Pattern Mining Approach to Categorization of Students' Performance using Apriori Algorithm

Sushil Kumar Verma
Department of Computer Applications
SATI Vidisha

R.S.Thakur, PhD
Department of Computer Applications
MANIT Bhopal

Shailesh Jaloree
Department of Computer Science & Mathematics
SATI Vidisha

ABSTRACT

Researchers are used of data mining to extract hidden information from raw data. Now data mining can be used in any domain such as education. Data mining is used in education to achieve quality education and to categorize the students' performance through the analysis of educational data which reside or store in educational organization's database. In this paper, we categorize the performance of students based on their previous records such as 12th marks, graduation marks, previous semester marks (PSM) , previous academic records (PAR- average of 12th and graduation marks), mid sem marks (MSM), attendance (ATT) and end semester marks (ESM). Based on these attributes we determine the performance of students in end semester using apriori algorithm. With the help of categorization of performance, the main advantage is that classify of weak students, so that teacher give the particular interest on weak students and they could better perform in the next semester exam.

Keywords

Data mining, Educational data mining, Association rule mining, Data Transform.

1. INTRODUCTION

Traditional statistical techniques and data base management tools are application oriented and are not suitable for analysis of large amount of data, because in the world a huge quantity of data are collected and stored daily. Study of such data is an important need [1] [2].

This need can be fulfill by the use of data mining. Data mining has become the area of growing significance because it helps in analyzing data and summarizing it in to useful information. The notion of data mining is the technique of extracting previously unknown information with the widest relevance from database, in order to use it in the decision making process [1] [2]. Data mining is the process of extracting previously unknown, valid, potentially useful and hidden patterns from large datasets. In order to get required benefits from large data and to find hidden relationships between variables, different data mining functionalities / techniques can be developed and used such as classification, prediction, clustering and association analysis. Association is the discovery of association relationships or correlations among a set of items. An association rule in the form of X -> Y is interpreted as « database tuples that satisfy X are likely to satisfy Y ». Association analysis is widely used in transaction data analysis for marketing, catalog design and other business decision making process [3] [4]. A lot of private and government institutes and universities have been opened in India recently. The main aim of these educational institutions is to provide standard and quality education to students especially for competitions. Lack of knowledge and useful information on the part of the management about students' behavior prevents them from achieving the standard and qualitative objective which in turn can be achieved effectively by the use of data mining techniques. Data mining comprises a lot of tasks that can be used efficiently to analyze the performance of the students. Due to increasing competition the students are required to be assessed on various parameters for which data mining techniques can be very handy. Educational Data Mining is an emerging field which relates to developing new methods to analyze information from data.

Association analysis can be used in educational field. Data mining is a new notion in education domain. Data mining used in different fields including higher educational environments. Educational data mining which extracts useful, previously unknown patterns from educational database to improve and enhance the student's performance and improve educational quality.

This paper is organized as follows. Section 2 explains related work in educational data mining. Section 3 expalins about association rule mining. Section 4 explains about experimental setup and section 5 shows experimental results. Section 6 is conclusion and future work.

2. RELATED WORK

The aim of this section is to provide a review of the past research efforts related to catégorization of students' performance and educational data mining. This section shows various existing proposed works in the area of educational data mining.

In 2006 Shyamala et.al. [5] developed a model to find similar patterns from the data and to make prediction about performance of student using data mining.

In 2007 Ranjan et.al. [6] described a framework for effective educational process using data mining techniques to uncover the hidden trends and patterns and making accuracy based predictions through higher level of analytical sophistication in students analysis process using data mining.

In 2010 M.Ramaawami et.al. [7] explained the performance of student could depend on different factors like demographic, academic, psychological, socio-economis and other factors. Based on these factors they constructed their analytical process to identify the different performance factors of students using association rule.

In 2011 J.Mamcenko et.al. [8] tried to store and collect the data based on questions given to the student and the answers given by the students. Based on questions they construct their data sets and applied association mining techniques in these data sets to discover or identify different learning patterns of every student.

In 2006 T.Hadzilacos et.al. [9] explained that in a complex educational system, sensitive learning patterns can be identified, applying the machine learning techniques on a data set of student. In their research work, they focused on association rule mining algorithms on demographic data and student course data to discover the relationship of tutors with their students using association.

