Volume 4, Issue 2, February 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

www.ijarcsse.com

Research Paper

Available online at: www.ijarcsse.com

Behavior in Moodle System Using Datamining Technics

Yassine CHAJRI*, Abdelkrime MAARIR
Department of Informatics
FST-BENI MELLAL
University Sultan Moulay Sliman
BENI MELLAL, MOROCCO

Belaid BOUIKHALENE
Department of Mathematics
FP-BENI MELLAL
University Sultan Moulay Sliman
BENI MELLAL, MOROCCO

Abstract— Educational data mining is an interesting discipline, concerned with developing methods to extract knowledge and discover patterns from the E-learning systems. This work is an application of data mining in learning management systems. Our objective is to introduce Educational Data Mining, by describing a step-by-step process using a variety of technics such as Attribute Weighing, Classification, Clustering and Association Rules to achieve the goal to discover useful knowledge from Moodle. for association rules we will present a comparison between two algorithms datat mining Apriori and FP-Growth to justify our choice of FP-Growth algorithm.Analyzing mining results enables teachers to better allocate resource and understand student's behavior.

Keywords— Educational Data mining, E-learning, Data Mining, Moodle, learning patterns

I. INTRODUCTION

Data mining, or knowledge discovery, is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing being proactive and making decisions based on knowledge[1]. Data mining tools scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations…

So data mining refers to extracting or "mining" knowledge from large amounts of data using advanced technics like Classification, Clustering and Statistics….

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in. Using Data mining techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students‟ performance and so on. The main objective of this paper is to use data mining methodologies to study students performance. Data mining provides many tasks that could be used to study the student performance.

This paper is arranged in the following way: Section II is devoted to the definition of a some key words that are the focus of our work, section III describes the data preparation and preprocessing steps. Section IV shows the experimental results for dataMining algorithms.

II. EDUCATIONAL DATA MINING (EDM)

A. E-Learning

e-learning is electronic learning, and typically this means using a computer to deliver part, or all of a course whether it's in a school, part of your mandatory business training or a full distance learning course[2].

In the early days it received a bad press, as many people thought bringing computers into the classroom would remove that human element that some learners need, but as time has progressed technology has developed, and now we embrace smartphones and tablets in the classroom and office, as well as using a wealth of interactive designs that makes distance learning not only engaging for the users, but valuable as a lesson delivery medium. E-learning Systems offer the facilitation of communication between students and educators, sharing resources, producing content material, preparing assignments, conducting online tests, enabling synchronous learning with forums, chats, news services, etc…

A lot of e-learning systems exist like Moodle, TopClass, Ilias, Claroline… in our case we will use moodle.

B. Moodle

Moodle "Modular Object-Oriented Dynamic Learning Environment" is an open source course management system, orginally developed by Martin Dougiamas [3].It is used by thousands of educational institutions around the world to provide an organized interface for e-learning, or learning over the Internet. Moodle allows educators to create online courses, which students can access as a virtual classroom. A typically Moodle home page will include a list of participants (including the teacher and students) and a calendar with a course schedule and list of assignments. Other Moodle features include online quizzes, forums, where students can post comments and ask questions, glossaries of terms, and links to other Web resources.

C. Data mining Definition and techniques

Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or “mining" knowledge from large amounts of data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making [4]. While data mining and knowledge discovery in database are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The sequences of steps identified in extracting knowledge from data are shown in Fig 1.

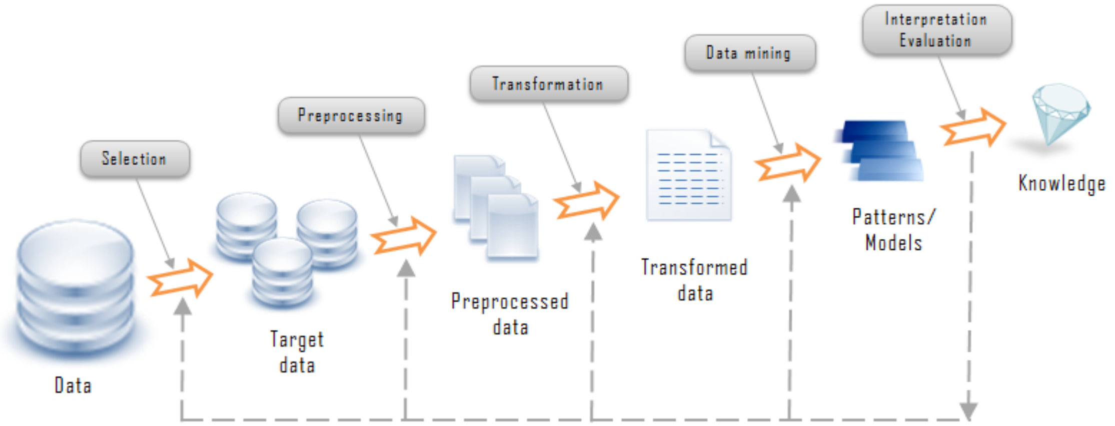

Fig. 1 The steps of extracting knowledge from data

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. These techniques and methods in data mining need brief mention to have better understanding.

