MINING STUDENTS DATA TO ANALYZE ACADEMIC PATTERNS FROM EDUCATIONAL DATASETS

Nazek Awadh Ba saleh
School of Computing, College of Arts and Sciences, UUM, 06010 Sintok, Kedah, Malaysia.
nab9090@gmail.com

ABSTRACT: *In educational institutions specially universities there is a lot of data being produced and gathered in databases that lead to the production of educational data. Mining educational data concerns with discovering knowledge and pattern from educational data using data mining methods and techniques. The main challenge in this area is to deeply analyze and transform such data into a valuable information and knowledge to get benefit from. Most related work concentrates on predictive analysis. This study aims to discover new valuable knowledge and patterns from postgraduate students' educational data of School of Computing (SOC) in University Utara Malaysia (UUM) using association rule as data mining techniques and RapidMiner modular environment as a tool. The results discern and discover patterns and provide valuable knowledge such as the relation between students' background information with students' performance.*

Keywords: Mining educational data, UUM, Association rule, FP-Growth, Apriori, student data

.

INTRODUCTION

The advent of technology has become an integral part of integrated world led to the availability of a huge amount of data which can be used effectively to produce vital information. Data analytics allow uncovering hidden patterns and other useful information in data and it can be used across various sectors such as transport, health-care, industry or education. Education is utilizing available technologies more and more each year and analyzing educational data are expected to be the biggest factor in what is going to shape the future of education [1]. Data analytics is similar with data mining in the context of the work with patterns to determine if the pattern in data can explain what is happening.

Data mining is referred to the discovery of new and potentially valuable information from an extensive level of data and it has been employed in various fields. Research interests in using data mining in the area of education is known as Educational Data Mining. Such education is regarding the development of methods to unearth knowledge and data patterns that stem from the educational surroundings which basically concerning micro-concepts entailed in learning. Educational data mining provides many advantages over traditional research paradigms like laboratory experiments, in-vivo experiments and design research.

The primary issue in mining educational data is the requirement of enhanced technological solutions to address them. [2] highlighted that the primary challenge in higher learning institutions is the deep analysis of information and the identification of valuable information to develop a strategy for further and future activities.

University Utara Malaysia (UUM) is one of the Malaysian universities, where more than one thousand students register every semester, and where students' data can be personal or academic, and these data are saved by university. School of Computing (SOC) is one of UUM's schools, which saves postgraduate students data such as master students, and this data are not analyzed before by SOC; so this amount of data can be mined, analyzed and transformed into knowledge that leverage the performance of the school to the maximum through the benefit taken from utilizing this data. The students' data can be mined and analyzed to discover useful information and knowledge that can be used to offer a constructive and helpful recommendations to the school to better understand students' behavior, to improve students' performance, and many other benefits.

Most of previous studies in mining educational data used data mining tasks such as classification or prediction to predict student performance and analyze students' data and these studies made use of small data set to discover knowledge from students' data. Also at UUM level very little attempts has been made to mine and analyze students' data and most of these attempts focus on undergraduate students' data with a small limited time period and so postgraduate students' data of SOC are not mined or analyzed previously.

This study aims to extract and discover new valuable knowledge and patterns and to find the relation between student background information with student performance from educational data of School of Computing of University Utara Malaysia using association rule mining as data mining technique to analyze and mine the postgraduate students' educational data to extract meaningful knowledge regarding the student performance.

RELATED WORK

Others [3], defined data mining as the use of advanced data analysis tools to discover previously unknown information in a large amount of data sets. Data mining can provide a number of applications used to search required information, such as observation, patterns, theoretical models, a set of rules and relationships.

The importance of quality in education is a big priority for every educational institution. Data mining has played a key role in assisting the difficult job of improving this quality [4] have applied the data mining techniques in their research to improve the quality of education. The goals were to achieve better understanding of student behavior and also their ability to learn various subjects in the first year university course Electrical engineering fundamentals as well as to predict the success of the next year students based on the success of prior year's students using algorithms and techniques that data mining offers.

Key uses of Educational Data Mining (EDM) include mining student performance and studying learning in order to recommend improvements to current educational practice. EDM can be considered one of the learning sciences, as well as an area of data mining [5].

