Data Preparation Strategy in E-Learning System using Association Rule Algorithm

Sunita B Aher
M.E. (CSE) -II
Walchand Institute of Technology
Solapur University India

Lobo L.M.R.J.
Associate Professor, Head, Department of IT
Walchand Institute of Technology
Solapur University India

ABSTRACT
Data preparation is the important step in Course Recommendation System which aims at predicting the course selected by student. In this paper we present the data preparation strategy for Course Recommendation System. Here we have used the real data of courses offered through Moodle package of the college & apply the data preparation strategy to it to find out the best combination of courses. We are using the open source data mining tool Weka to check the result.

KEYWORDS
Apriori Association Rule, Weka, Moodle

1. INTRODUCTION
Data Mining can be used to extract knowledge from e-learning systems such as Moodle, through the analysis of the information available in the form of data generated by their users. The main objective becomes finding the patterns of system usage by teachers and students and, perhaps most importantly, discovering the students' learning behavior patterns [1]. The Course Recommendation System in E-Learning is a system that suggests the best combination of courses in which the students are interested [8]. Here we are using Moodle as Learning Management System to collect the data regarding the course selection by student. We have created the student login & gave the access to the student. For collecting the data we are considering the student of three years of engineering course Second year, Third year, & Final year of Computer Science & Engineering and Information Technology. Student will enroll for those set of courses in which they are interested. This enrollment of courses information is stored in Moodle database which we use to find the best combination of courses. This architecture for Course Recommendation System is explained in [8].