In 2012 Azhar Rauf et.al. [10] suggested a method known as k- means clustering algorithm for students' data, it calculates initial centroids instead of random selection, due to which the number of iterations is reduced and elapsed time is improved for find the students' cluster

In 2003 Jaideep Vaidya et.al. [11] suggested a privacy preserving k-means clustering method over vertically partitioned data when different web sites for educational purpose contain different attributes for a common set of entities. This knowledge can be helpful for e-learning systems.

In 2011 S.A.Kumar et.al. [12] explained to predict the overall performance of student based on their internal assessment and they considered different course modules which were offered in a semester using claffication.

3. ASSOCIATION RULE MINING

The objective of Association rule mining is to find out the interesting rules from which knowledge can be derived. Let $I = \{i_1, i_2, ..., i_m\}$ is a set of items, $T = \{t_1, t_2, ..., t_n\}$ is a set of transactions, each of which contains items of the itemset *I*. Thus, each transaction t_i is a set of items such that $t_i \subset I$. An association rule is an implication of the form: $X \rightarrow Y$, where $X \subset I, Y \subset I$ and $X \cap Y = \phi$. *X* (or *Y*) is a set of items, called itemset [3] [14] [17]. **Support:** The support of the rule $X \rightarrow Y$ is the percentage of transactions in *T* that contain *X* ∩*Y*. It determines how frequent the rule is applicable to the transaction set *T*. The support of a rule is represented by the formula [2] [3] [15].

$$supp(X \rightarrow Y) = \frac{|X \cap Y|}{n}$$

Where |*X*∩*Y*| is the number of transactions that contain all the items of the rule and n is the total number of transactions.

Confidence: The confidence of a rule shows the percentage of transactions containing *X* which also contain *Y* [2] [3] [15].

$$conf(X \rightarrow Y) = \frac{|X \cap Y|}{|X|}$$

3.1 Pattern Mining Algorithm

Several algorithms have been developed to improve the efficiency of frequent pattern and association rule discovery. Most of the algorithms such as apriori, FP-tree, partitioning and samplimg algorithm have been developed [2] [16]. Pattern mining algorithm focus on either frequent itemset generation or discovering the association rules from the frequent itemset. Apriori provides solutions for both problems. Many of the algorithms have been developed for use with binary association rules, but they will equally work with quantitative association rules [2] [13].

3.2 Apriori Algorithm

The Apriori algorithm was the first attempt to mine association rules from a large dataset. It has been presented in [15] for the first time. The algorithm can be used for both, finding frequent patterns and also deriving association rules from them [2] [15].

Flow chart of apriori algorithm is shown in figure 1 [2] [3] [15].

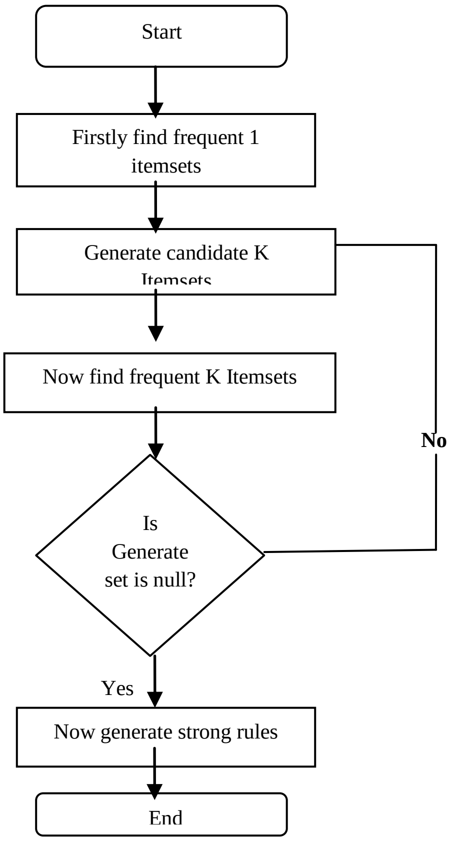
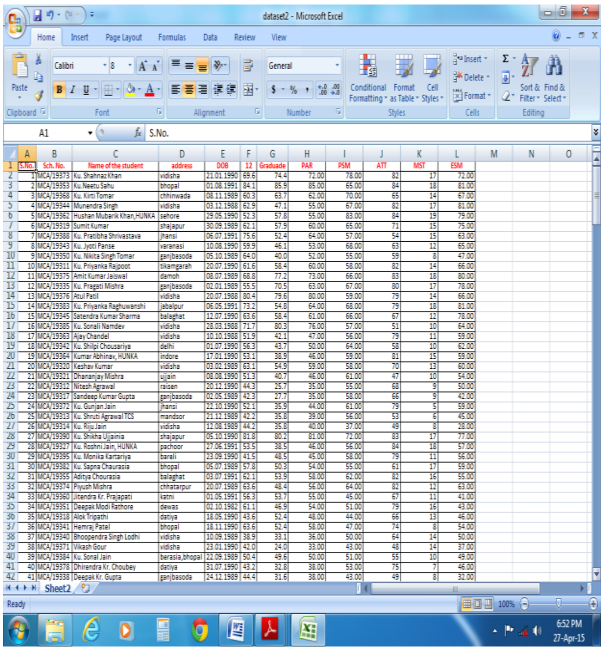

