Research and Application of Association Rule Mining Algorithm on Teaching and Learning of Distance Education Platform

Shaozhen Huang

Hainan Open University, Hainan, Haikou, 570100

Keywords: association rule; mining algorithm; distance education platform

Abstract: This paper introduces the basic principles of association rule mining algorithms, and in turn studies association rule mining algorithms based on the number of variables (dimensions) involved in mining, the level of abstraction of data, and the categories of processing variables (Boolean and numeric). This paper summarizes, analyzes and compares some typical algorithms. Finally, the research direction of association rule mining algorithms is prospected.

1. Introduction

Data mining refers to the extraction of implicit, previously unknown knowledge and rules that have potential value for decision-making from large databases or data warehouses. It is the product of the combination of artificial intelligence and database development. It is one of the most cutting-edge research directions for database and information decision systems in the world. The main algorithms of data mining include classification mode, association rules, decision tree, sequence pattern, cluster pattern analysis, neural network algorithm and so on. Association rules are a very important research topic in the field of data mining. They are widely used in various fields. They can test long-term knowledge patterns in the industry and discover hidden new laws. Effectively discovering, understanding and applying association rules is an important means to complete data mining tasks. Therefore, the research on association rules has important theoretical and practical significance.

2. Basic Principles of Association Rule Mining Algorithm

Association rule mining algorithmis a process of mining association rules from a given transaction database. The classic algorithm is the Apriori algorithm. The association algorithm of SQL Server 2005 uses the Apriori algorithm. The algorithm divides association rules into two steps. In the first step, all frequent items whose support degree is not less than the minimum support degree are mined from the database. In the second step, the association rule that the confidence degree is not less than the minimum confidence degree is generated using the frequent items that have been excavated.

Based on the information obtained from the previous scan of the data set, careful analysis of this combination can yield an improved algorithm. Mannila et al.considered this first, and they considered sampling as an effective way to discover rules. This idea was further developed by Toivonen, first using the samples extracted from the database to get some rules that might be established in the entire database, and then verifying the results for the rest of the database. Toivonen's algorithm is quite simple and significantly reduces the I/O cost, but a big drawback is that the resulting result is inaccurate, that is, there is a so-called data skew. Data distributed on the same page is often highly correlated and may not represent the distribution of patterns across the database, which may result in the cost of sampling 5% of the transaction data that may be similar to scanning the database. Lin and Dunham discussed an antiskew algorithm to mine association rules, where they introduced techniques to scan the database less than two times. The algorithm uses a sampling process to collect the number of relevant data to reduce the number of scan passes.

AprioriTid and AprioriHybrid algorithms proposed by R. Agraw al et al. In addition to the characteristics of the Apriori algorithm, the algorithm AprioriTid has another feature, that is, the

transaction database D is used to calculate the support number of the candidate strong set only in the first scan, and the other scans use the candidate database D generated by the previous scan. To calculate the support number of candidate strong item sets. In the last few scans, the size of D is much smaller than D, reducing the I/O operation time and improving the efficiency of the algorithm. Algorithm AprioriHybrid is the combination of algorithm Apriori and algorithm AprioriTid. Algorithm Apriori is used when candidate transaction database D can not be fully stored in memory, and algorithm AprioriTid is used when memory can fully accommodate candidate transaction database D.

Constraint-based rule mining the content of constraints can be: a. Data constraints. The user can specify which data to mine, not necessarily all of the data. b. Specify the dimension and level of mining. Users can specify which dimensions of the data and which levels on those dimensions are mined. c. Rule constraints. You can specify which types of rules are needed. Introduce the concept of a template, which the user uses to determine which rules are of interest: If a rule matches a contained template, it is interesting, but if a rule matches one Restricted templates (rextrictivetemplate) are considered to be of lack of interest. For the Apriori-based frequency set method, even if optimized, some inherent defects can not be overcome. Apriori's algorithms and their optimization algorithms may generate a large number of candidate sets. When there are 10,000 frequency sets of length 1, the candidate set of length 2 will grow exponentially, and the number of candidate sets will exceed 107. If you want to generate a very long rule, the intermediate element to be generated is also huge. This can be solved using the FP tree algorithm.

3. The Application of Association Rule Mining Algorithm in Distance Teaching Feedback

The collection of distance education teaching feedback information is the process of collecting information such as basic individual student information, learning behavior information, and teaching evaluation information. In the collection process, full use of information technology and network resources, a comprehensive collection of various types of teaching feedback information, according to the database input requirements, the various types of feedback information into the database. The student's personal basic information records the learner's basic characteristics and is obtained by submitting an electronic form when the user registers. Collect basic student information.

The learning behavior feedback information mainly includes the student's online learning time, homework completion status, participation in the online answering questions and test scores and other information. Collect this information in various subsystems of the e-learning platform. The online learning subsystem completes the collection of online learning time and other information in various chapters of the course; the operating subsystem completes the collection of information on the chapters' work performance, teacher comments, etc.; the Q&A subsystem completes the questions asked by the students in each chapter, the number of question answering, and other information. The acquisition; online test subsystem to complete the students' test scores and other information collection. The teaching evaluation information mainly includes information on the evaluation of teaching content and resources by students, the evaluation of teacher tutoring questions, and evaluation of the entire online learning platform. These evaluation information are mainly obtained through network survey and entered into the database.