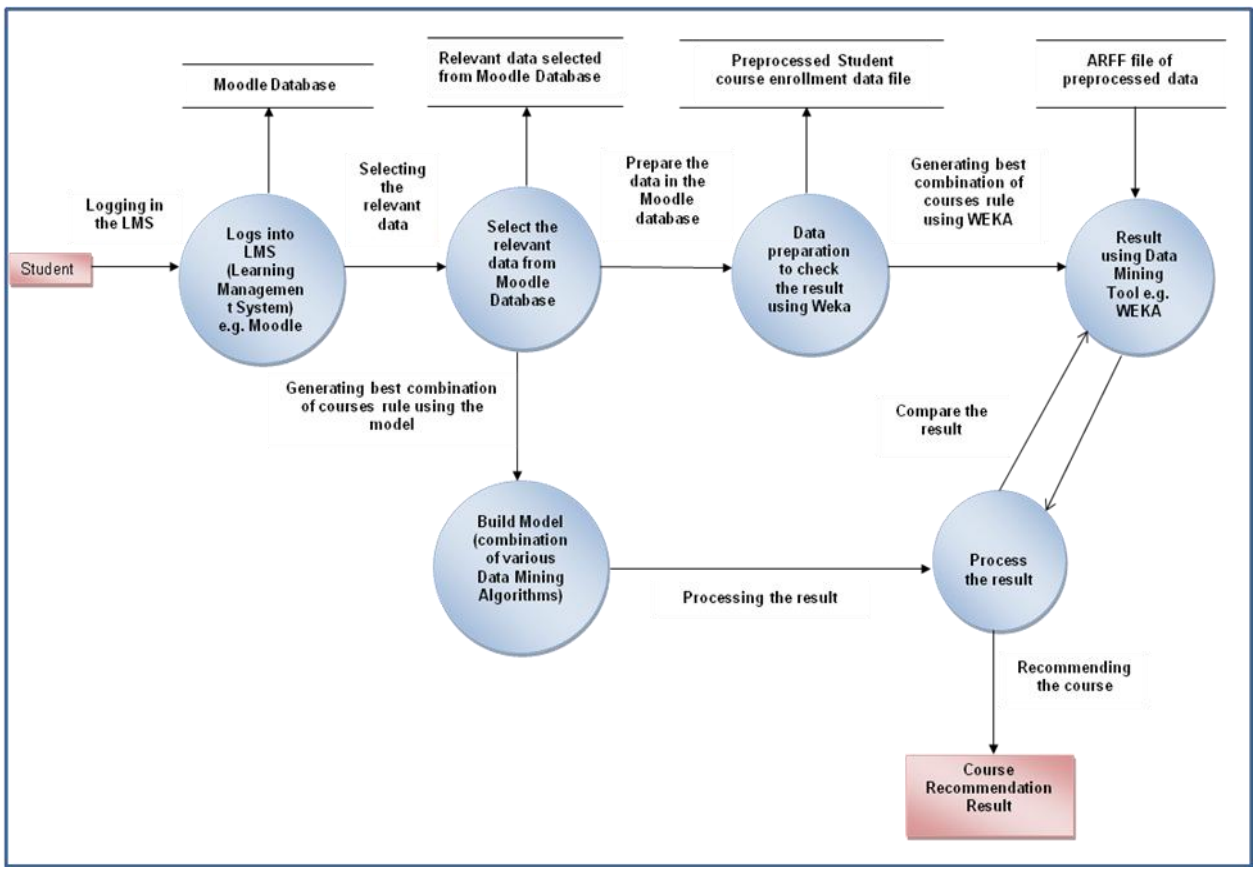

Figure 1: Data Flow Diagram for Course Recommendation System

2. LITERATURE REVIEW

Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás [2] compared different data mining methods and techniques for classifying students based on their Moodle usage data and the final marks obtained in their respective courses. They have developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors.

The research [3] proposes a framework of a personalized learning recommender system, which aims to help students find learning materials they would need to read. Two related technologies are developed under the framework: one is a multi-attribute evaluation method to justify a student's need, and another is a fuzzy matching method to find suitable learning materials to best meet each student need. The implementation of this proposed personalized learning recommender system can support students online learning more effectively and assist large class online teaching with multi-background students.

Seki, K., Tsukahara, W., & Okamoto, T [4] developed an integrated e-Learning environment which integrates learning history and learning content information to control each learner's learning. They have been developing an LMS based on Learning Ecological Model which is a model of Learning Environment focusing on learning content, learning objective, and learning style. By analyzing access log and report submission log, they obtained suggestions for improvement such as restriction of browsing period to make students access constantly, fragmentation of report submission deadline for students not keeping report assignment too.

C. Romero, S. Ventura and E. Garcia [5] described the full process for mining e-learning data step by step as well as how to apply the main data mining techniques used, such as statistics, visualization, classification, clustering and association rule mining of Moodle data.

Hamalainen, W., Suhonen, J., Sutinen, E., & Toivonen, H designed and implemented Data Mining System (DMS) to analyze the study records of two programming courses in a distance curriculum of Computer Science. Various data mining schemes, including the linear regression and probabilistic models, were applied to describe and predict student performance.

The purpose of study [10] was to identify and examine learning processes, based on data extracted from log files, which document the learners' action within an online learning environment. For this purpose, log files of four elementary school students, studied with a science Web-based module, were examined and analyzed. A Learnogram - graphical representation tool that visualizes students' learning process over time - was produced for each student. Based on the log files and the Learnograms, seven learning variables were defined and computed, reflecting the differences between the learning processes.

3. DATA PREPARATION STRATEGY

Here in this Course Recommendation System, we have considered the 13 course category which is shown in following table 1. Under each category there will courses. So there are about 82 courses.

Table 1: Course category & courses [8]