Start
Firstly find frequent 1 itemsets
Generate candidate K Itemsets
Now find frequent K Itemsets
No
Is Generate set is null?
Yes
Now generate strong rules
End

Fig. 1- Flow chart of Apriori Algorithm [3]

4. EXPERIMENTAL SETUP

We have considered student records dataset obtained from Samrat Ashok Technological Institute, Vidisha, (Madhya Pradesh) of course MCA (Master of Computer Applications) from session 2007 to 2010. This is dataset50. Size of the dataset50 is 50 student records which includes name, scholar no, DOB, 10th, 12th, graduation marks, PAR, PSM, MSM, ATT and ESM fields. In this experiment datasets are used to identify the performance of student. The details for students of MCA Batch (2007-2010) can be viewed in the fig. 2.



S.No.	Sch. No.	Name of the student	address	DOB	12	Graduade	PAR	PSM	ATT	MST	ESM
1	MCA/19373	Ku. Shahnaz Khan	vidisha	21.01.1990	69.6	74.4	72.00	78.00	82	17	72.00
2	MCA/19353	Ku.Neetu Sahu	bhopal	01.08.1991	84.1	85.9	85.00	65.00	84	18	81.00
3	MCA/19368	Ku. Kirti Tomar	chhinwada	08.11.1989	60.3	63.7	62.00	70.00	65	14	67.00
4	MCA/19344	Munendra Singh	vidisha	03.12.1988	62.9	47.1	55.00	67.00	82	17	81.00
5	MCA/19362	Hushan Mubarik Khan,HUNKA	sehore	29.05.1990	52.3	57.8	55.00	83.00	84	19	79.00
6	MCA/19319	Sumit Kumar	shajapur	30.09.1989	62.1	57.9	60.00	65.00	71	15	75.00
7	MCA/19388	Ku. Pratibha Shrivastava	jhansi	06.07.1991	75.6	52.4	64.00	57.00	54	15	63.00
8	MCA/19343	Ku. Jyoti Panse	varanasi	10.08.1990	59.9	46.1	53.00	68.00	63	12	65.00
9	MCA/19350	Ku. Nikita Singh Tomar	ganjbasoda	05.10.1989	64.0	40.0	52.00	55.00	59	8	47.00
10	MCA/19311	Ku. Priyanka Rajpoot	tikamgarah	20.07.1990	61.6	58.4	60.00	58.00	82	14	66.00
11	MCA/19375	Amit Kumar Jaiswal	damoh	08.07.1989	68.8	77.2	73.00	66.00	83	18	80.00
12	MCA/19335	Ku. Pragati Mishra	ganjbasoda	02.01.1989	55.5	70.5	63.00	67.00	80	17	78.00
13	MCA/19376	Atul Patil	vidisha	20.07.1988	80.4	79.6	80.00	59.00	79	14	66.00
14	MCA/19383	Ku. Priyanka Raghuwanshi	jabalpur	06.05.1991	73.2	54.8	64.00	68.00	79	18	81.00
15	MCA/19345	Satendra Kumar Sharma	balaghat	12.07.1990	63.6	58.4	61.00	66.00	67	12	78.00
16	MCA/19385	Ku. Sonali Namdev	vidisha	28.03.1988	71.7	80.3	76.00	57.00	51	10	64.00
17	MCA/19363	Ajay Chandel	vidisha	10.10.1988	51.9	42.1	47.00	56.00	79	11	59.00
18	MCA/19342	Ku. Shilpi Chousariya	delhi	01.07.1990	56.3	43.7	50.00	64.00	58	10	62.00
19	MCA/19364	Kumar Abhinav, HUNKA	indore	17.01.1990	53.1	38.9	46.00	59.00	81	15	59.00