Studies dedicated to mining and analyzing higher education data using prediction or classification include that of [6] who evaluated students' performance through classification task employing a decision-tree method. They used a data set of 50 students enrolled in Master of Computer Applications faculty in years from 2007-2010. The study was held in the VBS Purvanchal University and the outcomes were presented in the IF-THEN rule form.

Moreover, some authors [7] also employed the WEKA classification decision-tree as a data mining tool for the prediction of loyal students or otherwise in higher education in the hopes of improving the higher educational system in light of its quality.

Furthermore, in the context of India, some [8] concentrated on mining educational data to create predictive data mining model for the performance of students. They attempted to assist the low achieving students in higher secondary level by conducting a survey to produce a database, wherein primary data were gathered among regular students and secondary one from school and office of the CEO. They managed to gather information from five different schools for the year 2006 and following the processing of 772 student records they revealed their outcome. The classified the results through a decision tree (CHAID) – a version of the Automatic Interaction Detector, that used data mining technique of STATISTICA 7.

Another related study comes from [9], for mining educational data since they applied decision tree algorithms on the prior performance data set of students obtained from the VBS Purvanchal University in India from the faculty of Master of Computer Applications (2008-2011). They used the data set to produce a model that predicts students' performance and assists in the determination of which of the students will drop out of their courses, and which of them need attention – this could urge teachers to provide suitable advice to them.

In a related study, [10] developed a model of student performance predictors using the kernel method of data mining in an attempt to analyze the associations between the behavior of students and their success. They also made use of

Smooth Support Vector Machine (SSVM) classification for the predication of the final grade of the students.

Most of these studies in mining educational data used data mining tasks such as classification or prediction to predict performance and these studies made use of small dataset to discover knowledge from students' data and also it just made use of traditional techniques and tools to mine and analyze educational data such as WEKA. So in this study educational data have been analyzed and mined using association rule mining as a data mining technique to discover new valuable knowledge and patterns such as the relation between student background information with student performance and by using an expanded data sets to get more accurate results and RapidMiner also used as analytic tool.

Studies dedicated to mining and analyzing educational data to extracting information include that of [11] conducted study wherein they made use of WEKA data mining tool in the educational context to mine relevant rules that assist in extracting information through Apriori algorithm. They used academic records for a group of 101 students; after which they used K-means clustering to extract information in order to group the students categorically to profile their performance as Good, Satisfactory or Poor.

This study in mining educational data made use of association rule mining asa data mining technique with small dataset to discover knowledge from students' data and also it just made use of traditional tool to mine and analyze educational data which is WEKA. So in this study educational data have been analyzed and mined using association rule mining as data mining technique to discover new valuable knowledge and patterns such as the relation between student background information with student performance and by using an expanded data sets to get more accurate results and also RapidMiner was used as analytics tool.

Others [12] conducted a study in Pakistan metropolitan city to determine the factors that affect the academic performance of 600 secondary school students on 10th grade based on their result of 9th grade annual examination. They analyzed data by applying traditional techniques ANOVA using traditional statistical tool SPSS 16. Another study conducted also in Pakistan by [13] to examine the factors that affect private colleges' student academic performance. They also used traditional appropriate statistical package to analyze the data and get the result.

In the latter study, [2] highlighted the promising benefits of data mining applications in the context of university management. He concentrated on implementation of data mining techniques and methods for obtaining new knowledge from data collected by universities. He attempted to determine and discern data patterns that are invaluable in predicting the performance of university students according to their personal and pre-university characteristics. For data analysis, he made use of Bayes classifiers and a Nearest Neighbor classifier where he made use of the WEKA software.

These studies in analyzing educational data to discover knowledge from students' data made use of traditional tools to analyze educational data, such as WEKA and SPSS and made use of other techniques and method to analyze the data. So in this study educational data have been analyzed and mined using association rule mining as a data mining technique to discover new valuable knowledge and patterns such as the relation between student background information with student performance and RapidMiner was used as analytic tool.

In UUM level [14] conducted a study in UUM on mining programming data set of undergraduate student of Bachelor of Information Technology from Faculty of IT, UUM in the year 2004-2005 to explore the important factors that may impact students programming performance based on information from previous student performance. They used rough set data mining classification framework for extract hidden pattern inside data and discovering relationship in inaccurate data; the result compared to previous works which use the same dataset with different data mining task and techniques. Among these previous studies that use the same programming dataset of UUM are that of [15] which used decision tree as data mining classification technique to classify the performance of students in programming course using same programming dataset of UUM, also in the study of Mohsin, [16] the same programming dataset of UUM has been used to explore the important factors that may impact programming performance of students; they used directed association rule (AR) mining algorithm called apriori.

