Investigating Interesting Rules Using Association Mining for Educational Data

Dr.Vijayalakshmi M N[1]
[1] Associate Professor
Department of MCA
R.V.College of Engineering, Bangalore, India

S.Anupama Kumar[2]
[2]Assistant Professor,
Department of MCA
R.V.College of Engineering, Bangalore, India

Kavyashree BN[3]
[3] Student, 5th semester,
Master of Computer Applications
R.V.College of Engineering, Bangalore, India

***Abstract*—The different data mining techniques like classification, clustering, association mining can be applied on different applications to bring out new knowledge from it. This paper aims at applying association mining on educational data to understand the knowledge and performance of students. Apriori algorithm is implemented on student log data to bring out the interesting rules. There rules can be used to infer the performance of the students and to impart the quality of education in the educational institutions. The algorithm generated frequent item sets using support so as to understand the interest of the students in the course. Interesting rules are generated based on frequent item sets using confidence factor of the dataset. The rule helps the tutor to understand the knowledge and performance of the students in answering the questionnaire and hence understand the interest of the students in the course.**

Index Terms — Student log data, Association rules, Apriori algorithm

I. INTRODUCTION

Advancement of technology has resulted in generating huge amount of electronic data, and has enabled the data to be captured, processed, analyzed, and stored inexpensively. This capability has enabled industries and innovations. Data mining is a rapidly growing field that is concerned with developing techniques to assist managers and decision makers to make intelligent use of these repositories. Educational Data Mining (EDM) is a research area that deals with the development of methods to explore data originating in an educational context. Education
al data mining can be used to deduce new information from the large educational repositories and bring out improvement in education. This paper explains the employment of the association rule mining on student log data and infers new rules from the dataset. These rules can be used to understand the efficiency of the students in understanding the course and their attitude in answering the questions on line. This paper explains the implementation of large student datasets which require finding inherent regularities (associations) in data. The output of the algorithm is used to understand the behaviour of the student towards understanding the course.

II. BACKGROUD WORK

In [1], Magdalene delighta Angeline has discussed the implementation of Apriori algorithm to extract the set of rules, specific to each class and analyzes the given data to classify the student based on their performance in academics. Students are classified based on their involvement in doing assignment, internal assessment tests, attendance etc., which helps to predict the performance of the student based on the pattern extracted from the educational database. This would help to identify the average, below average students and to improve their performance.

In [2], the different learning styles in which a student can be categorized have been discussed. Apriori and Tertius rule mining algorithms are used to understand the behavior and attitude of the students towards diverse learning styles. Both the algorithms revealed interesting rules depending on the confidence factor of the dataset. Item sets were generated using the Apriori algorithm giving best rules. Using Tertius algorithm, the most excellent rules were generated based on the number of hypothesis, confidence value, true and false positive rates.

In [3], the authors have proposed an analysis and prediction of students' placements based on the historical information from the database by considering the students information at different confident levels and support counts to generate the association rules. The widely used algorithm in data mining ie, Apriori algorithm is specifically considered for the extraction of the knowledge.

Apriori is a classic algorithm for learning association rules. The author in [6] have explained how Apriori algorithm is applied on database containing academic records of various students and how Apriori helps to extract association rules in order to profile students based on various parameters like exam scores, term work grades, attendance and practical exams. The implemented algorithms offer an effective way of profiling students which can be used in educational systems.

Authors in [11] carried out study in EDM using Apriori algorithm. Based on the Apriori algorithm analysis and research, it points out the main problems on the application Apriori algorithm in EDM and presents an improved support-matrix based Apriori algorithm. Which overcome the limitations of Apriori algorithm on EDM and produce enhanced result.

An improved version of Apriori algorithm is proposed [12] to overcome the deficiency of the basic Apriori algorithm. The basic Apriori algorithm follows bottom up approach which suffers from increased number of data base scan. The new proposed method follows top down approach which reduces the number of database scans. The improved version Apriori algorithm is more efficient which takes less time, less memory and hence reflects in high efficiency.

In [13] dataset containing crimes data concerning women is used and showcases the implementation of the Apriori algorithm in mining association rules from same dataset. WEKA tool is used for extracting results. Apriori algorithm has been implemented to extract hidden information present in the crime records. The output of the algorithm identified what age group is responsible for crime and find where the real culprit is hiding.

In this paper Apriori algorithm is implemented on student data set to understand the behaviour of the students in new learning environment and also helps the tutor to understand the efficiency of the student in learning the subject.

