2011 2nd International Conference on Networking and Information Technology
IPCSIT vol.17 (2011) © (2011) IACSIT Press, Singapore

Knowledge Mining in Supervised and Unsupervised Assessment Data of Students' Performance

Anwar M. A. and Naseer Ahmed

[1]College of Engineering and Applied Sciences, Al Gurair University, Academic City Dubai, United Arab Emirates

[2]Institutional Research and Effectiveness, Al Gurair University, Academic City Dubai, United Arab Emirates

anwarma@yahoo.com, naseer@agu.ac.ae

Abstract. There are several statistical tools being used for students' performance analysis for information extraction and knowledge discovery. This paper presents data mining approach applied to discover students' performance patterns in supervised and unsupervised assessment instruments of a course in an engineering degree program. The interesting patterns emerging from this analysis promise to offer some helpful and constructive suggestions to educational administrators and decision makers in the sector of higher education for the improvement and revision of assessment methodologies, restructuring the curriculum, and trimming down the mismatch between the two modes of assessments.

Keywords-component; Association Rules, KDD, Educational Data Mining, Assessment Instruments

1. Introduction

The past several decades have witnessed a rapid growth in the use of data and knowledge mining as a means by which academic institutions extract useful hidden information in the student result repositories in order to improve students' learning processes. Data mining (also called data or knowledge discovery) is the method of analyzing data from different perspectives to discover interesting and helpful information. The information gained through data mining has been effectively used in various sectors ranging from finance, agriculture to health and education. There are many data mining tools [1, 2, 3] available that allow users to analyze data from many different aspects, categorize it, and discover the identified relationships. Technically, data mining is a technique of finding correlations or patterns among many fields in large databases. Educational data mining is fast becoming an interesting research area which allows researcher to extract useful, previously unknown patterns from the educational databases for better understanding, improved educational performance and assessment of the student learning process [4]. It facilities the exploration of unique information from students' result database in academic institutions.

The student's performance in a course is assessed through a variety of assessment instruments i.e. assignments, projects, laboratory work, semester end examinations etc. Some of these assessment instruments are unsupervised such as assignments, homework, and projects for which students are at a liberty to take help from textbooks or reference material or even from their peers or seniors. The assessments in this category are an essential part of learning process and can be regarded as a mean of preparing students for supervised assessments, for example, tests, presentations, oral examinations. The unsupervised assessments are administered under the constant vigilance of a teacher or an examiner with no outside help or assistance for a student undertaking this type of assessment. It is generally expected that students performing well in an unsupervised assessment would also perform well in a supervised assessment. Generally there is no fixed percentage allocated to different assessment instruments as it varies from course to course and from instructor to instructor. In our experience, an instructor might allocate about 20-30% of the total marks for unsupervised assessment instruments whereas the remaining marks are reserved for all other supervised assessments.

From the stand view of the e-learning scholars, data mining techniques have been employed to solve different problems in the educational environment. Some of these applications include students' classification based on their learning performance; detection of irregular learning behaviors; e-learning system navigation and interaction optimization; clustering according to e-learning system usage; systems' adaptability to students' requirements and capacities [5]. The selection of data mining tools and techniques mostly depends on the scope of the problem and the expected results from the analysis. For example, a classification approach is used [6] to classify students to predict their final year performance based on different parameters derived from the data in an educational web-based system. A clustering algorithm is used [7] to categorize students with similar behavioral characteristics. Association rule mining techniques have frequently been used to solve educational problems and carry out critical analysis in an academic environment for improving the learning process of students. These efforts are carried out in order to raise the standards and administration of educational processes by investigating the learning systems, learning resources arrangements, and students' results, curriculum restructuring, and institutional websites [8, 9, 10]. In one study [11], clustering and association rule mining techniques have been applied to students' data to mine the common factors affecting the learners' performance that can be utilized as a base for providing some clues and hints for future learners. In another study [12], students' actions logged during tutor session have been categorized, binned, and symbolized to discover student behavior patterns.