Another related study held in UMM was that of [17] that used data mining technique to analyzing academic achievement of UUM College of Art and Science (CAS) students. The study focused on undergraduate students who have completed their study in 2006, 2007 and 2008, and logistic regression and artificial neural network have been used for analyzing the student academic achievement.

In UUM very little attempts has been made to mine and analyze students' educational data and most of these attempts focus on undergraduate students' data with a small limited time period and so postgraduate students' data of SOC are not mined or analyzed previously. So this study will mine the postgraduate students' educational data of SOC to discover new valuable knowledge and patterns such as the relation between student background information with student performance. Postgraduate students' educational data of SOC distributed in two data sets the first dataset about student background information contains more than 1800 records and the second dataset about student performance information contains more than 12000 records the period from 1997 to 2012.

PROPOSED METHOD

This study includes four different stages which are: theoretical study, framework development, data collection, and implementation and result. Figure 1 below presents the methodological framework adopted in this study.

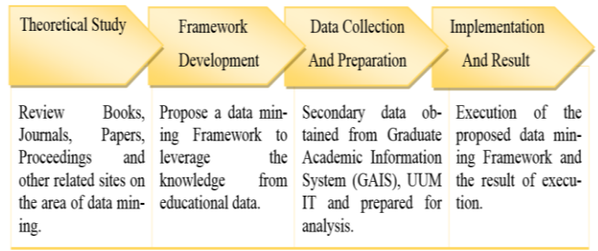

Figure 1: Research Methodological Framework

Theoretical Study

In this stage reviewing books, journals, proceedings, papers, online documentations and other related sites about the topic to understand the background of the study and see the limitation of previous studies. Based on that, research questions, objectives, and scopes are identified.

Framework Development

In this stage framework of the data mining was proposed. Figure 2 shows the flow of data and models in framework. Data goes for cleaning before mining and analyzing. Then this both datasets goes through joining to create a single large data set and all data flowing between nodes holds meta-information concerning the type of its columns in addition to the actual data. The new dataset goes through attribute selection, data conversion and association rules mining algorithm and then the result will be displayed.

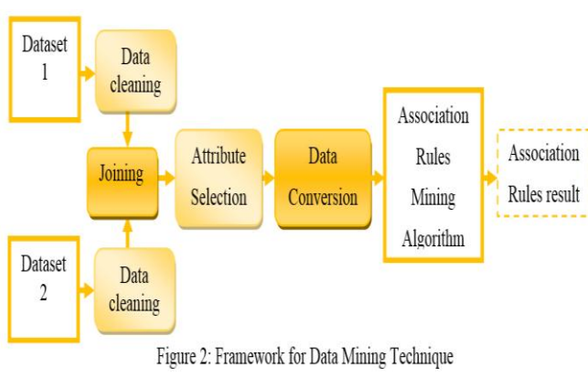

Figure 2: Framework for Data Mining Technique

Data Collection and Preparation

This study is based on the secondary data compilation. The relevant data that have been used for the analysis obtained from authentic source which is Graduate Academic Information System (GAIS), UUM IT. The data is for master students with Master of Science (Information and Communication Technology), Master of Science (Information Technology) and Master of Science (Technopreneurship) of SOC in the period from 1997 to 2012. The data that has been used was distributed in two data sets the first dataset contain more than 1800 records and the second dataset contain more than 12000 records and the following tables showed the details about every datasets and the type of data included.

Table 1: Dataset 1 Student Background Information

Attribute	Data type	Attribute	Data type
Matric	Integer	Qualification CGPA	Real
Name	String	Current program	String
Date of birth	Date	Permanent address	String
Type of Study	String	Organization	String
Year-ins	Date	Position	String
Date of register	Date	Years' experience	integer
Nation of origin	String	English course	String
Bachelor degree program	String	Point	Real
University granted bachelor degree	String	Result	Real
Qualification	String	Student status	String

Table 2: Dataset 2 Student Performance Information

Attribute	Data type	Attribute	Data type
Matric	Integer	Grade	String
Semester code	String	Total mark	Real
Subject code	String	GPA	Real
Subject name	String	CGPA	Real

The preparation of data for analysis in this study included data cleaning by fill in missing values, smooth noisy data and resolve inconsistencies. Noisy data, there are incorrect and inconsistence attributes values due to data entry problems, so these inconsistencies are corrected by making all inconsistence attributes consistence. There are some tuples that have more than one missing value so they are removed.

