Effectiveness of Association Rule Mining for Learning Efficiency Analysis

Jayshree Jha
Dept. of Information Technology
Atharva College of Engineering
Mumbai, India

***Abstract*— Educational Data Mining (EDM) is a promising interdisciplinary research area that deals with the development of methods to explore data originating in an educational perspective. EDM uses multiple computational approaches to analyze educational data in order to study educational questions. Different data mining algorithm has been applied in this domain. One of the most popularly applied techniques in EDM is association rules mining. This research paper surveys the most significant studies carried out in EDM using Association Rule Mining algorithm. Through the analysis of these studies it has been established that there is shortfall in research related to learning efficiency analysis of cognitive tutor using association rule mining. This paper examined the effectiveness of Association Rule Mining algorithm on improving the learning efficiency in the Cognitive Tutor prospectus. Association Rule Mining is used to identify the initial difficulty of knowledge Component (KC) and also the learning rate of these KCs. In order to quickly find all frequent patterns, an improved algorithm for mining association rules based upon Apriori is used.**

Keywords*— *Educational Data Mining, Association Rule Mining, Improved Apriori Algorithm, Cognitive Tutor.

I. INTRODUCTION

The benefit and usage of computers in learning and teaching has advanced significantly over the last decades through various e-learning environment. Nevertheless, the need for improvement and advancement has always been there. Students read, learn, experiment, analyse and explore content, and by doing such activities, they leave trail of information in computer log. There has been always an important research oriented question: how this vast log data can be utilized? The parameters related to e-learning habit of a student could be easily recorded in e-learning system. These data can be analyzed, and used by teacher to adapt tests to maximize performance. Data mining (DM) algorithm and techniques rise as a suitable answer as they are already used in various other research domain (e.g. medicine, business, market research, etc.) with vast amount of provided data. The similar kind of ideas were already successfully applied in e-commerce environment, the first and most popular application of DM, in order to determine clients' interests so as to be able to increase online sales. However, there has been comparatively less progress in this direction in educational domain. Presently this scenario is changing and there is an increasing interest in applying DM to the educational environment [1]. Educational Data Mining (EDM) is an emerging domain, concerned with developing methods for exploring the unique types of data that come from e-learning environment, and using those methods to better understand students, and the settings which they learn in [2]. Its main objective is to analyze these types of data in order to resolve educational research issues [3]. It becomes an imperative and significant research topic that discovering hidden and useful knowledge from such an extensive quantity of data to guide and develop education. The EDM process converts raw log data coming from educational settings into useful information that could potentially have a great impact on educational research and practice. New computer-supported interactive learning methods, intelligent tutoring systems, simulations, games have opened up opportunities to collect and analyze these student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn.

EDM research uses the five categories of data mining methods to accomplish its objectives: Prediction, Clustering, Relationship Mining and Distillation for human judgment. The data mining techniques being widely used in teaching system is association rules mining.

The rest of the paper is structured as follows: Section 2 provides background knowledge about use of Association Rule Mining in EDM. Based on Association Rule Mining, this research surveys the most relevant studies carried out in EDM as a literature survey in Section 3. Section 4 presents the proposed EDM framework using association rule mining for learning efficiency analysis of cognitive tutor. This section also presents the improved Apriori algorithm which uses bottom up approach and support matrix for mining frequent educational data patterns. In Section 5 experimental results has been discussed.

II. ASSOCIATION RULE MINING

Association rule mining is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute values [4]. In association rules we find the co-occurrences among item sets through finding the large item sets. Association rules can be de scripted

formally as follows, Assuming I=(i1, i2 ….., im) is a collection of m different attributes, in the given database D, each record T is a collection of a set of attributes of I. That is T⊆ I, T has a unique identifier TID. If X ⊆ I and X ⊆ T, T includes X. An association rule is the formula as X⇒ Y, X⊆ I, Y⊆ I and X ∩Y = Φ. It indicates that if X appears in a transaction, Y will be lead to appear in the same transaction inevitablely.