Aforementioned literature review illustrates that different types of investigations have been undertaken on students' assessment data to mine and discover a variety of essential knowledge. However, no study has been carried out to investigate an association between supervised and unsupervised assessment results for discovering possible hidden patterns. The knowledge discovered from such a study would potentially be quite valuable not only for carrying out the course and program assessment activities but also for determining the overall effectiveness of teaching and learning process.

In this paper we investigate this important topic of research and present an analysis of supervised and unsupervised assessment results using association rule mining techniques. The rules meeting the predefined support and confidence are mined to expose the hidden knowledge from the available data of the two types of assessment. These mined rules are analyzed to review the existing assessment processes and recommend constructive actions to academic planners. The results from such analysis have uncovered a number of important facts that are extremely helpful for curriculum planners, developers and academic managers in carrying out a range of essential activities such as assessing and evaluating the course and program curricula, learning methodologies and assessment processes. All of these activities, if done properly, play a pivotal role for the enhancement of students' performance which undoubtedly can be characterized as an ultimate goal of any academic program and the institution. In section 2, we present relevant information about knowledge discovery process along with the data mining and association rule that we have used for the discovery of hidden knowledge. The results of the analysis and the rules discovered from the present study are discussed in section 3. The conclusions of our work are given in section 4.

2. Knowledge Discovery Process

Knowledge Discovery (KD) process is one of the basic concepts of the field of Knowledge Discovery and Data mining (KDD). Figure 1 illustrates the knowledge discovery employed in the present study that we have adapted from Fayad et al. [13]. Solid-line arrows represent various important data processing steps leading towards the knowledge discovery whereas dotted-line arrows show the steps that may form an iterative cycle in order to refine the knowledge discovery process.

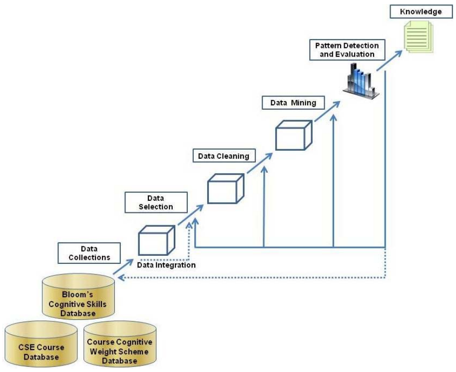

Figure 1. Knowledge Discovery Process

2.1. Selecting Mining Frequent Patterns and Associations

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes' value conditions that occur frequently together in a given dataset [14]. The preliminaries necessary to understand for performing data mining on any data are discussed below.

Let $I = \{I_1, I_2, I_3, ..., I_m\}$ be a set items. Let *D*, the task relevant data, be a set of database transactions where each transaction $T \subseteq I$. Each transaction is an association with an identifier, called transaction identification (TID). Let *A* be a set of items. A transaction *T* is said to contain *A* if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I, B \subset I$, and $A \cap B = \phi$.

Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule $A \Rightarrow B$ means that A and B exist together in 2% of all the transactions under analysis. The rule $A \Rightarrow B$ having confidence of 60% in the transaction set D means that 60% is the percentage of transactions in D containing A that also contains B.

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. If the relative support of an itemset I satisfies a prescribed minimum support threshold, then I is a frequent itemset.

The association rule mining can be viewed as a two-step process:

1) Find all frequent itemsets: Each of these itemsets will occur at least as frequently as a predetermined minimum support count.

2) Generate strong association rules from the frequent itemsets: The rules must satisfy minimum support and confidence. These rules are called strong rules.

2.2. Appriori Algorithm

Apriori is a seminal algorithm proposed by R. Agarwal and R. Srikant [15] in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. The following lines state the steps in generating frequent itemset in Apriori algorithm.