D. Educational Data Mining(EDM)

EDM was defined by Baker [5] as the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in. In another words EDM concerns with developing methods that discover knowledge from data come from educational environment. The data can be collected from historical and operational data reside in the database of educational institutes. And the data reside on the e-learning systems database. The E-learning data mining process consists of the same four steps in the general data mining process:

Collect data: students activities are stored in the database of E-learning system in our case is moodle database.

Pre-process the data: The data is cleaned and transformed into an appropriate format to be mined.

Apply data mining: The data mining algorithms are applied to create and execute the model that discovers the knowledge and patterns. Different Data Mining tools exist to apply DM algorithms. In this paper we will use RapidMiner.

Interpret, evaluate and deploy the results: The results obtained are interpreted and used by the instructor for further actions. The instructor can use the information discovered to make decisions about students.

III. PRE-PROCESSING MOODLE DATA

Moodle is an open-source course management learning system to help educators create effective online learning communities. It is an alternative to proprietary commercial online learning solutions, and is distributed free under open source licensing. The Moodle database has about 306 interrelated tables. But we do not need all this information and it is also necessary to convert it into the required format used by data mining algorithms.

So in this table we will present the most important tables:

Table I: IMPORTANT TABLES

Table	**Description**
Mdl_assign	Assignment information
Mdl_assign_submission	Work done by student
Mdl_course	Course Information
Mdl_quiz	Quiz information
Mdl_quiz_attempts	Quiz attempts information

Mdl_quiz_grade	Quiz grade details
Mdl_forum	Forum information
Mdl_forum_post	Posts in forums
Mdl_forum_discussion	Discussions in forums
Mdl_forum_read	Posts reads by student
Mdl_grade_grades	Student grades
Mdl_log	Logs every user's action

A. Select Data

It is necessary to choose which courses can be useful for us. So, we will chose only the courses that they use a higher number of Moodle activities and resources at least assignments, forums and quizzes. In our database we have chosen 30 students and 4 courses.

This query selects the courses that had at least assignments, forums and quizzes.

```
SELECT c.id
FROM mdl_course c, mdl_assignment a, mdl_forum f, mdl_quiz q
WHERE c.id=a.course
AND c.id=f.course
AND c.id=q.course
GROUP BY c.id
HAVING count (a.course)>0 and count (f.course)>0 and count (q.course)>0;
```

B. Create Summarization table

Our data are spread over several tables, a summary table has been created (see TABLE II) which integrates the most important information for our objective. This table (mdl_summary) has a summary per row about all the activities done by each student during the course and the final mark obtained by the student in the course.

TABLE II: MDL_SUMMARY TABLE

Attribute name	**Description**
Course	Identification of the course
Assignment_number	Number of assignments done
Quiz_number	Number of quizzes done
quiz_passed_number	Number of quizzes passed
quiz_failed_number	Number of quizzes failed
forum_posts_number	Number of posts in forum
forum_read_number	Number of reads in forum
total_time_assignment	Time spent on assignments
Total_time_quiz	Time spent on quizzes
Total_time_forum	Total time spent on forums student
Resource_view	Total Number of course materials and resources views
Final_mark	Student final grade

To create mdl_summary table, we create a stored procedure with a lot of queries. So we will present 2 examples of sql queries applied on moodle database.

1) Sql query for calculate the total time spent on quizzes

```
SELECT sum (att.timefinish - att.timestart) AS time_spent
FROM mdl_quiz q, mdl_quiz_attempts att, mdl_user u
WHERE q.course = 2
AND q.id = att.quiz
AND att.userid =u.id.
```

2) Sql query for calculate Number of posts in forum

```
SELECT count (fp.userid) AS n_posts
FROM mdl_forum f, mdl_forum_discussions fd, mdl_forum_posts fp, user u WHERE fp.discussion = fd.id
AND fd.forum = f.id  AND fp.userid=u.id  AND fd.course=2;
```

C. Data Discretization

Performing a discretization of numerical values may be necessary to increase interpretation and comprehensibility. Discretization divides the numerical data into categorical classes that are easier to understand for the instructor (categorical values are more user-friendly for the instructor). All numerical values of the summarization table mdl_summary have been discretized except the course identification number. For the final grade attribute discretized with three intervals and labels (FAIL if value is <5; PASS if value is >=5 and <8; and EXCELLENT if value is >=8) .For the other attribute we used equal-width method (divides the range of the attribute into a fixed number of intervals of equal length) in all the other attributes with four intervals and labels (ZERO, LOW, MEDIUM and HIGH).

