The Application of Association Rules and Interestingness in Course Selection System

Zhimin Chen, Wei Song, Lizhen Liu
Information and Engineering College
Capital Normal University
Beijing 100048, China
e-mail: hbu_chenzhimin@126.com, wsong@cnu.edu.cn, liz_liu@126.com

***Abstract*—With the development of social economy and the improvement of cultural and educational level, the enrollment of all the major colleges and universities in our country has been expanding constantly. The arrangement of courses of the school has become more and more complicated. It has become a common phenomenon that students of colleges and universities select courses by online course selection system. In this paper, we combine the association rule of the data mining technology with interestingness measure threshold, and apply them to course selection system. Firstly, we set the support threshold, the confidence threshold and the interestingness measure threshold. Then we mine the association rules in the course information database. Finally, according to the relevance of the courses and association rules, we filter out the better association rules in the resulting association rules. The proposed method can provide students with better course selection programs without teachers' arrangements, and it saves time and money for school.**

Keywords-association rules; interestingness measure; course selection system

I. INTRODUCTION

With the development of Internet technology, the network-based course selection system for students has been applied by many colleges and universities, course selection is an important part of the successful completion of their studies. The existing course selection system sometimes can not meet the needs of students, and it is not conductive to the schools' academic department for the analysis of the courses that will be opened. Based on this situation, developing a course selection system with simple operation to meet the needs of students and teachers now has become an important trend.

Data mining technology is a cross-discipline that integrating the theory and technology of the database, artificial intelligence, machine learning, statistics and other fields[1]. And data mining is an important part of the field of computer research and application, it gives us a new way to understand the data. Association rule[2] is an important model of data warehouse and data mining technology. The research of data mining technology in China is relatively late in the world, although the data mining technology has been applied to many fields of society, the application of data mining technology in the student selection system is not very widespread at present.

In recent years, the research of data mining technology has drawn wide attention from experts and engineers in the field of artificial intelligence and database around the world[3,4,5]. Mining association rules in database is a very important subject in data mining field. An example of association rules is that "some customers buy milk, and 90 percent of them buy bread at the same time", which has an intuitive meaning of the proportion of customers tend to buy other products when they buy a certain product[6]. In the same way, we apply the association rules in the course selection system to find the relevance among courses, so that we can provide students with reasonable suggestions.

Nowadays, many colleges and universities in China have their own online course selection system, these systems only provide the basic function of course selection, but there are few schools applying association rules to course selection system. We apply association rules to course selection system, and use interestingness measure threshold to analyze those association rules obtained by association rules. So we make it possible to analyze the relevance of the selected courses better, and to provide students with better course selection programs.

The rest of this paper is organized as follows. In Section II, we review the related work briefly. Then our proposed method is introduced in Section III. Section IV gives the experimental results and the corresponding analysis. Finally, Section V concludes the paper.

II. RELATED WORK

Data mining technology is an evolving technology that incorporates technologies and theories in many fields. Early on, there is a new term, Knowledge Discovery (KDD), since then people accepted KDD and used it to describe the whole process of data mining[7]. So far, the KDD[8] International Symposium sponsored by American Association for Artificial Intelligence has been held many times and the focus of the symposium has also been gradually applied by the Discovery Steering System, which focuses on the integration of multiple technologies and discovery strategies, as well as interdisciplinary penetration. Also, a lot of software products presented in the symposium have been applied in many countries.

Nowadays, many foreign academic papers have researched the data mining technology, and in some other fields, data mining technology has also been deeply researched and it is written into the special issue. Because the international markets' demands for data mining technology

978-1-5090-3619-6/17/$31.00 ©2017 IEEE

int software developing is increasing, so many well-known companies have researched data mining technology. Up to now, a lot of softwares with a wide applications of mature data mining technology have been developed.

Based on data mining, the system analyzes and gets rules. It discovers the neglected elements, predicts trends and makes decisions. Association rules can discover the relationships among data in different fields, and find out the dependencies among the fields to make support and confidence meet the given threshold. Mining association rules refers to mining the rules in data warehouse of the pattern, that is "some events' occurrence cause the occurrence of other events". For example, based on the information of selection and scores of students, we can draw the following conclusions: Comparing the scores of data structure course, students who have selected C++ program design generally have higher scores than those who have not selected C++ program design.

Through analysis, we find that there are association relationships and sequential relationships among the courses. Courses are sometimes sequential, the student learn a course may have an impact on learning another course, and this impact is hard to find, but the students themselves may not recognize it as well. By applying support[9] and confidence[10] of the association rules and interestingness measure threshold[11] to the course selection system, we can find out the relationships between the courses.

This paper applies the association rules of data mining technology to the course selection system. Based on the previous situation of students' selecting courses, through research and analysis, we can get the relationships among the courses and instruct the students to select courses using the course selection system, which should be welcomed by many teachers and students.