A. EDM using Association Rule Mining

The general KDD (Knowledge Discovery and Data Mining) process [5] has the next steps: collecting data, preprocessing, applying the actual data mining tasks and post-processing. These steps can be particularized for association rule mining in the EDM domain as shown in figure 1.

- Collecting data - Most of the current LMSs (Learning Management Systems) do not store logs as text files. Instead, they normally use a relational database that stores all the systems information: personal information of the users (profile), academic results, the user's interaction data, etc.

- Data pre-processing - Most of the traditional data pre-processing tasks, such as data cleaning, user identification, session identification, transaction identification, data transformation and enrichment, data integration and data reduction are not necessary in LMS. Data pre-processing of LMS data is simpler due to the fact that most LMS store the data for analysis purposes. LMSs also employ a database and user authentication (password protection) which allows identifying the users in the logs. Some typical tasks of the data preparation phase are: data discretization, derivation of new attributes and selection of attributes, creating summarization tables transforming the data format.

- Applying the mining algorithms - In this step it is necessary: 1) to choose he specific association rule mining algorithm and implementation; 2) to configure the parameters of the algorithm, such as support and confidence threshold and others; 3) to identify which table or data file will be used for the mining; 4) and to specify some other restrictions, such as the maximum number of items and what specific attributes can be present in the antecedent or consequent of the discovered rules.

- Data post-processing - The obtained results or rules are interpreted, evaluated and used by the teacher for further actions. The final objective is to putting the results into use. Teachers use the discovered information (in form of if-then rules) for making decisions about the students and the LMS activities of the course in order to improve the students' learning.

III. LITERATURE SURVEY

Zaiane et al. [6] have proposed an approach to build a software agent that uses data mining techniques such as association rules mining in order to build a model that represents on-line user behaviors, and uses this model to suggest activities or shortcuts.

Chun-Hsiung [7] et al. in their paper proposed to apply the algorithm of Apriori for Concept Map to develop an intelligent concept diagnostic system (ICDS). Jong et al. [8] in their research conducted a series of experiments to examine the learning logs recorded by the learning platform over several years to learn about the relationships among students' learning behaviors and learning achievement.

Bidgoli et al. [9] proposed a general formulation of interesting contrast association rules and developed an algorithm to discover a set of contrast rules. This tool can help instructors to design courses more effectively, detect anomalies, inspire and direct further research, and help students use resources more efficiently.

Ma et al. [10] in their paper have described a data mining based method for selecting the right students for remedial classes. The key component of this method is a new scoring function called SBA (Score Based on Association Rule). This method has yielded some exceptionally promising results. It outperforms the traditional method significantly.

Chandra et al. [11] in their paper had shown the potential of the association rule mining algorithm for enhancing the effectiveness of academic planners and level advisers in higher institutions of leaning.

Merceron et al. [12] have used association rule mining on data from the Logic-ITA, a web based learning environment to practice logic formal proofs. Association rule mining has been used to find mistakes often occurring together while solving exercises.

Li et al. [13] have designed an intelligent tutoring system based on data mining technology that could return the learners feedback about knowledge points. In order to quickly find all frequent patterns, i.e., knowledge points, an improved algorithm for mining association rules based on FP-growth is presented.

Garcia et al. [14] in this paper describes a collaborative educational data mining tool based on association rule mining and collaborative filtering for the ongoing improvement of e-learning courses and allowing teachers with similar course profiles to share and score the discovered information.

Omar et al. [15] proposed a framework for personalizing e-learning that necessitates careful attention towards individual learning styles. Proposed framework focuses on identifying learning patterns of learners and the sequence of choosing learning resources in relation to their learning styles.

All the previous research in the area of EDM have applied Association Rule Mining for Either finding

relationship between learners' behavior pattern or associating content with user types to built recommendations model. Some of the research has also used Association Rule Mining for finding students mistakes that co-occur and also for finding out students' learning problem. There is shortfall in research related to learning efficiency analysis of cognitive tutor using association rule mining. This research uses association rule mining algorithm to evaluate an existing cognitive model and give suggestions for further improvements.