TABLE IIII: Summary Table (CATEGORICAL VERSION)

Attribute	**categories**
Assignment_number	Zero,Low,Medium,High
quiz_passed_number	Zero,Low,Medium,High
quiz_failed_number	Zero,Low,Medium,High
forum_posts_number	Zero,Low,Medium,High
total_time_assignment	Zero,Low,Medium,High
Total_time_quiz	Zero,Low,Medium,High
Resource_view	Zero,Low,Medium,High
Final_mark	FAIL,PASS,EXCELENT

D. Transform the data

The data must be transformed to the required format of the data mining algorithm. But in our case we use RapidMiner that accepts a several type of inputs like database table, Csv, Excel, ARFF…

In our case we used database table input.

IV. **APPLYING DATA MINING ALGORITHMS AND INTERPRETING RESULTS**

A. Information Gain

Before starting applying data mining algorithms it is worth to find out the weights for each attribute.

Information gain (IG) method used for attributes weighting, IG gives a good indication for how much the students are involved in a particular activity and this method is used in classification algorithm.

TABLE IIIV: Weighting with Information Gain IG

Attribute	**Information gain**
Total_time_quiz	0
forum_read_number	0
quiz_failed_number	0.010
forum_posts_number	0.015
total_time_assignment	0.370
quiz_passed_number	0.426
Quiz_number	0.484
Assignment_number	0.657
Resource_view	1

In the table IV the resource view attribute has the highest weight (with IG=1.00), this is a good indication that students are viewing the courses materials frequently. In the second place we found Assignment_number and the third place is taken by the number of quiz taken by student. In the last places we see total time spent on quiz and number of forum reads with IG = 0.

B. Classification

Classification consists of predicting a certain outcome based on a given input[6]. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. RapidMiner has several classification algorithms available. The C4.5 algorithm is used to characterize students who passed or failed the course. The C4.5 is an algorithm for generating decision trees and inducing classification rules from the tree[7]. In this case, our objective is to classify students into different groups with equal final marks depending on the activities carried out in Moodle. We have executed the C4.5 on mdl_summary with categorical version. We obtain a set of IF-THEN-ELSE rules from the decision tree that can show interesting information about the classification of the students.

In our decision tree(see Fig 2), if the resource view are zero or low the student classify as FAIL, if the Resource_view is high and quiz_passed_number is medium the student classify as excellent. If Resource_view is medium and assign number is medium the student classify as PASS.