Sr No	Course Category	Course Name & Code
1	Basic Science (BS)	1.Engineering Physics(PHY) 2.Engineering Mathematics-I (M-I) 3.Engineering Mathematics-II (M-II) 4.Applied Mathematics-I (AM-I) 5.Applied Mathematics-II (AM-II) 6.Engineering Chemistry (CHEM)) 7.Discrete Mathematical Structure (DMS) 8.Applied Mechanics (MECH)
2	Basic Foundations of Engineering (BFE)	1.Engineering Graphics(EG) 2.Basic Electronics And Computer Programming (BECP) 3.Workshop Practice (WP) 4.Basic Civil Engineering (BCE) 5.Basic Electrical Engineering (BEE) 6.Basic Mechanical Engineering (BME)
3	Humanities & Social Science (HSS)	1.Communication Skills (COMM) 2.Environmental Sciences (ES) 3.Management Information System (MIS)
4	Hardware Design & Engineering (HDE)	1.Switching Theory And Logic Design (STLD) 2.Microprocessor (MP) 3.Advanced Microprocessor (AMP) 4.Computer Organization (CO) 5.Advanced Computer Architecture (ACA) 6.VLSI Technology (VLSI) 7.Microcontroller (MICRO) 8.Digital Signal Processing (DSP) 9.Mobile Computing (MC) 10.Natural Language Processing (NLP)
5	Theoretical Computer Science (TCS)	1.Data Structures –I (DS-I) 2.Data Structures –II (DS-II) 3.Design & Analysis of Algorithm (DAA) 4.Formal System And Automata (FSA) 5. Theory Of Computation (TOC)
6	System Software (SS)	1.System Programming (SP) 2.Compiler Construction (CC) 3.Operating System – I (OS-I) 4.Operating System – II (UNIX) (OS-II) 5.Distributed Systems (AOS)
7	Networks (NT)	1.Computer Networks – I (CN-I) 2.Computer Networks – II (CN-II) 3.Information Security (IS) 4.Network Engineering (NE) 5.Network Security (NS) 6.Computer Network Administration (CNA)
		1.Database Engineering (DBE)

8	Database (DB)	2.Advanced Database System (ADS) 3.Advanced Database Design (ADD) 4.Oracle (SQL) 5.Bioinformatics(BIO)
9	Software Engineering & Principles (SEP)	1.Software Engineering (SE) 2.Object Oriented Modeling And Design (OOMD) 3.Software Testing & Quality Assurance (STQA) 4.Information Technology (IT) 5.Principles Of Management & Engineering Economics (PMEE) 6. Component Technology (CT)
10	Image Processing, Graphics And Artificial Intelligence Application (IP&AIP)	1.Computer Graphics (CG) 2.Image Processing (IP) 3.Embedded System (ESYS) 4.Pattern Recognition (PR) 5.Information Retrieval (IR) 6.Artificial Intelligence (AI) 7.Artificial Neural Network (ANN) 8.Human Computer Interfaces (HCI) 9.Expert System (ES)
11	Web Technology (WT)	1.Hyper Text Markup Language (HTML) 2.Dynamic Html (DHTML) 3.JavaScript (JS) 4.Extensible Markup Language (XML) 5.Java Server Pages (JSP) 6.Servlets (STS) 7.Active Server Pages (ASP) 8.ASP.NET (ASP.NET) 9.Hypertext Preprocessor (PHP)
12	Principles of Programming Language(PPL)	1.Visual Basic (VB) 2.C-Programming (CP) 3.Advanced C (AC) 4.Object Oriented Design & Programming (C++) (OODP) 5.Java Programming (JP) 6.Advanced Java (AJ) 7.Visual C++ (VC++) 8.C-Sharp (CS)
13	Information Retrieval & Extraction (IRE)	1.Data Warehousing (DW) 2.Data Mining (DM)

In this data preparation strategy, first we select the data from Moodle database where the student's enrollment for specific set of courses is stored. Here we integrate data from various Moodle database into one table. Finally we prepare the table in which the row represents the student roll number to indicate that particular student is interested in that course & column represents the course name as mentioned in table 1. As we are using open source data mining tool Weka, we have to convert that table in ARFF file & check the result. If result contains the rule having "no" only then we delete those course from the table having lowest student count & those student, having lowest course count. Check the result again using Weka & if still it contains the association rule having "no" then increase course count. Repeat this process until we will get the association rule having "yes" only. The flowchart for this data preparation step is shown in figure 2.