Data extraction refers to the integration of various types of data from a transaction-oriented real-time operating database to a data warehouse oriented data mining analysis, which includes the integration of basic student information, learning behavior information, and teaching evaluation information. Data cleaning mainly includes null value processing and noise processing. For example, in the teaching evaluation database, it can be seen that some students do not evaluate the courseware resources, while others do not have evaluation results. Therefore, if there is no evaluation content or score for the students, the records of this part should be deleted. As for the content that only gives an evaluation, there is no evaluation of the teaching results. For these vacancies, the content of the evaluation can be filled. The Apriori algorithm needs to discretize the data of the classification attributes and continuous attributes. First, pre-admission education is

handled as follows: A1, junior college, A2, secondary school, A3, high school and below. Second, the proficiency of the computer is handled as follows: B1, not well understood; B2, basic understanding; B3, proficient. Third, the online time processing is as follows: C1, the online time is less than the learning time that the course must reach; C2, the online time meets the learning time that the course must reach. Fourth, the number of Q&A sessions is as follows: D1, less than 3 times; D2, 3-5 times; D3, 5 times or more. Fifth, all kinds of results are handled as follows: X1, 100-90; X2, 89-80; X3, 79-60; X4, 60 points or less.

Choosing the right data mining tool can make the mining more effective. The Business Intelligence Development Studio components of SQL Server 2005 are powerful, highly visual, and easy to use. They provide an integrated environment for creating and using data mining models that include data mining algorithms and tools that make it easier to implement various types of data mining. project. Therefore, this study was selected as a data mining tool. Using the association rule algorithm it provides, set the minimum supportability of the main parameters to 0.2, the minimum confidence level to 0.6, and the maximum item set to 2. The minimum support value range is 0 to 1. If this value is set too low, the algorithm may take a long time to process and require a lot of memory. The minimum confidence value ranges from 0 to 1. The largest item set specifies the value of the largest item set. Decreasing the maximum item set value will reduce the processing time because when the size of the candidate set reaches this limit, the algorithm does not need to iterate further on the data set.

4. Interpretation of Mining Results and Recommendations

From the rules, it can be seen that the student's online learning time is less than the study time that the course must reach, and the probability that the total rating is unsatisfactory is 80%. Limited online learning time is a major factor in the disqualification of remote students. This is because most of the online students are working adult learners. The contradictions in work and study and the contradictions in family education make students' study time limited. Teachers should consider this issue when arranging online courses.

As can be seen from the rules, the degree of computer proficiency is not well understood, and the probability of unsatisfactory performance is 80%. At present, a considerable number of distance education students do not have the knowledge and skills necessary for online learning, which has caused them difficulties in learning. Therefore, the distance education department can organize training in this area so that learners can master these knowledge and skills so that they can learn normally.

It can be seen from the rules that students have low ratings of courseware evaluation and teacher’s Q&A, and students’ overall evaluation performance is often not ideal. Multimedia courseware is the main resource for students' online learning. The distance education department should take various measures to improve the quality of courseware. Teachers should also pay full attention to online counseling. For example, teachers can use BBS, Email, chat rooms, etc. to communicate with students in order to answer the difficulties encountered by students in a timely manner.

As can be seen from the rules, students with better test scores have completed online assignments and have higher scores. Even if students who have less online learning time through the online job test, can clearly grasp the lack of knowledge, so as to carry out targeted review, or can achieve better results. This shows that online homework is a better method and indicator for measuring student knowledge. The important role of homework and peacetime tests in improving student achievement should be fully utilized.

It can be seen from the rules that students with low academic qualifications often have low evaluations of courseware. At present, most courseware in distance education is produced for students with higher education level. Because distance education learners have different academic levels before entering the school, it is difficult for the unified courseware to meet the needs of individualized learning. The production of courseware should try to consider these factors. For example, students with lower academic qualifications may provide additional preliminary

knowledge resources or provide specialized courseware for students with lower academic qualifications.

5. Conclusion

In this paper, the data mining technology is used to mine the collected teaching feedback information. According to the mining results, the teacher can adjust the teaching strategy in time and formulate the teaching content and teaching activities suitable for the individuality of the students. This paper not only broadens the application fields of data mining technology, but also provides deep data support that is different from previous simple statistics for the collection and processing of distance education teaching feedback information.

Acknowledgements

Fund Project: Hainan Higher Education Research Funding Project
Project No.: Hnky2018-62

References

[1] Chen Wenqing, Xu Yu. Improvement and implementation of Apriori algorithm for mining association rules [J]. Microcomputer Development (now renamed: Computer Technology and Development), 2005, 15(8): 155-157.

[2] Yang Jianbing. Improved algorithm of association rules in data mining and its implementation [J]. Microcomputer Information, 2006 (7): 195-197.

[3] Liu Lindong, Zeng Xiaoning .Application of Apriori Algorithm in Online Examination System [J] .Journal of Guangdong Institute of Education, 2005(5):103-108.

[4] Yang Yongbin. Application of Data Mining Technology in Education [J]. Computer Science, 2006(12): 284~286

[5] Luo Ke, Wu Jie. Research on measurement criteria of association rules [J]. Control and Decision, 2003, (5): 277~281