III. METHOD OF MINING ASSOCIATION RULES

In this section, we sketch our proposed method firstly, and then depict the algorithm in detail.

A. Overview

We can get the information of courses from the database of the educational administration information system, the information of the students choosing the courses and the information of the students' scores. Our goal is mining the association rules that among courses according to the previous course selection information, so that we can propose more reasonable course selection programs for the students.

Overall, our method is to setting the minimum support and minimum confidence between courses firstly. And then according to Apriori algorithm, we can mine association rules among courses. According to the association rules and interestingness measure threshold, we propose reasonable course selection programs to students.

B. Association Rules and Interestingness Measure

Association rules mining is one of the most important functions of data mining. It was proposed by Agrawal et al. in 1993[12,13,14] firstly. The most classical algorithm of asscociation rules is Apriori algorithm[15]. The key idea of Apriori algorithm is to use the known sets of high-frequency data items to extend to other high-frequency data sets. The support indicates the importance of association rules in the entire database, while the confidence level of association rules reflects its reliability.

Suppose $i = \{i_1, i_2, \quad , i_m\}$ is the collection of items, the task-related data D is the collection of database transactions, and each transaction T of D is a collection of items, which makes that $T \subseteq I$. Each transaction has an identifier that called TID. Suppose A is an item set, and transaction T contains A if and only if $A \subseteq T$. The form of association rules is like $A \Rightarrow B$, where $A \subset I$, $B \subset I$, and $A \cap B = \phi$ [16].

The support of association rules in D is the percentage that both X and Y are included in Transaction D, which named probability. The confidence is the percentage of Y, which is the conditional probability when D has already contained X in the transaction, which named conditional probability. The association rules are considered interesting if the minimum support threshold and the minimum confidence threshold are met, and these thresholds are artificially set according to the mining requirements [17].

The rule $X \Rightarrow Y$ in transaction set D is constrained by support and confidence, confidence indicates the strength of the rules, and support indicates the frequency of occurrence in the rules. The support of data set X is the ratio of the number of transactions X and the total number of transactions in D. To facilitate the following description, the support of data set X is represented by the number of X contained by database D. Support is defined as the ratio of the transactions containing $X \cup Y$ in D, representing the ratio that the number of transactions containing both X and Y and the total number of transactions for D; the definition of confidence for rule of $X \Rightarrow Y$ is: when the transactions contains X, the possibility of the transactions containing Y.

The minimum support represents the minimum statistic reliability of the data item set. The minimum confidence represents the minimum reliability of the rule. If the data item set X meets the requirement of $X.\sup port \geq \min confidence$, then X is considered as a large item set. Generally, the minimum confidence and the minimum support are given by the users. When both confidence and support are larger than their corresponding threshold, the rule is called the strong association rules, otherwise it is called weak association rules. The task of mining association rules is to find out those strong association rules[18] from the database whose support and confidence are larger than the given value.

Support:

$$\sup port(X \to Y) = \sup port(X \cup Y) = P(X \cup Y) \qquad (1)$$

Confidence:

$$confidence(X \to Y) = \frac{\sup port(X \cup Y)}{\sup port(X)} * 100\% = P(X \mid Y) \qquad (2)$$

We define the interestingness measure of $X \to Y$ as follows[11].

$$IM = \mathrm{c(x,y)} = \frac{\mathrm{P(XY)}}{\mathrm{P(X)P(Y)}} = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X \mid Y)}{P(X)} \quad (3)$$

$c(x, y)$ represents the correlation that between x and y. If $c(x, y) > 1$, x and y are positive correlation. If $c(x, y) < 1$, x and y are negative correlation. If $c(x, y) = 1$, x and y are independent of each other. In this paper, according to the setting of thresholds, we can retain the rules of interest [11].

We can understand $IM = \frac{P(Y \mid X)}{P(Y)}$ as follows. If $IM > 1$, the emergence of X can lead to the emergence of Y. And the larger of the IM's value, the greater of the driving force of X to Y, which is what we need. If $IM < 1$, the probability of occurrence of X reduces the probability of occurrence of Y, which is undesirable in analysis of association. If $IM = 1$, the probability of occurrence of X and the probability of occurrence of Y are independent of each other.

We combine confidence with interestingness, and we use the confidence threshold to find association rules, and then use interestingness measure threshold to find interesting association rules. These association rules satisfy the requirement of the relationship of the courses that can not be drawn in the original course selection system.

C. Algorithm

Apriori algorithm is one of the most classical algorithms of association rules, and the process can be divided into three steps:

(1) We pretreatment the experimental data, according the specific requirements, we give the corresponding operation to the database, then constitutes a standardized database D.

(2) Then, we find all the item sets that satisfy the minimum support in D, which is the largest item set. As the general situation, the database we are facing is relatively large, so that this step is the core of the algorithm.