IV. PROPOSED EDM FRAMEWORK

This research paper presents a complete Education Data Mining framework based on association rule mining technology that could be used to examine the learning efficiency of a Cognitive Tutor Curriculum. In proposed framework Association Rule Mining is used to identify the initial difficulty of knowledge Component (KC) and also the learning rate of these KCs. With the help of generated association rules student performance can be evaluated and it can be shown that how student performance improves with practice. These rules will also help in identifying overpractice and underpractice KCs. Framework of the proposed model consists of following components:

- Data Acquisition from cognitive Model
- Data Preprocessing
- Rule Mining using improved Apriori algorithm
- Rule Analysis

B. Data Acquisition from Cognitive Model

Cognitive Model used in this research for data acquisition is Area Unit of the Geometry (1996 - 1997) Cognitive Tutor accessed via Data Shop [41]. This dataset has the data from the area unit of the Geometry course conducted during the school year 1996-1997. This dataset stores complete answers of students, including wrong answers. The cognitive model implemented in the Tutor has 10 skills or Knowledge Component for "Textbook New" KC model. In DataShop terminology, Knowledge Components (KCs) are used to represent pieces of knowledge, concepts or skills that students need to solve problems. When a specific set of KCs are mapped to a set of instructional tasks (usually steps in problems) they form a KC Model.

C. Data Preprocessing

For this research focus is put on three target variables, Knowledge Component/Skill, Opportunities and Success (Incorrect/Correct), in order to find the association rules involving these attributes. Dataset table obtained from datashop is fragmented according to Knowledge Component (KC) into ten different tables representing each KC. A constant difficulty in using any of the association rule mining algorithms is that they can only operate on binary data sets. Thus, in order to analyze quantitative or categorical attributes, some modifications are required. In this research, equal-frequency binning for discretizing 'Opportunity' attribute is used and it is mainly categorized into nominal values: LOW, MEDIUM and HIGH.

D. Rule Mining using Improved Apriori

Association rule mining using improved Apriori is executed on preprocessed dataset. Improved algorithm is used to find association between opportunity and Success for each Knowledge Component or Skill. This research uses an Improved Apriori algorithm based on Bottom up approach and Support matrix to identify frequent item set [19].

Improved Apriori Algorithm for EDM

Presented algorithm uses Bottom up Approach to find the frequent item set from largest frequent Item set to smallest frequent item set which help in mining pattern easily. This algorithm works in two phases: - Support Matrix Generation and Bottom Up approach to mine frequent items set.

Phase1:- (Generation of sorted Support matrix) Steps to generate sorted Support Matrix are as follows:-

Step 1:- Scan database to generate Boolean matrix A1. Rows in matrix represent transaction. Columns in matrix represent items. Each cell will have the values either 1 or 0 for representing presence of items in the transaction. Entry value 1 indicates the corresponding item is present in transaction and value 0 indicates the corresponding item is not present in transaction. Simplify the matrix A1. Remove the items from the A1 for which support is less than Minsupport.

Step 2:- Calculate the Support value of each item. Boolean Matrix A1 is transformed to Support Matrix A2, by replacing each entry value of 1 by the Support value of corresponding item and inserting two more columns to the Support Matrix A2 to hold total Support value and Count of elements in each row respectively.

Step 3:- The Support Matrix A2 will be rearranged in descending order in accordance with total Support value and non-zero entry will be replaced by 1 which leads to generation of Sorted Support Matrix A3.

Phase2:- Bottom Up approach to mine frequent items.

Step 1: Select first transaction from A3 and compare its total Support value and count with next transaction total support and count respectively.

Step 2 : If the next transaction total support value and count is greater than or equal to the first transaction then do the BITWISE AND operation between the transaction, if the resultant of AND operation is equal to first transaction structure then increase the support count of first transaction item set by 1. Continue this process with remaining transaction.

