.

Data mining is a new type of data analysis method, which extracts valuable information from human data in the information world for human use[1, 2]. In data mining, association rules[3] are one of the main techniques. Association rule mining is to automatically extract many related rules from a large amount of real data through computers. Among the association rule mining techniques, the most classic algorithm is Apriori algorithm which uses the iterative method of layer-by-layer search to find out the relationship of item sets in the database and form rules.

At present, the teaching plans of the most colleges and universities are written by the president of the college. With many years of teaching experience, they will decide what courses to arrange for students and the order of these courses according to related rules. However, due to subjectivity, they will inevitably ignore the importance of the valuable resources of student achievement accumulated over the years. We can conduct data mining on the scores of students in various subjects, find out the association rules of these courses and objectively understand the relationship of them, in order to help the preparation of future teaching plans, better provide reference and guidance for the teaching management department in training program development and teaching reform.

Improved Technologies of Apriori Algorithm

Based on Apriori algorithm, a large number of experts and scholars have conducted in-depth research, and proposed AprioriTid algorithm[4], FP-Tree algorithm[5], Partition algorithm[6], DHP algorithm[7], DIC algorithm[8], PFUP[9]and other improved algorithms, which significantly makes Apriori algorithm be more efficient.

Apriori algorithm improvement. In order to improve the efficiency of Apriori algorithm and reduce the number of times of scanning the database, based on the logic operations of 0-1 matrix and the mapping set, the implicit Information Is Mined By Strong Association Rules. This Is The Principle Of An Improved Apriori Algorithm Based On Mapping And Logic Operations. .

The Improved Apriori Algorithm Greatly Reduces The Number Of Times Of Database Scans. Based on Logic and Mapping Operations, it only needs to perform two overall database scans,

which greatly improves the efficiency of the algorithm.

Improved algorithm application implementation.

Input: database D; minimum support threshold *min_sup*
Output: Frequent item set L of D
STEP1: got Items( ); //Obtain all items for the first scan
second scan(); //The second scan builds 0-1 matrix
gotL_1( ); //Statistics obtains L_1
gotC_2( ); // Logical and operation obtain candidate item set 2
gotL_2( ); //obtain L_2
STEP2: for (k=2; $L_{k-1} \neq \Phi$; k++)
{
L_C=CutBefore(L_k); //Pruning before connection
C_{k+1}=conltems(L_C); //Candidate set C_{k+1} is obtained from the frequent set connection after pruning
L_{k+1}=CutAfter(C_{k+1}); //Pruning after connection
}

Application structure of improved Apriori algorithm in teaching management

In this paper, the improved Apriori algorithm is used to extract and analyze the data and form the decision basis.

Modeling process. In order to more clearly describe the application process of Apriori algorithm, the data mining of the score modeling of the first five semesters of a certain college is used to show the algorithm modeling process.

(1) Data collection, pre-processing and cleaning

Data is collected first, and noise and inconsistent data are eliminated. Records with missing scores are recorded as the average score of the course and records with more scores are recorded as the first score.

(2) Data integration and data filtering

If there are multiple data sources, the multiple data sources should be combined together. This example uses a single data source, and it is only needed to collect the scores of each semester in one file (*.xls). Table 1 shows some origin data.

Table 1 Some origin data

Computer Basics	C Language Programming	Data Structure	Operating System	Database System	…
91	82	95	90	94	…
86	84	83	86	84	…
92	81	95	89	83	…
⋮	⋮	⋮	⋮	⋮	⋮

(3) Data selection

Extract data related to the analysis task from the database.

For this test, the scores of courses, students' numbers and names have no effect on the data analysis. The courses can be roughly divided into basic courses, management courses and computer courses. This test is mainly for the analysis of computer courses, so only the scores of these courses are retained in the data source.

(4) Data conversion and build 0, 1 matrix

Converse the data into a form which is suitable for mining. Apriori algorithm is association rule algorithm of Boolean type, so continuous students' scores are conversed to discrete Boolean type

data (0, 1).

Since the scoring standards are inconsistent of different courses, the conversion method of this test is 1 if the score is greater than the average score of the course; and is 0 if the score is less than the average score of the course. The result after data conversion is shown in Table2.

Table 2 Result after data conversion

Computer Basics	C Language Programming	Data Structure	Operating System	Database System	…
1	1	1	1	1	…
1	1	1	1	1	…
1	0	1	1	1	…
⋮	⋮	⋮	⋮	⋮	⋮

Experiment and Test

Data set preparation. This test uses the majority of the scores of the courses from the first semester of the first year to the first semester of the fourth year of 830 students from a major as a data set. Complexity and the need to achieve universality in the amount of data are also considered in the process of selecting data sets. Since Apriori algorithm can only process discretized data, the data set is also discretized and some other processes are prepared. Items in the data set include student numbers, names and course names. The name is indicated in Chinese Pinyin, and the course name is abbreviated by the initials of the English word. Student scores are divided into five grades, which are Excellent(>=90), Good(>=80), General(>=70), Pass(>=60) and Fail(<60).

After that, each student becomes a transaction, that is, the student number, name and the scores of each course form a transaction. There are 49 courses in this test and each course can be divided into 5 grades. The number of elements of a transaction is obtained by the following formula: 49*5=245. Apriori algorithm has practical effects in the case of high dimensionality. This data set dimension can be said to be a relatively high-dimensional data set. Therefore, for this design application, the data sets adopted are basically satisfactory or can be said to be data sets with relatively high requirements.