(3) Finally, we generate the rules that meet the minimum confidence. Establishing a rule set, and explaining and outputing it.

The above is a traditional mining method of association rules, which only considers support and confidence. We combine this method with interestingness measure threshold, and apply them to the course selection system. We compute its confidence and interestingness of every possible association rule of a large set of items. Confidence is denoted as C, and interestingness is denoted as IM. They will appear four possible cases as follows:

(1) c <minconfidence, it indicates that the correctness of the rules is not high, we eliminate them.

(2) c> minconfidence, $IM = im = 1$, it indicates that the actual value of the rules is not high, we eliminate them.

(3) c> minconfidence, $IM \geq \mathrm{im} > 1$, it indicates that the rules have a high practical value, we can output them.

(4) c> minconfidence, $IM < 1$, it indicates that the rules of the negative rules may have a high practical value, but this paper does not discuss this situation, so that it is not rule that we need, we eliminate it.

In the above description, minconfidence represents the minimum confidence threshold, and im indicates the interestingness measure threshold.

The algorithm combining interestingness with association rule mining algorithm is described as follows. We input minsupport, minconfidence and im. Then the program outputs all interesting strong association rules. Firstly, a large item set is generated using the classic Apriori algorithm already described above. And then we use the large item set to generate the association rules with interest degree constraints.

IV. Experiments and Analyses

In this section, extensive experiments are conducted to demonstrate the effectiveness of our proposed method.

A. Data Processing

Each name of course corresponds to a code, the code of course is denoted by Cno, and the correspondence relationship is shown in Table I.

TABLE I. Code of Course

Cno	Name of Course
N1	C++ Program Design
N2	Advanced Mathematics
……	……
N12	Computer Networking
N13	Data Structure
N14	Introduction to Algorithms
……	……
N25	Information Security and Cryptology
N26	Network Attacks and Defense
……	……

We should process the previous scores of the students that stored in the database. Because the attribute of scores is quantity, we use the following methods to deal with student scores: if the score is more than 85 points, we will use "A" to represent it; if the score is more than 60 points and less than 85 points, we will use "B" to represent it; if the score is less than 60 points, we will use "C" to represent it; and the numbers after the name of courses represent the courses code. Each student is a transaction, which contains the score of each student's course. The processed data is shown in Table II.

B. Experimental Settings

In order to get the association rules, we combine the Apriori algorithm with the interestingness measure threshold set in this paper. After the association rules are obtained by the algorithm, we use the tinterestingness measure threshold to filter the association rules. Finally we can get a more reasonable recommendation program.

TABLE II. DATABASE OF SCORES

Cno / Sno	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	…	N20	…	N29	…
2161002001	C1	0	0	A4	0	B6	B7	0	0	0	0	0	B13	B14	…	0	…	0	…
2161002020	0	0	0	0	0	B6	B7	0	0	0	0	0	0	0	…	0	…	0	…
…	…																		
2161002101	0	0	0	A4	0	A6	A7	0	0	0	0	0	0	0	…	0	…	0	…
…	…																		

We set the minimum support to 0.21 and set the minimum confidence to 0.56, then try to find the frequent item sets which can meet the minimum support that are stored in the database. Using the association rules which were generated by frequent item sets to find all non-empty subsets of each frequent item set in D. If the ratio obtained by $support(D)/\sup port(d)$ is equal to or bigger than the minimum confidence, an association rule is generated, which is $d \to (D-d)$. The database of transaction structure is shown in Table III.

TABLE III. THE DATABASE OF TRANSACTION STRUCTURE

Sno	Score
2161002021	B1,A5,B6,C7,B11,A13,C15,A22,A26,B28,B2,A30,B33
2161002032	B1,A3,B6,B8,C11,A12,B15,C18,B22,C25,A2,B27,C31,A33
…	…
2161002078	B1,A4,A6,A8,C12,B13,C15,A18,B20,B29,C3,A31,B21
…	…

C. Experimental Results

We do not list all of the association rules, the partial association rules mined by above data is shown in Table IV.

TABLE IV. ASSOCIATION RULES

Rule	Support	Confidence
N1=>N13	0.63	0.95
N2=>N3	0.31	0.61
N13=>N14	0.24	0.74
N12=>N25	0.23	0.62
N25=>N26	0.27	0.68
N12=>N26	0.22	0.57

The second rule is shown in Table IV indicates that selecting advanced mathematics firstly is help to learing data structure for students. The value of support is 0.31, and the value of confidence is 0.61. In the database, the scores of discrete mathematics bigger than 85 points accounted for 45%, the scores of data structure bigger than 85 points accounted for 35%, while the scores of both of them bigger than 85 points accounted for 12%.