Step 3: At the end check the total support count of first transaction if it is greater than or equal to the Minimum support count extract the item set of that transaction and all

its subset and move it to frequent Item set. The same process will be repeated for remaining transaction.

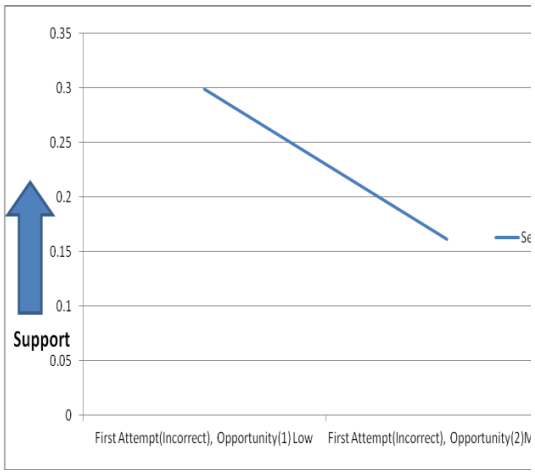
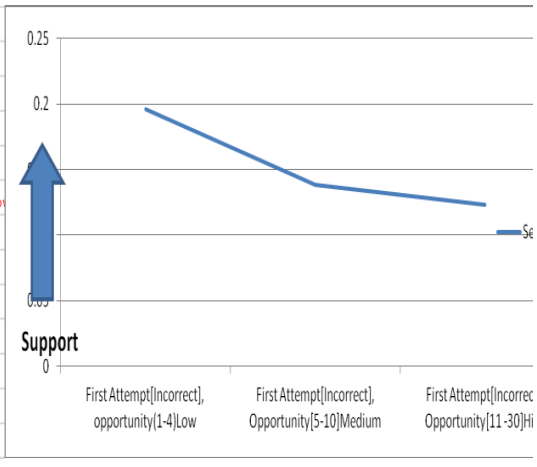
E. Rule Analysis

Improved algorithm is used to find association between Opportunity and Success for each Knowledge Component or Skill.

V. EXPERIMENTAL RESULT

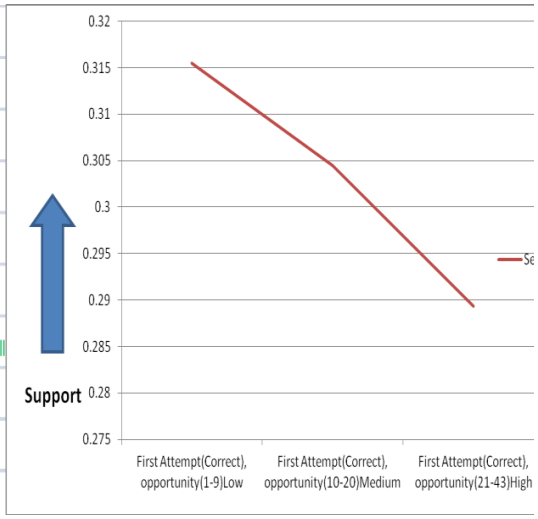
Association Rules Generated for some of the Knowledge Components (KCs) are listed below in Table 1.