Let *Ck* be a candidate itemset of size *k* and *Lk* as a frequent itemset of size *k*. The main steps of iteration are:

- Find frequent set L_{k-1}
- Join step: C_k is generated by joining L_{k-1} with itself (cartesian product L_{k-1} x L_{k-1})
- Prune step (apriori property): Any $(k - 1)$ size itemset that is not frequent cannot be a subset of a frequent *k* size itemset, hence should be removed
- Frequent set L_k has been achieved

2.3. Task Relevant Data Collection

We analyzed the results data of a three credit hour programming course taught as a part of an engineering degree program at our university. This course had four categories of assessment instruments; assignment (including homework) 20%, laboratory work 20%, class tests and quizzes 20%, and final examination 40%. The assignment falls under the category of unsupervised whereas laboratory work, class tests, and final examination fall under the category of supervised assessment. The students' data from three past course offerings in different semesters was extracted from university's student information system. The score of each student was transformed into transactions (assignment, laboratory work, class test, final examination) where student number will serve as TID.

2.4. Data Preprocessing

The real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results [16]. Therefore, data preprocessing is an important task in data mining. The data we used was in the percentages as discussed above and needed to be transformed to same level of achievement in each assessment. Hence, all scores in different assessment instruments were transformed to a number calculated out of 100. A snapshot of this data, we designate it raw data, is shown in Table 1. Table 1 is the raw data in all the assessments conducted in the course whereas Table 2 is the

TABLE I. RAW ASSESSMENT DATA

A1 and A2 represent marks in two assignments, Q1 and Q2 represent marks in two quizzes, FL is used for final laboratory examination, T1 and T2 represent two one hourly tests, and FE represents final comprehensive examination

A1	A2	Q1	Q2	FL	T1	T2	FE
19	38	5.5	14	11	29	30	61
19	15	7	9	13	23	36	62
19	34	5.5	11	10	23	22	41
0	16	4	1	5	7	12	45
20	34	6.5	10	13	30	27	59
20	32	8	17	13	28	34	87
20	35	6.5	15	11	37	24	78
20	34	8	13	10	22	23	53
19	34	3	10	5	19	16	48
20	40	7.5	12	7	26	33	57
19	34	9	11.5	0	29	21	45
19	34	7	10.5	10	32	21	50
20	30	7	14	5	29	29	60
20	40	5	4	7	0	36	27
0	14	6	11.5	4	6	22	25
0	20	7	1	0	18	16	4
20	30	4	2	3	27	11	28
20	34	2.5	9	6	25	36	30
20	26	9	14	9	35	20	62
20	20	10	16	10	33	40	64
20	36	3	10	6	21	17	31
20	14	3	12.5	6	24	12	30
20	14	5	6	10	29	6	20
20	20	7.5	7	0	14	16	33

TABLE II. PREPROCESSED DATA (STAGE 1)

A% percentage of marks in assignments, Q% percentage of marks in quizzes, FL% percentage of marks in final lab, T% percentage of marks in tests, and FE% percentage of marks in final examination

A%	Q%	FL%	T%	FE%
95.0	65.0	55.0	73.8	61.0
56.7	53.3	65.0	73.8	62.0
88.3	55.0	50.0	56.3	40.5
26.7	16.7	25.0	23.8	44.5
90.0	55.0	65.0	71.3	59.0
86.7	83.3	65.0	77.5	86.5
91.7	71.7	55.0	76.3	78.0
90.0	70.0	50.0	55.6	53.0
88.3	43.3	25.0	43.1	47.5
100.0	65.0	32.5	73.8	57.0
88.3	68.3	0.0	61.9	44.5
88.3	58.3	50.0	65.6	49.5
83.3	70.0	25.0	72.5	59.5
100.0	30.0	35.0	45.0	26.5
23.3	58.3	17.5	35.0	24.5
33.3	26.7	0.0	42.5	4.0
83.3	20.0	15.0	47.5	28.0
90.0	38.3	30.0	76.3	29.5
76.7	76.7	45.0	68.8	62.0
66.7	86.7	50.0	91.3	63.5
93.3	43.3	27.5	47.5	30.5
56.7	51.7	30.0	44.4	30.0
56.7	36.7	47.5	43.1	20.0
66.7	48.3	0.0	37.5	32.5

pre-processed data (Stage-1) calculated out of 100. Symbols A, Q, FL, T, and FE are used to identify assignments, quizzes, laboratory work, test, and final examination respectively.