Therefore,

$$IM = \frac{0.12}{0.45 \times 0.35} = 0.762 < 1 \quad (4)$$

so the two courses are negatively related. According to the method in this paper, this rule can be deleted. Those rules like it can also be deleted.

We get the conclusion from the experimental results that the greater the interestingness measure threshold is, the fewer the number of rules is, and the more rules will be deleted.

According to the above tables of association rules and the meaning of each rule, we can draw the following conclusions: the students learn C++ program design firstly is helpful to the students learning data structure, the support is 0.63, the confidence is 0.95. And the students learn data structure firstly is helpful to the students learning introduction to algorithms, the support is 0.24, the confidence is 0.74. It is recommended that students learn C++ program design firstly, and then learn data structure, and then learn introduction to algorithms. The students learn computer networking firstly is helpful to the students learn introduction to network security and learn network attacks and defense, the support are 0.23 and 0.22 respectively and the confidence are 0.62 and 0.57 respectively. Therefore, students should be recommended to learn the computer networking firstly, and then learn introduction to network security and network attacks and defense. The students learn information security and cryptology firstly is helpful to the students learning network attacks and defense, the support is 0.27, the confidence is 0.68. It should be recommended that students learning information security and cryptology, and then learning network attacks and defense.

V. CONCLUSIONS

In this paper, we apply the association rules and interestingness measure to the course selection system for students to recommend prerequisite courses of some courses, it can help students master the professional knowledge more easily and complete their studies more successfully. With the increasing of data volume and the continuous development of data information, the application of association rules in data mining to the selection system in colleges and universities has gradually become a development trend, which can promote the further reform, improvement and development of education management.

ACKNOWLEDGMENT

This work was supported in part by National Science Foundation of China under Grants No. 61303105 and 61402304, the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046, the Beijing Natural Science Foundation under Grants No. 4154065, the Beijing Educational Committee Science and Technology Development Planned under Grants No.KM201 610028015, Science and technology innovation platform, Teaching teacher, and Connotation Development of Colleges and Universities.

REFERENCES

[1] Chen X, Li H. Application of Association-Rule in the Course Selecting System Based on Credit System[J].Modern Computer,2008,6,285.

[2] Agrawal R, Imielinski T, Swami A.Mining Association Rules between Sets of Items in Large Databases[C]//Proceedings of the 1993 ACM SIGM OD Conference. Washington D C:[s.n.],1993:207-216.

[3] Chen M S, Han J, Yu P S. Data mining: an overview from a database perspective[J]. IEEE Transactions on Knowledge and data Engineering, 1996, 8(6): 866-883.

[4] Agrawal R, Imielinski T, Swami A. Database mining: A performance perspective[J]. IEEE transactions on knowledge and data engineering, 1993, 5(6): 914-925.

[5] Piatetsky-Shapiro G. Advances in knowledge discovery and data mining[M]. Menlo Park: AAAI press, 1996.

[6] Tie zhixin, Chen qi. A Summary of Mining Association Rules[J]. Journal of Computer Applications, 2000, 17(1): 1-5.

[7] Zaki M J, Parthasarathy S, Ogihara M. New Algorithms for Fast Discovery of Association Rules[C]//KDD. 1997, 97: 283-286.

[8] Ester M, Kriegel H P, Sander J. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Kdd. 1996, 96(34): 226-231.

[9] Liu B, Hsu W, Ma Y. Mining association rules with multiple minimum supports[C]//Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999: 337-341.

[10] Dasseni E, Verykios V S, Elmagarmid A K. Hiding association rules by using confidence and support[C]//International Workshop on Information Hiding. Springer Berlin Heidelberg, 2001: 369-383.

[11] Qu Shouning, Xu Dejun, Wu Tong, Wang Qin. Interestingness Measure and Its Application in Learning Guidance System Based on Association Rules[J]. Journal of Computer Applications, 2007, 2: 246-248.

[12] Agraw R, Weika R.Fast algorithms for mining associ-ation rules in large databases[C]//Proc.20thInt'lConf.Very LargeDatabases, 1994:478-499.

[13] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases[C]//Acm sigmod record. ACM, 1993, 22(2): 207-216.

[14] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc. 20th int. conf. very large data bases, VLDB. 1994, 1215: 487-499.

[15] Chai S, Yang J, Cheng Y. The research of improved Apriori algorithm for mining association rules[C]//2007 International Conference on Service Systems and Service Management. IEEE, 2007: 1-4.

[16] Chen Dingquan, Zhu Weifeng. Association rules and library bibliography recommendation[J]. Intelligence Theory and Practice, 2009 (6): 81-84.

[17] Jiang Shengyi, Li Xia, Zheng Qi. Data mining principle and practice: Electronic Industry Press, 2011-8-1

[18] Lu Lina, Chen Yaping, Wei Hengyi. Resaarch on the algorithm apriori of mining association rules[J].Mini-Micro System,2000,9:940-943.