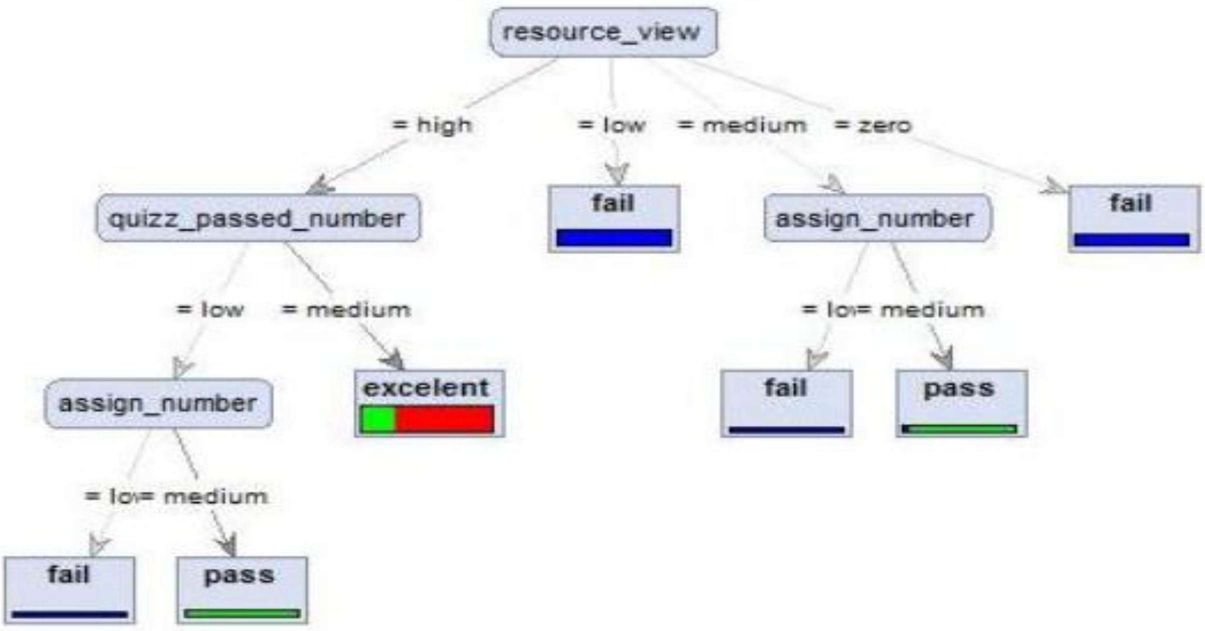

Fig. 2 Decision Tree

C. Model Validation

It's necessary to see the validation of the model by applying technics of measurement of validation and performance.
By X-validation we got a good results like the total accuracy have 85.67% and for the other results are regrouped in the table below.

TABLE V
X-Validation results

label	**Class recall**	**Class precision**
fail	97.83%	100%
Pass	41.03%	91.43%
excellent	100%	73.10%

D. Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes[8]. In e-learning, clustering has been used for: finding clusters of students with similar learning characteristics and to promote group-based collaborative learning as well as to provide incremental learner diagnosis[9].
The RapidMiner has several clustering algorithms. The KMeans (MacQueen, 1967) [10], one of the simplest and most popular clustering algorithms, has been used here and it is an algorithm that clusters objects based on attributes in k-partitions. In result we have two clusters a non-active group (cluster_0) and an active group (cluster_1).

TABLE VI
clustering with k-mean algorithm

Attribute	**Ctuster_0**	**Cluster_1**
Assignment_number	2.067	4.244
total_time_assignment	185.644	15.011
Quiz_number	1.712	3.650
quiz_passed_number	1.951	3.783
quiz_failed_number	0.344	0.128

Total_time_quiz	506.436	330.689
forum_posts_number	0.282	0.406
forum_read_number	0	0
Resource_view	50.47	500

Cluster_0 is characterized by inactive students in Moodle with a low number of actions in the system.
Cluster_1 is characterized by active student in moodle with high assignment number and resource view.

E. Association Rules

Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [11]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared later. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two subproblems. One is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second problem is to generate association rules from those large itemsets with the constraints of minimal confidence. Suppose one of the large itemsets is Lk, Lk = {I1, I2, … , Ik}, association rules with this itemsets are generated in the following way: the first rule is {I1, I2, … , Ik-1}⇒ {Ik}, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second subproblem is quite straight forward, most of the researches focus on the first subproblem.

The first sub-problem can be further divided into two sub-problems: candidate large itemsets generation process and frequent itemsets generation process. We call those itemsets whose support exceed the support threshold as large or frequent item-sets, those itemsets that are expected or have the hope to be large or frequent are called candidate itemsets [12].

1) Apriori:

The Apriori algorithm is one of the most influential algorithms used for mining association rules, which was proposed by R. Aglawal et al. in 1994. According to the principles of the Apriori algorithm in [13], it is composed of two steps, one is extracting all the frequent itemsets; the other is generating all the strong association rules from frequent itemsets [14]. In fact, the essence is to iteratively generate the set of candidate itemsets of length (k+1) from frequent itemsets of length-k and check their corresponding occurrence frequencies in the database to obtain frequent itemsets of length (k+1) at each level. Therefore it can be seen that there are two main reasons to low efficiency of the Apriori algorithm: It is required to generate lots of candidate itemsets for generating each frequent itemsets; It is essential to scan database many times for generating each frequent itemsets [15].

2) FP-Growth:

One of the currently fastest and most popular algorithms for frequent item set mining is the FP-growth algorithm [16]. It is based on a prefix tree representation of the given database of transactions (called an FP-tree), which can save considerable amounts of memory for storing the transactions. The basic idea of the FP-growth algorithm can be described as a recursive elimination scheme: in a preprocessing step delete all items from the transactions that are not frequent individually, i.e., do not appear in a user-specified minimum number of transactions. Then select all transactions that contain the least frequent item (least frequent among those that are frequent) and delete this item from them. Recurse to process the obtained reduced (also known as projected) database, remembering that the item sets found in the recursion share the deleted item as a prefix. On return, remove the processed item also from the database of all transactions and start over, i.e., process the second frequent item etc. In these processing steps the prefix tree, which is enhanced by links between the branches, is exploited to quickly find the transactions containing a given item and also to remove this item from the transactions after it has been processed [17].