There are many algorithms available in the literature that are employed to mine association rules implementing the above stated two-step process. In this study we use Apriori algorithm to generate hidden patterns in the students' result data from the two different types of assessment instruments; supervised and unsupervised. These hidden patterns will provide a strong base and valuable information to academic planners and implementers for revising curriculum as well as teaching and assessment methodologies in order to improve student's performance in all types of assessment administered in delivery of a course.

TABLE III. TRANSFORMED ASSESSMENT DATA

(a) Preprocessed data (Stage 2)					(b) Preprocessed data (Stage 3)				
A	Q	FL	T	FE	A	Q	FL	T	FE
A	D	F	C	D	A-A	Q-D	FL-F	T-C	FE-D
F	F	D	C	D	A-F	Q-F	FL-D	T-C	FE-D
B	F	F	F	F	A-B	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
A	F	D	C	F	A-A	Q-F	FL-D	T-C	FE-F
B	B	D	C	B	A-B	Q-B	FL-D	T-C	FE-B
A	C	F	C	C	A-A	Q-C	FL-F	T-C	FE-C
A	C	F	F	F	A-A	Q-C	FL-F	T-F	FE-F
B	F	F	F	F	A-B	Q-F	FL-F	T-F	FE-F
A	D	F	C	F	A-A	Q-D	FL-F	T-C	FE-F
B	D	F	D	F	A-B	Q-D	FL-F	T-D	FE-F
B	F	F	D	F	A-B	Q-F	FL-F	T-D	FE-F
B	C	F	C	F	A-B	Q-C	FL-F	T-C	FE-F
A	F	F	F	F	A-A	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
B	F	F	F	F	A-B	Q-F	FL-F	T-F	FE-F
A	F	F	C	F	A-A	Q-F	FL-F	T-C	FE-F
C	C	F	D	D	A-C	Q-C	FL-F	T-D	FE-D
D	B	F	A	D	A-D	Q-B	FL-F	T-A	FE-D
A	F	F	F	F	A-A	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
F	F	F	F	F	A-F	Q-F	FL-F	T-F	FE-F
D	F	F	F	F	A-D	Q-F	FL-F	T-F	FE-F

2.5. Data Cleaning

It is fundamental truth that incorrect or inconsistent data can lead to false conclusions and hence wrong inferences and decisions. Therefore, high quality data needs to pass a set of quality criteria; accuracy, integrity, completeness, validity, consistency, uniformity, density, and uniqueness. Data cleaning routines attempts to fill in missing values, smooth out noise, and correct inconsistencies in the data. There are a number of data cleaning techniques [16] in the literature such as fill missing values, binning, regression, and clustering. We used the following criteria to clean our data:

- If a student did not sit in the final examination then zero is entered in his score. We removed all such tuples from our result data.
- If a student is absent in one or two assessment instrument then his score was replaced by average of the students score in that assessment.
- If a student is absent in more than two assessment instruments then all such tuples were removed.

2.6. Data Transformation

The result data in each assessment instrument was preprocessed to grades (Stage-2) A(>= 90), B(>= 80), C(>=70), D(>= 60), and F(<60) as shown in Table 3(a). These grades were concatenated with the type of assessment for example an A-A represents A grade in assignment and FE-B represents a B grade in final examination. A sample snapshot from such a transformation is shown in Table 3(b). The final pre-processed form (Stage-3) of assessment transaction, for example, is highlighted in the Table 3(b) by a rectangular box.