All data preparation are done manually record by record through datasets which are stored in an excel file. After data preparation the first dataset about student background information remains contain more than 1800 records and the second dataset about student performance information becomes contain less than 12000 records around 11790 records. All datasets records that are prepared in this phase are used in the analysis and mining process.

Implementation and Result

After constructing the data mining framework, which was mentioned in the Framework Development stage and after getting and collecting the data for analysis as mentioned in the Data Collection stage. This stage includes the details of execution and running of the analysis using RapidMiner. Association rule mining, which is a descriptive data mining task has been used for the analysis and mining the data. There are different algorithm for Association rule mining such as Apriori, Charm, FP-growth, Closet and MagnumOpus. In this study Apriori and FP-Growth both have been used to get the largest possible number of best rules.

A priori

Takes transactional data in the form of one row for each pair of transaction and item identifiers. It first generates frequent itemsets and then creates association rules from these item sets. It can generate both association rules and frequent itemsets. A priori supports many different configuration settings and the search type is breadth first search [18] Frequent pattern growth also labeled as FP growth is a tree based algorithm to mine frequent patterns in the database, no candidate frequent itemset is needed rather frequent patterns are mined from fp tree and the search type is divide and conquer [19]. FP-Growth process similar to Apriori,

W-FPGrowth

W-FPGrowth is a WEKA extension. In W-FPGrowth, similar as FP-Growth and W-Apriori from retrieving to discretize by frequency then NominaltoBinominal operator was called to change the nominal attributes to binominal attributes which is allowed in W-FPGrowth. Then W-FPGrowth has been called to find frequent pattern and generate the rules.

For all algorithms, the minimum support is set to be 0.1 and confidence is 0.5. The result of the execution of different Association rule mining algorithms ha been displayed and output rules were discovered the valuable knowledge and patterns in the data.

RESULT AND ANALYSIS

The experiment mines, both students data sets together and to identify the only unique characteristic that affect student performance, the experiment is divided into 2 tries which are 10 and 20 best possible rules that can effect student performance that can be generated by using different algorithms for association rule mining; Apriori and FP-Growth.

A priori

In this analysis, the two data sets which are student background dataset and student performance dataset are joined and analyzed using association rules algorithm, namely W-Apriori which is from the WEKA extension in RapidMiner software. The targets of the mining are to see the effect of student background in their performance. As mention earlier, 2 tries were created to identify the interesting relationship among the attributes. The morning started with 10 best roles and the confidence level was set to 0.5 to get most possible rules.

Apriori identifies that there is ainteresting relationship between student background information and their performance. Out of the 10 rules, two of them show that there is a positive relation between Qualification CGPA, and that a student will get high GPA and CGPA.

In Second try (20 best rules), the confidence level was also set as 0.5 and the number of maximum rule can be generated by Apriori is set to 20 rules. A priori identifies that there is an interesting relationship between student background information and their performance. Out of the 20 rules, seven of them represent the relation, three of the rules show that there is a relation between Bachelor of Computer Science as Bachelor degree program, with that a student will get the highest GPA and CGPA. Two other rules show that there is a positive relation between Qualification CGPA, and that a student will get high GPA and CGPA. And the last two rules show that there is a relation between Type of Study is Sambilan (part time), with that a student will get higher GPA and CGPA.

FP-Growth

In this analysis also, the two data sets which are student background dataset and student performance dataset are joined and analyzed using association rules algorithm namely FP-Growth in RapidMiner software. The targets of the mining are to see the effect of student background in their performance.

FP-Growth identifies that there is an interesting relationship between student background information and their performance. Out of the resulted rules, four of them represent the relation and to interpret the resulted rules in the association rules mining, two of the rules indicate that there is a positive correlation between the premises Qualification CGPA and the Type of Study both with the conclusion GPA and CGPA; the other two rules indicate that there is a positive correlation between the Qualification CGPA of the postgraduate students with that a student will get high GPA and CGPA.