3) The comparative study of apriori and FP-growth algorithm:

Apriori and FP-Growth are two data mining algorithms. Indeed, it is possible to generate, from these two methods, the rules for associating a database. The main common feature of these two algorithms is that they use the generation of frequent itemsets to find association rules. However, Apriori requires multiple scan database, generates a large number of itemsets and calculates the carrier each time. In addition, it is very expensive to manage this amount of itemsets, knowing that they must test the frequency of each itemset. The objective of the FP-Growth algorithm is to reduce the number of routes from the database to perform, significantly reduce the number of itemsets generation and facilitate the calculation of support. To do this, FP-Growth takes a cutting strategy to decompose the data mining tasks. That is why he uses the Growth Pattern Fragment method to avoid the costly process of generating and testing of candidates, used by Apriori.

based on this comparison we chose to use FPGrowth algorithm, because the results of this algorithm are a set of frequent item sets which would be used as input for Create Association Rule operator, this operator has created several Association Rules. The most important ones are presented in Table VII:

TABLE VII
Most important Association rules

Condition	**Result**
assign_number = low	**final_grade = fail, quizz_passed_number = low**
final_grade = fail	**assign_number = low, quizz_passed_number = low**
quizz_passed_number = low	**assign_number = low, final_grade = fail**
assign_number = low, final_grade = fail	**quizz_passed_number = low**
assign_number = low, quizz_passed_number = low	**final_grade = fail**
final_grade = fail, quizz_passed_number = low	**assign_number = low**
quizz_number = medium, resource_view = high	**quizz_failed_number = zero, final_grade = excellent**

V. CONCLUSION

In this research, a data mining model for Moodle data was proposed based on several technics: Attribute Weighting (Weighting by Information Gain,), Clustering (KMeans), Classification (Tree Decision), Association Mining (FPGrowth, Create Association Rule) was proposed. By educational data mining educator can apply clustering technics in order to obtain the exact groups of students. And these groups can also be used to create a classifier in order to classify students. The classifier shows what the main characteristics of the students in each group are, and it allows new online students to be classified. Finally, the instructors can apply association rule mining to discover if there is any relationship between these characteristics and other attributes. These rules can not only help to classify students, but also to detect the sources of any incongruous values obtained by the students.

ACKNOWLEDGMENT

After this work, we aim to apply the algorithms presented in this paper on the spatial data on the web and develop others in making comparisons.

REFERENCES

[1] Djamel Abdelkader ZIGHED, Ricco RAKOTOMALALA: Extraction des Connaissances à partir des Données.

[2] Alejandro Pena-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works.

[3] Martin Dougiamas, How we built a community around open-source software

[4] Brijesh Kumar Baradwaj, Saurabh Pal ; Mining Educational Data to Analyze Students Performance in International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011 .

[5] Baker, M.,(2010). Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), vol. 7, pp. 112-118. Oxford, UK: Elsevie.

[6] Fabricio Voznika, Leonar Doviana, Data mining Classification

[7] Nicolas Baskiotis , Michèle Sebag, C4.5 Competence Map: a Phase Transition-inspired Approach

[8] Jiawei Han, Micheline Kamber and Anthoney K .H.Tung, Spatial Clustering Methods In Data Mining : A Survey in School Computing Science, Simon Fraser University

[9] Cristóbal Romero , Sebastián Ventura, Enrique García : Data mining in course management systems: Moodle case study and tutorial.

[10] Vance Faber : Clustering and the Continuous k-Means Algorithm.

[11] Sotiris Kotsiantis, Dimitris Kanellopoulos,Association Rules Mining: A Recent Overview ; GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[12] Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216.

[13] R. Agrawal, T. Imelinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Database," Proc. ACM-SIGM OD International Conference, pp. 208-216, 1993.

[14] A. Salleb and C. Vrain, "An application of association rules discovery to geographic information systems," Proc. The 4th European Conference on Principles of Data Mining and Knowledge Discovery PKDD, pp. 613-618, 2000.

[15] Y Jaya Babu, G J Phani Bala, Siva Rama Krishna T Extraction Spatial Association Rules From the Maximum Frequent Itemsets based on Boolean Matrix : International Journal of engineering Science & Advanced Technology Volume - 2, Issue - 1, 79 – 84.

[16] J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on the Management of Data (SIGMOD’00, Dallas, TX). ACM Press, New York, NY, USA 2000.

[17] Christian Borgelt, An Implementation of the Fpgrowth Algorithm in Department of Knowledge Processing and Language Engineering School of Computer Science, OttovonGuerickeUniversity of Magdeburg Universit atsplatz 2, 39106 Magdeburg, Germany