3. Results and Rules Analysis

A literature review indicates that most of the authors emphasized on linking students' overall performance [12, 13, 17] to their knowledge mostly for prediction or finding the impact of students failing in one course or the other. This kind of work is not helpful to see the relationship between various assessments conducted within one course and to validate the students' learning process in that course. In this paper, we mined knowledge in the form of association rules, Table 4, to investigate the relationship among various assessment instruments administered in one course.

TABLE IV. ASSOCIATION RULES MINED

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support (a)	Support (c)
1	87.5	A-A=>	FL-F	8	21
2	75	A-A=>	FE-F	8	18
3	100	A-B, FE-F=>	FL-F	6	21
4	100	A-B, FL-F=>	FE-F	6	18
5	85.71	A-B=>	FE-F, FL-F	7	17
6	85.71	A-B=>	FE-F	7	18
7	85.71	A-B=>	FL-F	7	21
8	100	A-F=>	Q-F	6	15
9	84.62	FE-F, FL-F, Q-F=>	T-F	13	12
10	91.67	FE-F, FL-F, T-F=>	Q-F	12	15
11	76.47	FE-F, FL-F=>	Q-F	17	15
12	100	FE-F, Q-F, T-F=>	FL-F	11	21
13	78.57	FE-F, Q-F=>	FL-F, T-F	14	12
14	78.57	FE-F, Q-F=>	T-F	14	12
15	92.86	FE-F, Q-F=>	FL-F	14	21
16	91.67	FE-F, T-F=>	FL-F, Q-F	12	13
17	91.67	FE-F, T-F=>	Q-F	12	15
18	100	FE-F, T-F=>	FL-F	12	21
19	77.78	FE-F=>	Q-F	18	15
20	94.44	FE-F=>	FL-F	18	21
21	100	FL-F, Q-F, T-F=>	FE-F	11	18
22	84.62	FL-F, Q-F=>	FE-F, T-F	13	12
23	84.62	FL-F, Q-F=>	T-F	13	12
24	100	FL-F, Q-F=>	FE-F	13	18
25	91.67	FL-F, T-F=>	FE-F, Q-F	12	14
26	91.67	FL-F, T-F=>	Q-F	12	15
27	100	FL-F, T-F=>	FE-F	12	18
28	80.95	FL-F=>	FE-F	21	18
29	100	Q-F, T-F=>	FE-F, FL-F	11	17
30	100	Q-F, T-F=>	FE-F	11	18
31	100	Q-F, T-F=>	FL-F	11	21
32	93.33	Q-F=>	FE-F	15	18
33	86.67	Q-F=>	FE-F, FL-F	15	17
34	86.67	Q-F=>	FL-F	15	21
35	91.67	T-F=>	FE-F, FL-F, Q-F	12	13
36	91.67	T-F=>	FL-F, Q-F	12	13
37	91.67	T-F=>	FE-F, Q-F	12	14
38	91.67	T-F=>	Q-F	12	15
39	100	T-F=>	FE-F, FL-F	12	17
40	100	T-F=>	FE-F	12	18
41	100	T-F=>	FL-F	12	21

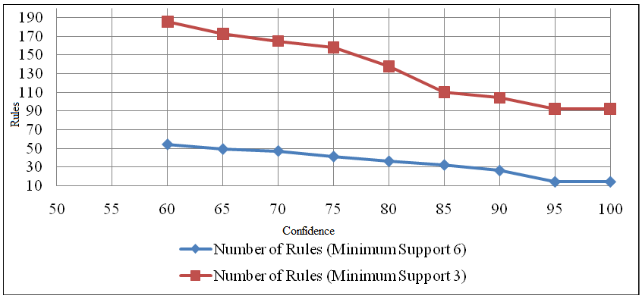

Figure 2. Relationship between support, confidence, and rules

The analysis of our study for a 3-credit engineering degree course taught over three semesters is presented in Table 4. The association rules depicted in Table 4 are mined using a data mining tool XLMiner [2]. This tool allows mining the association rules by setting various minimum support thresholds.