W-FPGrowth

In this analysis, the two data sets which are student background dataset and student performance dataset are joined and analyzed using association rules, namely W-FPGrowth which is WEKA-FPGrowth in RapidMiner software. The targets of the mining are to see the effect of student background in their performance. As mentioned earlier, 2 tries were created to identify the interesting relationship among the attributes.

The mining started with 10 best rules and the confident level was set to 0.5 to get most possible rules. W-FPGrowth identifies that there is interesting relationship between student

background information and their performance. Out of the 10 rules, two of them represent the relation and to interpret the two rules in the association rules mining result by W-FPGrowth, the two rule indicates that there is a positive correlation between the Qualification CGPA of the postgraduate students with that a student will get high GPA and CGPA.

In Second try (20 best rules), the confidence level was also set as 0.5 and the number of maximum rule can be generated by W-FPGrowth is set to 20 rules. W-FPGrowth identifies that there is interesting relationship between student background information and their performance. Out of the 20 rules, seven of them represent the relation, three of the rules show that there is a positive relation between Bachelor of Computer Science as Bachelor degree program, with that a student will get the high GPA and CGPA. Two other rules show that there is a positive relation between Qualification CGPA, and that a student will get high GPA and CGPA. And the last two rules show that there is a positive relation between Bachelor of Information Technology as Bachelor degree program, with that a student will get the high GPA and CGPA.

DISSCUSION

The findings and results discovered that among the students' background information Qualification CGPA has a high positive relation with CGPA and GPA. So the students with high Qualification CGPA usually has high achievement GPA and CGPA in their postgraduate master study. Also among the students' background information Type of Study has a positive relation with GPA and CGPA. So students who are Type of Study of them is Sambilan (part time) in master are usually has high achievement GPA and CGPA in their postgraduate master study, this can be justified is that because part time student take less subject per semester so maybe they more focus on and achieve more than full time student. Also among the students' background information Bachelor degree program has a positive relation with CGPA and GPA. So the students with Bachelor degree program Bachelor of Computer Science usually has high achievement GPA and CGPA in their postgraduate master study and also the students with Bachelor degree program Bachelor Of Information Technology usually has high achievement GPA and CGPA in their postgraduate master study, this can be justified is that because the students with Bachelor of Computer Science or Bachelor Of Information Technology have more knowledge about the subjects that offered in master since their background study is IT or Computer Science. Qualification CGPA, Type of Study and Bachelor degree program are important factors that affect student performance which represented by GPA and CGPA.

By comparing the results; W-Apriori and W-FBGrowth gave similar result which out of the twenty rules seven of them gave a relation between students' background information and their performance. W-Apriori shows that the high achievement Qualification CGPA, Type of Study Sambilan (part time) in master and Bachelor degree program Bachelor of Computer Science all these led to high GPA and CGPA. W-FBGrowth shows that the high achievement Qualification CGPA and Bachelor degree program is Bachelor of Computer Science or Bachelor Of Information Technology all these lead to high GPA and CGPA; Bachelor degree program is Bachelor Of Information Technology are not shown in the result of W-Apriori. For FP-Growth result which out of the thirty four rules just four of them gave a relation between students' background information and their performance and it shows that the high achievement Qualification CGPA, Type of Study Sambilan (part time) in master lead to high GPA and CGPA. Therefore W-Apriori and W-FBGrowth are seem better than FP-Growth.

In this study many of students' backgrounds information did not have a relation with their performance. Nation of origin, Qualification, Years' experience and University granted bachelor degree, all these background information are expected to have a relation with the GPA and CGPA but the study shows that all these did not affect and did not have a relation with the master students' performance.

Al traditional association rule mining algorithms only find positive associations between items. Positive associations refer to associations between items existing in transactions. In addition to the positive associations, negative associations can provide valuable information. Therefore, in this study the results just show the positive association between items; Qualification CGPA, Type of Study Sambilan (part time) in master and Bachelor degree program lead to high GPA and CGPA; the negative association which lead to low GPA and CGPA are not shown in this study because Apriori and FB-Growth algorithms are among traditional association rule mining algorithms which only find positive associations between items..

CONCLUSION

Academic achievement has been identified as major concerns in the universities and other educational institutions. Usually, students' academic achievement is measured by their grade point average in every semester which is GPA and at the end of their period of study, CGPA is calculated. In this study, GPA and CGPA is used as a target to investigate and analyzed the relationship and correlation between students' background information with students' performances. Association rule mining are used in this study as data mining technique with different algorithms which are Apriori and FP-Growth. The experiments using Apriori and FP-Growth are executed to identify which of students' background information are strongly effect and have a relation with students' performance.