III. DATASET USED

E-learning systems accumulate a vast amount of information which is very valuable for analyzing students' behaviour. They can record the student activities involved, such as reading, writing, taking tests and even communicating with peers. [14]

Student dataset used in this work was collected using e-learning portal with student register number, name and many questionnaires answered by students along with timestamp. The questionnaire was circulated among the students online and the responses for the same were collected. The data set was then pre-processed before implementing the algorithm. The dataset is stored as a database using MYSQL. The Table I gives the description of the variables present in the dataset.

Table I: Description of table dataset

Attribute name	Data type	Size
Student name	varchar	20
Question1	varchar	100
Response1	Varchar	100
Question2	Varchar	100
Response2	Varchar	100
Question3	Varchar	100
Response3	Varchar	100
Question4	Varchar	100
Response4	Varchar	100
Question5	Varchar	100
Response5	Varchar	100
timestamp	Date	20

Table I: Description of table dataset

Data has been accumulated from various sources hence it is pre-processed in order to discard the inappropriate data and retain relevant data. The data pre-processing includes removal of irrelevant data, records containing null values or unknown values using In this research work, Depth first Search method has been used during pre processing and the records stored as candidate data set in a data base after pre processing.

IV. IMPLEMENTATION

Data mining have key objective to identify potential useful, novel and patterns in existing data using association rules. [7]

Apriori algorithm is used to extract the set of rules, specific to each class and analyzes the given data to classify the student based on their performance in academics which helps to predict the performance of the student based on the pattern extracted from the educational database. As described in [5] association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all frequent itemsets in a database.
2. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

The support of an item (or set of items) is the number of transactions in which an item occurs and count refers to the number of items present in the data set. [5]

$$support = \frac{(X \cup Y).count}{n} \quad (1)$$

The confidence of a rule is the ratio of support of the rule to the support of its antecedent

$$confidence = \frac{(X \cup Y).count}{X.count} \quad (2)$$

Where X is number of itemset, denoted by X.COUNT, in a data set T is the number of transactions in T that contain X [4]. The Pseudo code of the algorithm is given below:

```
Procedure Apriori (T, minSupport)
{
 //T is the database and minSupport is the minimum support
L1= {frequent items};
for (k= 2; Lk-1 !=∅; k++)
{
 Ck= candidates generated from Lk-1
//that is Cartesian product Lk-1 x Lk-1
```


for each transaction t in database
do{
#increment the count of all candidates in Ck that are contained in t
L_k = candidates in Ck with *minSupport*
}//end for each
}//end for
return U_k L_k;
}

To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself [10]

A. Implementation of Apriori

The following fig 1 describes the implementation of Apriori algorithm on dataset along with pre-processing and generation of strong rules.

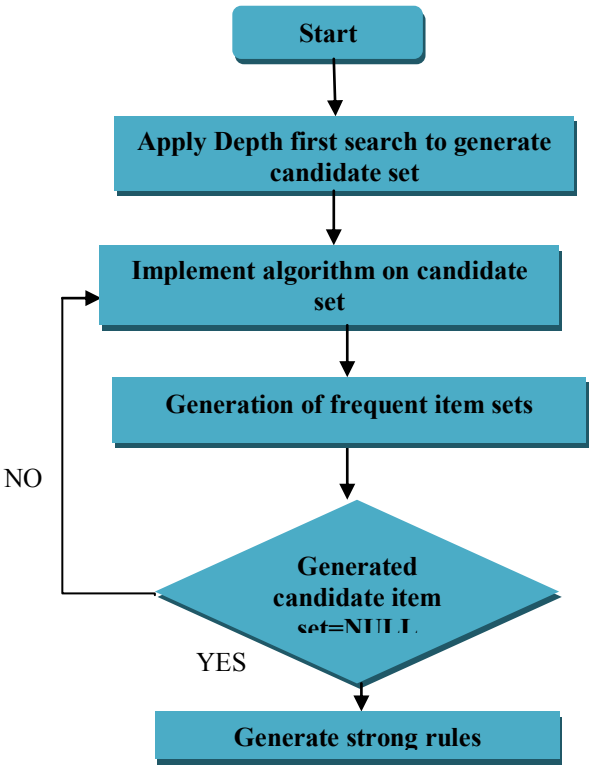

Fig 1: implementation of Apriori algorithm

The algorithm starts identifying the frequent individual items in the log database and extends the iteration to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database [3]. The algorithm uses minimum support uniformly to determine the frequent strong rules. The analysis of the frequent item sets and the rules clearly helps the tutor to understand the performance of the students in the particular course.

itemset in student dataset and changes the confidence level of each rule extracted from the frequent student itemset. The following items were generated by assigning the minimum support as 24 for the given dataset:

Frequent itemset:

1) [8/22/2012 20:20:30, brute force method, Fk-1XFk,Fk-1XFk-1]
2) [8/22/2012 20:20:30, data services, data migrator]
3) [8/22/2012 20:20:30, anti-monotone, data services, data migrator]
4) [brute force method, Fk-1XFk,Fk-1XFk-1, data services, data migrator]

The following inferences are made from the frequent item sets generated:

1) Many students have answered the questionnaire on 8/22/2012
2) Students have answered brute force have answered Fk-1XFk, Fk-1XFk-1 also as an answer to the question.
3) Students have answered data services on 8/22/2012 20:20:30 have answered data migrator as an answer to the questions
4) The interest of the students towards answering the questions in an online environment can be understood using the time at which they have answered.

These frequent item set where used to generate strong with the minimum confidence is set to be 0.5. Table II gives rules generated along with the confidence of the rules.

Table II: Generated rules and its confidence

Rules generated	Confidence
[anti-monotone, data services, data migrator]=>[brute force method, Fk-1XFk,Fk-1XFk-1]	0.96
[8/22/2012 20:20:30]=>[anti-monotone, data services, data migrator]	1.0
[8/22/2012 20:20:30, data services, data migrator]=>[anti-monotone]	1.0
[data services, data migrator]=>[anti-monotone]	0.512

The Table II clearly shows that frequent items along with time stamp have maximum confidence factor which forms

V. CONCLUSION

The paper explains the discovery of hidden knowledge, unexpected patterns and new rules from databases. The frequent item sets generated helps us to understand the knowledge of the student in the subject through the answers

and the interest of the students in answering the questions through online. The rule set helps us to understand the ability of the students in understanding the subject. Association mining can be further used in many other educational domains and help the educational community to grow.

REFERENCES

[1] D. Magdalene Delighta Angeline "Association rule generation for student performance analysis using Apriori algorithm"-, The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, pp 12-16

[2] S.Anupama Kumar , Dr. Vijayalakshmi MN, "Appraising the significance of self regulated learning in higher education using neural networks" , International journal of Engineering research and development ,Vol 1, Issue 1, May 2012, ISSN: 2278-067X, pp 09 – 15

[3] [Mr. Shreenath Acharya , Ms. Madhu N, "Discovery of students' academic patterns using data mining techniques", International Journal on Computer Science and Engineering (IJCSE), Vol. 4, June 2012 , ISSN : 0975-3397

[4] Lecture notes: "association analysis: basic concepts and algorithms", Website:"www. users .cs.umn.edu/~kumar/dmbook/ch6.pdf, pp 04

[5] http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab8-Apriori.pdf

[6] Parack. S, Zahid Z, Merchant F, "Application of data mining in educational databases for predicting academic trends and patterns", published in, technology enhanced education (ICTEE), 2012 IEEE international conference, Jan 2012 , pp 01-04

[7] zijian zheng, ron kohavi, llew mason, "real world performance of association rule algorithms", blue martini software, 2600 campus drive, san mateo, USA , 2001

[8] Almahdi mohammed ahmed, Norita md norwawi, Wan Hussain wan ishak, "Identifying of student and organization matching pattern using Apriori algorithm for practicum placement", 2009 International Conference on Electrical Engineering and Informatics

[9] Lecture notes: K. Ming Leung, "Data Mining: Motivations and Concepts",Website:www.cis.poly.edu/~mleung/FRE7851/f07/introDataMining1

[10] Dolf Zantinge and Pieter Adriaans (2003), "data mining", pearson education.

[11] Jayshree Jha, Leena Ragha, "Educational Data Mining using Improved Apriori Algorithm", International Journal of Information and Computation Technology, Volume 3, 2013, pp 411-418

[12] Suneetha K R, Krishnamoorti R, "Web Log Mining using Improved Version of Apriori Algorithm", International Journal of Computer Applications 29(6), September 2011, pp 23-27

[13] Divya Bansal, Lekha Bhambhu, "Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women ",International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 9, September 2013 ISSN: 2277 128X, pp 54-62

[14] Enrique García, Cristóbal Romero, Sebastián Ventura, Carlos de Castro, Toon Calders,"Association Rule Mining in Learning Management Systems", http://wwwis.win.tue.nl/~tcalders/pubs /CH7handbook.pdf

[15] Agathe Merceron, Kalina Yacef, "Revisiting interestingness of strong symmetric association rules in educational data", Proceedings of the International Workshop on Applying Data Mining in e-Learning 2007