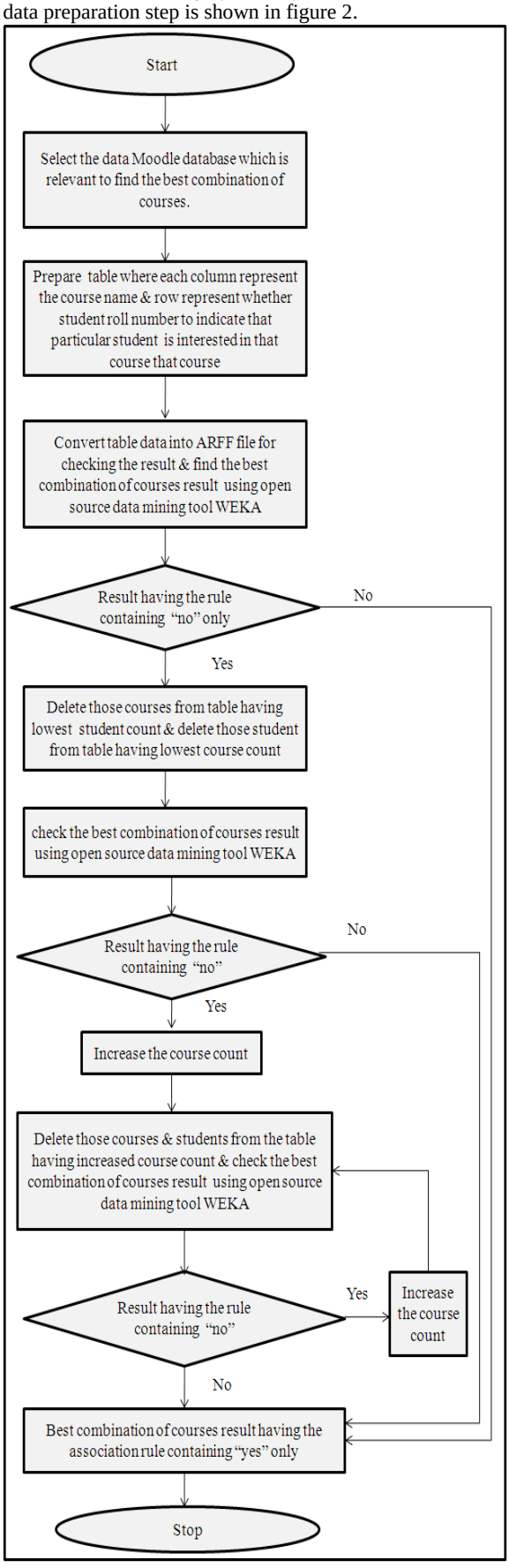

Figure 2: Data Preparation Strategy

4. APRIORI ASSOCIATION RULE

Association rules are used to find the frequent pattern, association or correlation in transaction database. Association rule mining can be used in Basket Data Analysis, Educational Data Mining, Classification, Clustering etc. Association Rule algorithms are Apriori, Sampling, Partitioning & Parallel Algorithm

Apriori Association rule is used to mine the frequent patterns in database. Support & confidence are the normal method used to measure the quality of association rule.

Support for the association rule X->Y is the percentage of transaction in the database that contains XUY [9]. Confidence for the association rule is X->Y is the ratio of the number of transaction that contains XUY to the number of transaction that contain X [9]. The Apriori association rule algorithm is given below [9].

Apriori Association Rule Algorithm

Input : Database of Transactions D={t_1,t_2, …, t_n}
Set if Items I= {I_1 ,I_2 ,….I_k}
Frequent (Large) Itemset L
Support,
Confidence.

Output : Association Rule satisfying Support & Confidence

Method :

C_1 = Itemsets of size one in I;
Determine all large itemsets of size 1, $L_{1;}$
i = 1;
Repeat
i = i + 1;
C_i = Apriori-Gen(L_{i-1});
Apriori-Gen(L_{i-1})

1. Generate candidates of size i+1 from large itemsets of size i.
2. Join large itemsets of size i if they agree on i-1.
3. Prune candidates who have subsets that are not large.