TABLE I. ASSOCIATON RULE GENERATED FOR KNOWLEDGE COMPONENTS

ASSOCIATION RULES	GRAPH
KC-Equi-tri-height Itemset Cou | Association Rules generated | Support 1- itemset | Opportunity(3 - 6)High | 0.356322 1- itemset | Opportunity(2)Medium | 0.321839 1- itemset | Opportunity(1) Low | 0.321839 1- itemset | First Attempt(Incorrect) | 0.551724 | Overall success rate is very low 1- itemset | First Attempt(correct) | 0.448276 2 - itemset | First Attempt(correct), Opportunity(3 - 6)High | 0.264368 2 - itemset | First Attempt(Incorrect), Opportunity(2)Medi | 0.16092 2 - itemset | First Attempt(correct), Opportunity(2)Mediun | 0.16092 2 - itemset | First Attempt(Incorrect), Opportunity(1) Low | 0.298851 | Intrinsic difficult skill	
KC-Trapezoid Itemset Count | Association Rules generated | Support 1- itemset | opportunity(1-4)Low | 0.3125 1- itemset | Opportunity[5-10]Medium | 0.354910714 1- itemset | Opportunity[11-30]High | 0.332589286 1- itemset | First Attempt[Incorrect] | 0.457589286 | Overall success rate is very lo 1- itemset | First Attempt(Correct) | 0.542410714 2 - itemset | First Attempt[Incorrect], opportunity(1-4)Low | 0.196428571 2 - itemset | First Attempt(Correct), opportunity(1-4)Low | 0.116071429 | intrinsically difficult skill 2 - itemset | First Attempt[Incorrect], Opportunity[5-10]Medi | 0.138392857 2 - itemset | First Attempt(Correct), Opportunity[5-10]Mediu | 0.216517857 2 - itemset | First Attempt[Incorrect], Opportunity[11-30]Hig | 0.122767857 2 - itemset | First Attempt(Correct), Opportunity[11-30]High | 0.209821429	

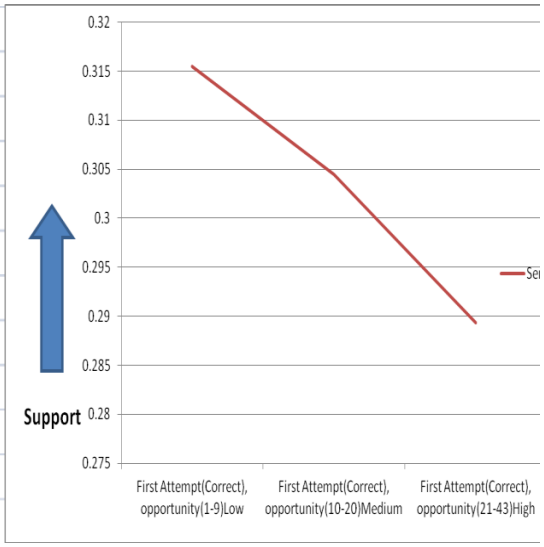
KC- Parallelogram

Itemset Count	Association Rules generated	Support	
1- itemset	opportunity(1-9)Low	0.3528226	
1- itemset	opportunity(10-20)Medium	0.3306452	
1- itemset	opportunity(21-43)High	0.3165323	
1- itemset	First Attempt(Correct)	0.9092742	overall high success rate
2 - itemset	First Attempt(Correct), opportunity(1-9)Low	0.3155242	student mostly have this skill
2 - itemset	First Attempt(Correct), opportunity(10-20)Mediu	0.3044355	
2 - itemset	First Attempt(Correct), opportunity(21-43)High	0.2893145	