20	MCA/19320	Keshav Kumar	vidisha	03.02.1989	63.1	54.9	59.00	58.00	70	13	60.00
21	MCA/19321	Dhananjay Mishra	ujjain	08.08.1990	51.3	40.7	46.00	61.00	47	10	54.00
22	MCA/19312	Nitesh Agrawal	raisen	20.12.1990	44.3	25.7	35.00	55.00	68	9	50.00
23	MCA/19317	Sandeep Kumar Gupta	ganjbasoda	02.05.1989	42.3	27.7	35.00	58.00	66	9	42.00
24	MCA/19372	Ku. Gunjan Jain	jhansi	22.10.1990	52.1	35.9	44.00	61.00	79	5	59.00
25	MCA/19313	Ku. Shruti Agrawal TCS	mandsor	21.12.1989	42.2	35.8	39.00	56.00	53	6	45.00
26	MCA/19314	Ku. Riju Jain	vidisha	12.08.1989	44.2	35.8	40.00	37.00	49	8	28.00
27	MCA/19390	Ku. Shikha Ujjainia	shajapur	05.10.1990	81.8	80.2	81.00	72.00	83	17	77.00
28	MCA/19327	Ku. Roshni Jain, HUNKA	pachoor	27.06.1991	53.5	38.5	46.00	56.00	84	18	57.00
29	MCA/19395	Ku. Monika Kartariya	bareli	23.09.1990	41.5	48.5	45.00	58.00	79	11	56.00
30	MCA/19382	Ku. Sapna Chaurasia	bhopal	05.07.1989	57.8	50.3	54.00	55.00	61	17	59.00
31	MCA/19355	Aditya Chourasia	balaghat	03.07.1991	62.1	53.9	58.00	62.00	82	16	55.00
32	MCA/19374	Piyush Mishra	chhatarpur	20.07.1989	63.6	48.4	56.00	64.00	82	12	63.00
33	MCA/19360	Jitendra Kr. Prajapati	katni	01.05.1991	56.3	53.7	55.00	45.00	67	11	41.00
34	MCA/19351	Deepak Modi Rathore	dewas	02.10.1982	61.1	46.9	54.00	51.00	79	16	43.00
35	MCA/19318	Alok Tripathi	datiya	18.05.1990	43.6	52.4	48.00	44.00	66	13	46.00
36	MCA/19341	Hemraj Patel	bhopal	18.11.1990	63.6	52.4	58.00	47.00	74	8	54.00
37	MCA/19340	Bhoopendra Singh Lodhi	vidisha	10.09.1989	38.9	33.1	36.00	50.00	64	14	50.00
38	MCA/19371	Vikash Gour	vidisha	23.01.1990	42.0	24.0	33.00	43.00	48	14	37.00
39	MCA/19384	Ku. Sonal Jain	berasia,bhopal	22.09.1989	50.4	49.6	50.00	51.00	55	10	49.00
40	MCA/19378	Dhirendra Kr. Choubey	datiya	31.07.1990	43.2	32.8	38.00	53.00	75	7	46.00
41	MCA/19338	Deepak Kr. Gupta	ganjbasoda	24.12.1989	44.4	31.6	38.00	43.00	49	8	32.00

Fig. 2- Dataset50 of Students Records from MCA (2007-2010)

The student data record that is being followed by the transformation is shown in table 1. The data is then ready to be mined using association rule mining. Here we show only 10 records in table 1.

5. EXPERIMENTAL RESULTS

The analysis using association analysis is being done with the help of WEKA tool. WEKA formally called Waikato Environment for Knowledge Analysis. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces. Result Generated by WEKA software is presented using fig. 3 and the table 2 shows the rules generated by WEKA.