Count C_i to determine $L_{i;}$
until no more large itemsets found;

5. RESULT

While collecting the data we have considered the different year of engineering courses i.e. second year, third year & final year of engineering. Here we have collect the data from two engineering courses: Computer science & Engineering and Information Technology. The result we obtained using Apriori Association Rule which is existing algorithm in open source data mining tool Weka before preprocessing is shown in table 2.As the rules contains only "no" in the result, we prepared the data collected from the Moodle database. As we are recommending the courses, we have to find out the rules containing "yes" only. After preparation of data, we get result using Apriori Association Rule which is shown in table 2. As we increase the support count, we get the refined rule. The meaning of the rule is that if student is interested in Data Structure-II then he/she is interested in Data Structure-I. Due to above rule, we can recommend to new student who has recently enrolled for Data Structure-II course, the Data Structure-I as a course to be opted. Association rule obtained here also match with the in general real world interdependencies among the course.

Table 2: Result using Apriori association rule algorithm

Courses considered	**Result using machine learning algorithm e.g. Apriori Association Algorithm**
Before Data preparation strategy	
	Minimum support: 0.95 **Minimum metric <confidence>: 0.9** **Number of cycles performed: 1** **Best rules found:** 1. INFORMATION_RETRIEVAL=no → MICROCONTROLLER=no conf:(1) 2. VLSI_TECHNOLOGY=no →MICROCONTROLLER=no conf:(1) 3. PATTARN_REGOGNITION=no →MICROCONTROLLER=no conf:(1) 4. NATURAL_LANGUGE_PROCESSING=no →MICROCONTROLLER=no conf:(1) 5. VLSI_TECHNOLOGY=no INFORMATION_RETRIEVAL=no →MICROCONTROLLER=no conf:(1) 6. PATTARN_REGOGNITION=no INFORMATION_RETRIEVAL=no → MICROCONTROLLER=no conf:(1) 7. VLSI_TECHNOLOGY=no PATTARN_REGOGNITION=no →MICROCONTROLLER=no conf:(1) 8. EMBEDDED_SYSTEM=no →MICROCONTROLLER=no conf:(1) 9. ARTFICIAL_NEURAL_NETWORK=no →MICROCONTROLLER=no conf:(1) 10. NATURAL_LANGUGE_PROCESSING=no INFORMATION_RETRIEVAL=no →MICROCONTROLLER=no conf:(1)
After Data preparation step	
STLD, DS-I, DS-II, CP, JP	**Minimum support: 0.55 (52 instances)** **Minimum metric <confidence>: 0.9** **Number of cycles performed: 9** **Best rules found:**

	1. DATA_STRUCTURE_II=yes →DATA_STRUCTURE_I=yes conf:(0.98) 2. JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.98) 3. C_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 4. DATA_STRUCTURE_II=yes JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 5. SWITCHING_THEORY_AND_LOGIC_DESIGN=yes →DATA_STRUCTURE_I=yes conf:(0.97) 6. DATA_STRUCTURE_II=yes C_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 7. C_PROGRAMMING=yes JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 8. SWITCHING_THEORY_AND_LOGIC_DESIGN=yes DATA_STRUCTURE_II=yes →DATA_STRUCTURE_I=yes conf:(0.96) 9. SWITCHING_THEORY_AND_LOGIC_DESIGN=yes JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.96) 10. DATA_STRUCTURE_II=yes C_PROGRAMMING=yes JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.96)
	Minimum support: 0.6 (56 instances) **Minimum metric <confidence>: 0.9** **Number of cycles performed: 8** **Best rules found:** 1. DATA_STRUCTURE_II=yes →DATA_STRUCTURE_I=yes conf:(0.98) 2. JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.98) 3. C_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 4. DATA_STRUCTURE_II=yes JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 5. SWITCHING_THEORY_AND_LOGIC_DESIGN=yes →DATA_STRUCTURE_I=yes conf:(0.97) 6. DATA_STRUCTURE_II=yes C_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 7. C_PROGRAMMING=yes JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97)
	Minimum support: 0.7 (66 instances) **Minimum metric <confidence>: 0.9** **Number of cycles performed: 6** **Best rules found:** 1. DATA_STRUCTURE_II=yes →DATA_STRUCTURE_I=yes conf:(0.98) 2. JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.98) 3. C_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 4. DATA_STRUCTURE_II=yes JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97) 5. SWITCHING_THEORY_AND_LOGIC_DESIGN=yes →DATA_STRUCTURE_I=yes conf:(0.97) 6. DATA_STRUCTURE_II=yes C_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97)
	Minimum support: 0.6 (145 instances) **Minimum metric <confidence>: 0.9** **Number of cycles performed: 8** **Best rules found:** 1. DATA_STRUCTURE_II=yes →DATA_STRUCTURE_I=yes conf:(0.98) 2. JAVA_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.98) 3. C_PROGRAMMING=yes →DATA_STRUCTURE_I=yes conf:(0.97)