In future, the study will be further enhanced and other methods will be applied as well in order to obtain more satisfied outcome, which means that the other data mining techniques such as classification, rough set, decision tree and etc. will be used. Then a prediction model can be built to predict student performance by giving their background information. Also increase the data sets to cover years after 2012 and also cover other UUM schools which may lead to new different results that will help UUM to understand their student behavior and performance.

REFERENCES

[1] Siemens, G., & Long, P. "Penetrating the Fog: Analytics in Learning and Education." *EDUCAUSE review*, **46**(5): 30 (2011)

[2] Kabakchieva, D. (2013). "Predicting student performance by using data mining methods for classification. " *Cybernetics and Information Technologies*, **13**(1): 61-72 (2013)

[3] Beikzadeh, M. R., Phon-Amnuaisuk, S., & Delavari, N. "Data mining application in higher learning institutions." *Informatics in Education-An International Journal*, **7** (1): 31-54 (2008)

[4] Vranić, M., Pintar, D., & Skočir, Z. (2007, June). "The use of data mining in education environment." *Telecommunications, 2007. ConTel 2007. 9th International Conference on* (pp. 243-250). IEEE

[5] Goyal, M., & Vohra, R. "Applications of data mining in higher education." *International journal of computer science*, **9**(2): 113 (2012)

[6] Baradwaj, B. K., & Pal, S. "Mining educational data to analyze students' performance. " In *International Journal of Advanced Computer Science and Applications (IJACSA)*, **2**(6): 63-69 (2011)

[7] Shannaq, B., Rafael, Y., & Alexandro, V. Student relationship in higher education using data mining techniques. " *Global Journal of Computer Science and Technology*, **10**(11) (2010)

[8] Ramaswami, M., & Bhaskaran, R. "A CHAID based performance prediction model in educational data mining. " In *International Journal of Computer Science Issues (IJCSI)*, **7**(1): 11-18 (2010)

[9] Yadav, S. K., Bharadwaj, B., & Pal, S. " Data mining applications: A comparative study for predicting student's performance. " *International Journal of Innovative Technology & Creative Engineering*, **12**(1): (2012)

[10] Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. " Prediction of student academic performance

by an application of data mining techniques. " In *International Conference on Management and Artificial Intelligence IPEDR*, **6**,110-114 (2011)

[11] Parack, S., Zahid, Z., & Merchant, F. "Application of data mining in educational databases for predicting academic trends and patterns. " *Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on* (1-4).

[12] Farooq, M. S., Chaudhry, A. H., Shafiq, M., & Berhanu, G. Factors affecting students' quality of academic performance: a case of secondary school level. " *Journal of quality and technology management*, **7**(2):1-14(2011)

[13] Mushtaq, Irfan, and Shabana Nawaz Khan. "Factors Affecting Studentsâ€™ Academic Performance." *Global journal of management and business research*, **12** (9): (2012).

[14] Mohsin, M. F. M., Norwawi, N. M., Hibadullah, C. F., & Wahab, M. H. A. " Mining the student programming performance using rough set. " *Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on* (478-483). IEEE.

[15] Hibadullah, C. F., & Norwawi, M. N. " Classification of student's performance in programming course using decision tree. " *The fifth international conference on information Technology in Asia, Kuching* (315-317) (2007)

[16] Mohsin, M. F. M., Zaiyadi, M. F., Norwawi, N. M., & Wahab, M. H. A. Wahab, M. H. A. " Pattern extraction for programming performance evaluation using Directed Apriori. "*Compilation of Papers* **2**,(2009)

[17] Nor Asiah, A. R. " Analyzing Academic Achievement of CAS's Students Using Data Mining. " Universiti Utara Malaysia, Kedah, Malaysia. (2009)

[18] Mishra, R., & Choubey, A. " Comparative Analysis of Apriori Algorithm and Frequent Pattern Algorithm for Frequent Pattern Mining in Web Log Data. " *International Journal of Computer Science and Information Technologies*, **3**(4): (2012)

[19] Hunyadi, D. "Performance comparison of Apriori and FP-Growth algorithms in generating association rules. " In *Proceedings of the European Computing Conference ISBN* , 978-960, (2011)