Here, a small part of the data set is intercepted, as shown in Table 3, where Num represents students' numbers and Name represents students' names. From the beginning of C column, the English abbreviation is used to indicate the names of courses.

Table 3 A small part of the data set is intercepted

No.	A	B	C	D	E	F	……
1	Num	Name	Ccb	Aml	La	Cel	……
2	3130101	Xia Dongdong	10000	00010	10000	00100	……
3	3130102	Xu LiLi	10000	00010	10000	01000	……
4	3130103	Huang Qinglan	01000	00010	10000	00100	……
5	3130104	Cheng Ying	01000	00100	00010	00010	……
6	3130105	Dong Yan	01000	00010	00100	00010	……
7	3130106	Li Lihua	10000	01000	01000	00010	……
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Test results. Take the score database of the students in Computer School as an example. About 15% of the students have excellent scores in *Discrete Mathematics* and *Data Structure*. About 70% of the students have excellent scores in both *Data Structure* and *Discrete Mathematics*. The relationship between the two is as follows:

Data Structure (excellent) → Discrete Mathematics (excellent), support=15%,

confidence=70%

The above correlation results tell us that in order to improve *Data Structure* which is a very important compulsory course for Computer School, the course of *Discrete Data* must be given priority to.

At the same time, the association rules summarized through analysis and excavation are not completely useful, and may even be misleading to people.For example, among the students of Computer Science of 2013, the ones who achieve excellent results in *College Sports 4* and *Legal Basis* account for about 40%. Among them, the ones who achieve excellent results in *Legal Basis* account for about 86%, and they also achieve excellent results in *College Sports 4*. At this point, if it is concluded that the results of *College Sports 4* are dependent on *Legal Basis*, this argument is obviously wrong. In this case, the most effective way to deal with this is to perform a simple correlation analysis to eliminate this relationship which seemingly includes rules, where correlation is measured by the following formula Eq. 1.

$$Corr_{A,B} = \frac{P(A \cup B)}{(P(A) * P(B))} = \frac{P(B \mid A)}{P(B)} \qquad (1)$$

It is easy to calculate the result from the above formula. When the correlation coefficient between event A and event B is less than 1, it is called negative correlation. Negative correlation rules should be removed when the final rule is formed. We should also follow objective reality. For example, there is a PP2<60→UNIX<60 in the conclusion, that is, *Situation and Policy* fails with score => UNIX. However, when we calculate CorrA, B, the probability of failing UNIX due to failure of *Situation and Policy* is 0.78, and the probability of failing UNIX is 0.2, at this time CorrA, B>1, the rule is positively correlated. However, from the perspective of association of events, this rule has no basis. It can be seen that objective events are still a criterion to be grasped.

Conclusion

Apriori algorithm for association rule data mining is a classic in the association rule algorithm, but it does have many performance bottlenecks that need to be solved by optimization algorithms. In this paper, Apriori algorithm is improved by using methods, such as hashing, division and sample selection. It is used for student performance analysis and has achieved good results. However, due to the limitations of experimental conditions, it is still not available on a large scale, so the follow-up work is still very complicated, so that it can be programmed and commercialized in the future, and can be used for the analysis of various data of colleges on a large scale. At the same time, we also hope to make further improvements in the efficiency of Apriori algorithm.

Reference

[1] Chengqi Zhang, Shichao Zhang. Association rule mining model and algorithms. Springer I.2002.2307:33-39.

[2] Tao Jianwen, Yao Qifu. Personalized Learning Recommendation System based on Web Usage Mining[J]. Journal of Computer Applications, 2007, 27, 7: 1809-1816. (In Chinese)

[3] Agrawal R., Shafer J.C. Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 962-969.

[4] Rakesh Agrawal, Ramakrishnan Srikant. Fast Algorithm for Mining Association Rules. *Proceedings of 20th Int. Conf. Very Large Data Bases*(*VLDB*), Jorge B. Bocca, Matthias Jarke and Carlo Zaniolo, eds. Morgan Kaufmann Press,1994: 487-499.

[5] Jiawei Han, Jian Pei, Yiwen Yin. Mining Frequent Patterns without Candidate Gerneration. *Proceedings of ACM SIGMOD Intl. Conference on Management of Data*, Weidong Chen, Jeffrey Naughton, Philip A. Bernstein eds. ACM Press, 2000: 1-12.

[6] A. Savasere, E. Omiecinski, S. Navathe. An Efficient Algorithm for Mining Association Rules in Large Databases[A]. *Proceedings of 21th Int'1 Conf. Very Large Data Bases*. San Francisco: Morgan Kaufmann, 1995, 432-444.

[7] Joog Soo Park, Ming-Syan Chen, Philip S.Yu. An Effective Hash Based Algorithm for Mining Association Rules[A]. Proceedings of ACM SIGMOD International Conference on Management of Data[C]. Michael J. Carey and Donovan A. Schneider eds. San Jose, Califormia,1995,175-186.
[8] Shao Fengjing, Yu Zhongqing. Principles and Algorithms of Data Mining. Beijing: China Water Resources and Hydropower Press, 2003: 123-135. (In Chinese)
[9] Jaw Han, Micheline Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.