6. CONCLUSION AND FUTURE WORK

Here we explain how to prepare the course data obtained from Moodle database for Course Recommendation System. Before data preparation step, we get the result using Apriori association rule containing "no" only which are not correct association rules. As we are recommending the course to the student, we need the association rule containing "yes" only. After preparation of data, we are getting the correct association rules at predicting the course selection by student. Association rule obtained using Apriori Association Algorithm also match with in general the real world interdependencies among the course. Future work includes finding out the algorithm which may be the combination of various data mining algorithm which could be applied on the huge amount of data obtained from Moodle course.

7. REFERENCES

[1] Castro, F., Vellido, A., Nebot, A., & Mugica, F. (in press). Applying data mining techniques to e-learning problems: A survey and state of the art. In L. C. Jain, R. Tedman, & D. Tedman (Eds.), Evolution of Teaching and learning paradigms in intelligent environment. Studies in Computational Intelligence (Vol. 62). Springer-Verlag.

[2] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás: Data Mining Algorithms to Classify Students-accessed from http://sci2s.ugr.es/keel/pdf/specific/congreso/Data%20Mining%20Algorithms%20to%20Classify%20Students.pdf on 15-03-2012

[3] J Lu: A personalized e-Learning Material Recommender system: Proceedings of the 2nd International Conference on IT for application, ISBN: 0646423134

[4] Seki, K., Tsukahara, W., Okamoto, T.: System Development and Practice of e-Learning in Graduate School. In: The Fifth IEEE International Conference on Advanced Learning Technologies, ICALT'05. IEEE Computer Society, Los Angeles, CA (2005) 740-744.

[5] C. Romero, S. Ventura and E. Garcia. Data Mining in Course Management Systems: MOODLE Case Study and Tutorial. Computers and Education, 2007. Num. 51. pp. 368-384.

[6] Hamalainen, W., Suhonen, J., Sutinen, E., & Toivonen, H. (2004). Data mining in personalizing distance education courses. In World conference on open learning and distance education, Hong Kong (pp. 1–11).

[7] Sunita B Aher and Lobo L.M.R.J.. Data Mining in Educational System using WEKA. *IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT)* (3):20-25, 2011. Published by Foundation of Computer Science, New York, USA (ISBN: 978-93-80864-71-13)

[8] Sunita B Aher and Lobo L.M.R.J. Article: A Framework for Recommendation of courses in E-learning System. *International Journal of Computer Applications* 35(4):21-28, December 2011. Published by Foundation of Computer Science, New York, USA

[9] "Data Mining Introductory and Advanced Topics" by Margaret H. Dunham

[10] Galit Ben-Zadok, Arnon Hershkovitz, Rachel Mintz and Rafi Nachmias: "Examining online learning processes based on log files analysis: A case study" Research, Reflections and Innovations in Integrating ICT in Education.