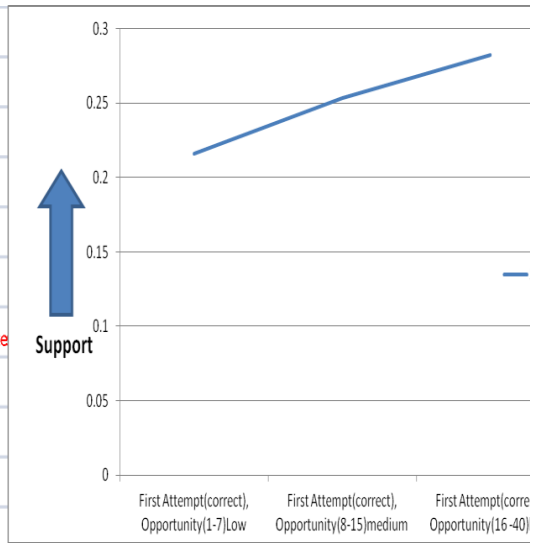
KC-Pentagon

Itemset Count	Association Rules generated	Support	
1- itemset	opportunity(1-6)Low	0.357798165	
1- itemset	Opportunity[7-13]Medium	0.328440367	
1- itemset	Opportunity[14 -27] High	0.313761468	
1- itemset	First Attempt[Incorrect]	0.352293578	Overall success rate is low
1- itemset	First Attempt(Correct)	0.647706422	
2 - itemset	First Attempt[Incorrect], opportunity(1-6)Low	0.185321101	
2 - itemset	First Attempt(Correct), opportunity(1-6)Low	0.172477064	intrinsic difficult skill
2 - itemset	First Attempt[Incorrect], Opportunity[7-13]Mediu	0.104587156	
2 - itemset	First Attempt(Correct), Opportunity[7-13]Mediu	0.223853211	
2 - itemset	First Attempt(Correct), Opportunity[14 -27] High	0.251376147	



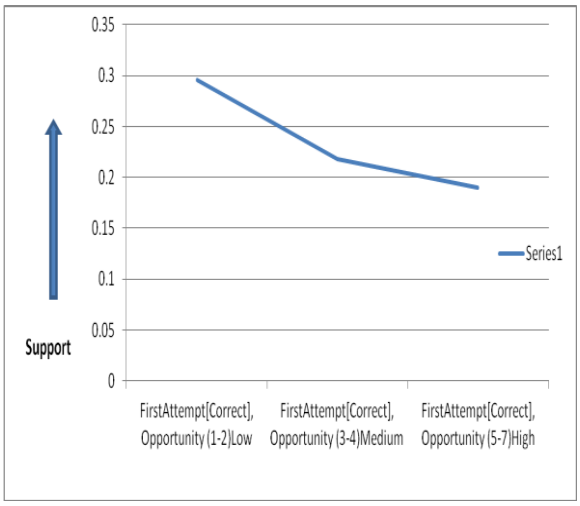
KC-Circle-area

Itemset Count	Association Rules generated	Support	
1- itemset	Opportunity(1-7)Low	0.357348703	
1- itemset	Opportunity(8-15)medium	0.309798271	
1- itemset	Opportunity(16 -40)high	0.332853026	
1- itemset	First Attempt(Incorrect)	0.247838617	
1- itemset	First Attempt(correct)	0.752161383	Overall good succe
2 - itemset	First Attempt(Incorrect), Opportunity(1-7)Low	0.141210375	
2 - itemset	First Attempt(correct), Opportunity(1-7)Low	0.216138329	
2 - itemset	First Attempt(correct), Opportunity(8-15)medium	0.253602305	
2 - itemset	First Attempt(correct), Opportunity(16 -40)high	0.282420749	



KC-Composed-by-addition

Itemset Count	Association Rules generated	Support
1- itemset	Opportunity(1-6)Low	0.359756098
1- itemset	Opportunity(7-14)medium	0.3125
1- itemset	Opportunity(15 - 35)High	0.327743902
1- itemset	Result(Incorrect)	0.260670732
1- itemset	Result(Correct)	0.739329268
2 - itemset	Result(Correct), Opportunity(1-6)Low	0.291158537
2 - itemset	Result(Correct), Opportunity(7-14)medium	0.237804878
2 - itemset	Result(Incorrect), Opportunity(15 - 35)High	0.117378049
2 - itemset	Result(Correct), Opportunity(15 - 35)High	0.210365854



F. Result Analysis

By analyzing the result of generated association rules in table 1, it has been found that for most of the KC students' performance improves with practice which goes by the popular belief. Apart from this it has been discovered that some of the KCs are intrinsically difficult with overall low or average success rate but have high learning rate for example Trapezoid-area and Pentagon area. These KCs are under-practiced KC and more problems need to be added for these KCs to improve overall success rate. On the other hand there are KCs for which generated association rules are showing no apparent learning. Parallelogram area and Composed-by-addition are such KCs. Parallelogram area KC is over practiced KC as it is an intrinsically easy skill and also it has good success rate. So, number of practice for this KC can be reduced and thus students' time can be saved. Composed by addition KC is more of concerned and is targeted for improvement as success rate for this KC is 73% which can still be improved. Thus, we can say that Composed-by-addition is a KC with no apparent learning and need improvement. By decomposition of this skill into subskill results can be improved.