Minimum support: 0.15 (8 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 17

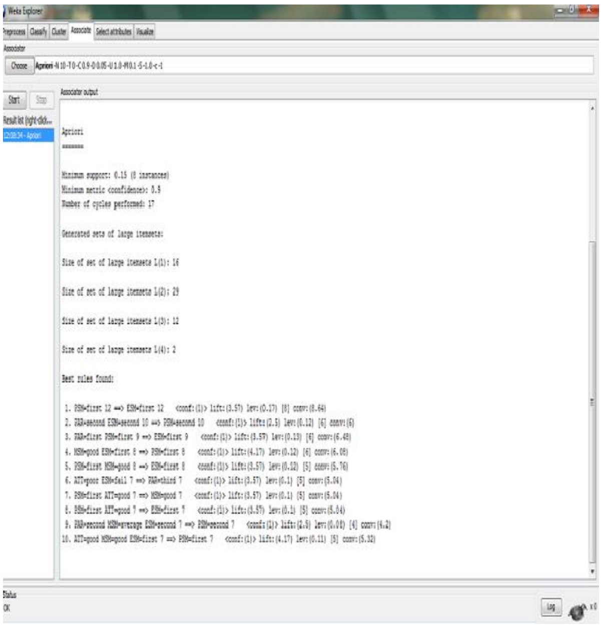
Generated sets of large itemsets:

Size of set of large itemsets L (1): 16

Size of set of large itemsets L (2): 29

Size of set of large itemsets L (3): 12

Size of set of large itemsets L (4): 2



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
Apriori
=======

Minimum support: 0.15 (8 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 29

Size of set of large itemsets L(3): 12

Size of set of large itemsets L(4): 2

Best rules found:

 1. PSM=first 12 ==> ESM=first 12    <conf:(1)> lift:(3.57) lev:(0.17) [8] conv:(8.64)
 2. PAR=second ESM=second 10 ==> PSM=second 10    <conf:(1)> lift:(2.5) lev:(0.12) [6] conv:(6)
 3. PAR=first PSM=first 9 ==> ESM=first 9    <conf:(1)> lift:(3.57) lev:(0.13) [6] conv:(6.48)
 4. MSM=good ESM=first 8 ==> PSM=first 8    <conf:(1)> lift:(4.17) lev:(0.12) [6] conv:(6.08)
 5. PSM=first MSM=good 8 ==> ESM=first 8    <conf:(1)> lift:(3.57) lev:(0.12) [5] conv:(5.76)
 6. ATT=poor ESM=fail 7 ==> PAR=third 7    <conf:(1)> lift:(3.57) lev:(0.1) [5] conv:(5.04)
 7. PSM=first ATT=good 7 ==> MSM=good 7    <conf:(1)> lift:(3.57) lev:(0.1) [5] conv:(5.04)
 8. PSM=first ATT=good 7 ==> ESM=first 7    <conf:(1)> lift:(3.57) lev:(0.1) [5] conv:(5.04)
 9. PAR=second MSM=average ESM=second 7 ==> PSM=second 7    <conf:(1)> lift:(2.5) lev:(0.08) [4] conv:(4.2)
10. ATT=good MSM=good ESM=first 7 ==> PSM=first 7    <conf:(1)> lift:(4.17) lev:(0.11) [5] conv:(5.32)
```

Fig.3- The Results Generated Using WEKA

The interpretation of the association rules (see table 2) for different confidence values depicts that the students' performance will be first in ESM if he / she got the first division in PSM or first in PSM and PAR or First in PSM and ATT is good in this semester or first in PSM and MSM was good. Also performance will be affected by the poor performance in unit test. So we can interpret that to get the good performance of student have to be good in their attendance and unit test. Also graduation performance will also have an impact on the student's unit test performance.

Table 1- Data Preprocessed (Transformed)

S.No.	PAR	PSM	ATT	MSM	ESM
1	'first'	'first'	'good'	'good'	'first'
2	'first'	'first'	'good'	'good'	'first'
3	'first'	'first'	'average'	'average'	'first'
4	'second'	'first'	'good'	'good'	'first'
5	'second'	'first'	'good'	'good'	'first'
6	'first'	'first'	'average'	'average'	'first'
7	'first'	'second'	'poor'	'average'	'second'
8	'second'	'first'	'average'	'average'	'first'
9	'second'	'second'	'poor'	'poor'	'third'
10	'first'	'second'	'good'	'average'	'first'

Table 2- Rule Generated by WEKA software in Dataset50

S.No.	Rules Generated by WEKA	Conf	Lift	Lev:	Conv:
1	If PSM=first then ESM=first	1	3.57	0.17	8.64
2	If PAR=second & ESM=second then PSM=second	1	2.5`	0.12	6

3	If PAR=first & PSM=first then ESM=first	1	3.57	0.13	6.48
4	If MSM=good & ESM=first then PSM=first	1	4.17	0.12	6.08
5	If PSM=first & MSM= good then ESM= first	1	3.57	0.12	5.76
6	If ATT= poor & ESM= fail then PAR= third	1	3.57	0.1	5.04
7	If PSM=first & ATT=good then MSM=good	1	3.57	0.1	5.04
8	If PSM=first & ATT=good then ESM=first	1	3.57	0.1	5.04
9	If PAR= second & MSM= average & ESM=second then PSM=second	1	2.5	0.08	4.2
10	If ATT=good & MSM=good & ESM=first then PSM=first	1	4.17	0.11	5.32

6. CONCLUSION AND FUTURE WORK

The icon of the institute is greatly involved by the results of the students. In this paper, a simple methodology based on association analysis algorithm is being used to determine set of weak students in current semester by comparing the performance of students of previous semesters on the basis of marks obtained at graduate level, marks obtained in previous semester, current semester attendance and current semester mid sem marks. This methodology will assist the academic planners to take the steps at the initial level in the subsequent years such that the overall result of the institute can improve. This will result in excellent placements and hereby increasing the quality intake of students. Hence this model will play important role for academic planners to determine the set of students that requires extra effort from institute side. Apriori algorithm used in this paper. In future we will apply apriori algorithm on fuzzy set of student records to improve the performance.

7. REFERENCES

[1] Brijesh Kumar Baradwaj, Sourabh Pal, "Mining Educational Data to Analyze Student’s Performance", IJACSA, Volume 2, Nov - 2011.

[2] Han, J. and Kamber, "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems", Jim Gray, Series Editor, 2006.

[3] Agrawal, Rakesh, Imielinski, Tomasz, Swami, Arun, “Mining Association Rules between Sets of Items in Large Databases”, Proceedings of the ACM SIGMOD International Conference on Management of Data, 1993.

[4] J. Han and M. Kamber, " Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.

[5] Shyamala K. and Rajagopalan S. P., “Data Mining Model for a better Higher Educational System", Information Technology Journal, Vol. 5, No. 3, pp. 560-564, 2006.

[6] Ranjan J. and Malik K., "Effective Educational Process: A Data Mining Approach", Vol. 37, Issue 4, pp 502-515, 2007.

[7] M.Ramaawami and R.Bhaskaran, "A CHAID based performance prediction model in educational data mining", 2010.

[8] J.Mamcenko, Jelena, Irna Sileikiene, Jurgita Lieponiene and Regina Kulvietiene, "Analysis of E-Exam data using data mining techniques", proceeding of 17th international conference on information and software technologies, pp. 215-219, 2011.

[9] T.Hadzilacos, D.Kalles, C.Pierrakeas and M.Xenos, "On small data sets revealing big differences", Advance in artificial intelligence, pp.512-515, 2006.

[10] Azhar Rauf, Sheeba, “Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity”, Middle-East Journal of Scientific Research, Vol. 12, Pp. 959-963, 2012.

[11] Jaideep Vaidya, "Privacy Preserving K-Means Clustering over Vertically Partitioned Data", proceeding of SIGKDD, Washington, DC, USA, August 24-27, 2003.

[12] S.A.Kumar and M.N.Vijayalakshmi, "Efficency of decision tree in predicting student’s academic performance", first international conference on computer science, Engineering and applications, Vol. 2, pp.335-343, 2011.

[13] Ceglar, Aaron, Roddick, John F., "Association mining: ACM Computing Surveys ", Volume 38 , Issue 2, 2006.

[14] Liu, Bing, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer, 2007.

[15] Agrawal, Rakesh, Srikant, Ramakrishnan, “Fast Algorithms for Mining Association Rules”, Proceedings 20th Int. Conf. on Very Large Data Bases, VLDB, 1994.

[16] Borgelt, Christian, "An Implementation of the FP-growth Algorithm”, ACM Press, New York, NY, USA, 2005.

[17] Ceglar, Aaron, Roddick, John F., "Association mining: ACM Computing Surveys ", Volume 38 , Issue 2, 2006.

[18]