VI. CONCLUSION

In this paper we have used Association Rule Mining Algorithm to identify the initial difficulty of knowledge Component (KC) and also the learning rate of these KCs. With the help of generated association rules we were able to evaluate the student performance in particular KC/skill and it is shown that how student performance improves with practice. These rules also helped in identifying overpractice and underpractice KCs. Moreover, based on the association rule mining results we were able to make suggestion for improvement in KC Composed-by-addition. Thus, we were able to use association rule mining algorithm to evaluate an existing cognitive model and give suggestions for further improvements.

ACKNOWLEDGMENT

My sincere thanks to all the people who have contributed in carrying out this work.

REFERENCES

[1] D. Krpan and S. Stankov, "Educational data mining for grouping students in e-learning system," in Proceeding of Information Technology Interfaces, June 2012.

[2] [Online]. Available: www.educationaldatamining.org

[3] C. Romero, "Educational data mining: A review of the state of the art," IEEE transaction on systems, MAN, and Cybernetics-part c: Application and Reviews, November 2010.

[4] Z. Zheng, R. Kohavi, and L. Mason, "Real world performance of association rules," in International Conference on Knowledge Discovery & Data Mining, 2001.

[5] E. García, C. Romero, S. Ventura, and T. Calders, "Drawbacks and solutions of applying association rule mining in learning management systems," in Proceedings of the International Workshop on Applying Data Mining in e-Learning, 2007.

[6] O. R. Zaiane, "Building a recommender agent for e-learning systems," in Proceedings of the International Conference on Computers in Education (ICCE'02), 2002.

[7] C. H. Lee, G. G. Lee, and Y. Leu, "Application of automatically constructed concept map of learning to conceptual diagnosis of e-learning," in Expert Systems with Applications, 2009.

[8] P. Markellou, L. Mousourouli, and A. Tsakalidis, "Using semantic web mining technologies for personalized e-learning experiences," in Web-based Education, 2005.

[9] B. Jong, T. Chan, and Y. Wu, "Learning log explorer in e-learning diagnosis," in IEEE transactions on education, 2007.

[10] N. Selmoune and Z. Alimazighi, "A decisional tool for quality improvement in higher education," in Information and Communication Technologies: From Theory to Applications, 2008.

[11] B. Bidgoli, P. Tan, and W. Punch, "Mining interesting contrast rules for a web-based educational system," in International Conference on Machine Learning and Application, 2004.

[12] Y. Ma, B. Liu, C.Wong, P. Yu, and S. Lee, "Targeting the right students using data mining," in international conference on Knowledge discovery and data mining, 2000.

[13] E. Chandra and K. Nandhini, "Knowledge mining from student data," European Journal of Scientific Research, 2010.

[14] A. Merceron and K. Yacef, "Revisiting interestingness of strong symmetric association rules in educational data," in Proceedings of the International Workshop on Applying Data Mining in e-Learning, 2007.

[15] Y. Li and S. Zhao, "An association rule mining approach for intelligent tutoring system," in 2nd International Conference on Computer Engineering, 2010.

[16] E. Garcia, C. Romero, S. Ventura, and C. Castro, "A collaborative educational association rule mining tool," Internet and Higher Education 14, 2011.

[17] H. Omar, I. Petrounias, and F. Anwar, "A framework for using web usage mining to personalise e-learning," in Seventh IEEE International Conference on Advanced Learning Technologies, 2007.

[18] [Online]. Available: http://pslcdatashop.org (Koedinger et al., 2010).

[19] J. Jha and L.Ragha, "Educational Data Mining using Improved Apriori Algorithm", International Journal of Information and Computation Technology, Volume 3, 2013, pp 411-418.