It is observed that by lowering the minimum support threshold there is a marked increase in the number of association rules generated by XLNimer tool. A relationship between minimum support, minimum confidence, and rules generated by the tool are illustrated in Figure 2.

The analysis of the generated rules presented in Table 4 show that rule 1 (support = 8, confidence 87.5%) indicates that students who performed excellent in assignment failed to perform even satisfactorily in the final laboratory examination. The rule 2 (support = 8, confidence 75%) is also strong and extracts knowledge that students who performed excellent in assignment failed in final examination. Similarly, rule 6 (support = 7, confidence 85.71%) and rule 7 (support = 7, confidence 85.71%) show that students who performed very good i.e. scored B grade failed to score similar grades in the final examination and the final laboratory work, respectively. A similar trend is observed in the rules generated with minimum support 3 and minimum confidence greater than 80%. We could not find a single rule with minimum support 3 and 6 and minimum confidence greater that 75% that verifies that students performing excellent in the unsupervised assessment instruments surely performed well in the supervised assessment instruments. The discovered rules are strangely contradictory to the fact that if a student's performance is excellent in the unsupervised assessment (homework or assignments) then he/she must perform better in the supervised assessment instruments such as tests, laboratory works, and/or final examination. This could be due to a variety of reasons; (i) assignments or homework were not developed properly, (ii) there might be an impedance mismatch in the unsupervised and supervised assessments, (iii) the students were not able to apply the knowledge and skills gained through unsupervised assessments in the final laboratory examination and/or class tests or final examination, (iv) the students might have copied the assignments and homework either from the resources available on the Internet or from their friends. In this course, only 20% of the marks were allocated to the unsupervised categories of assessment as compared to 80% allocated for supervised assessment components of the course. This might be another possible explanation for the strange results uncovered from this study.

We believe that the generated association rules are of great help for the curriculum planners and academic mangers. They can certainly use the hidden knowledge and patterns discovered in the present study for redesigning the curriculum and/or changing teaching and assessment methodologies to ensure that the students are fully equipped or capable to perform in a supervised assessment at a level at least equivalent, if not better, than what they perform independently in an unsupervised environment. This would benefit students in improving the quality of their learning experience as well as in enhancing the quality of students' performance by minimizing the impedance mismatch between different types of assessments administered in a particular course.

4. Conclusion

The paper presented the potential use of one of the data mining approaches called association rule mining algorithm in enhancing the quality and experience of students' performances in higher education. The analysis reveals that there are more students who got excellent grades in supervised assessment but failed to attain similar level of performance in the unsupervised assessments. All these and alike hidden patterns could serve as an important feedback for instructors, curriculum planners, academic managers, and other stakeholders in making informed decisions for evaluating and restructuring curricula and associated assessment methodologies with a view to improve students' performance in supervised and unsupervised assessment instruments.

5. Acknowledgements

The authors wish to acknowledge the financial support provided by the Al Ghurair University.

6. References

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.resample.com/xlminer/index.shtml

[3] http://www.knime.org/

[4] A.Y.K. Chan, K.O. Chow, and K.S. Cheung. "Online Course Refinement through Association Rule Mining", Journal of Educational Technology Systems, Volume 36, Number 4/2007 – 2008, pp 433 – 444, 2008.

[5] B. Dogan and A. Y. Camurcu. "Association Rule Mining form an Intelligent Tutor", Journal of Educational Technology Systems, Volume 36, Number 4/2007 – 2008, pp 444 – 447, 2008.

[6] B. Minaei Bidgoli, B.A. Kashy, G. Kortemeyer, and W. F. Punch. “Predicting Students Performance: an application of data mining methods with the educational web-based system LON-CAPA”, in proceedings of ASEE/IEEE Frontier in Education Conference, Boulder, CO: IEEE, 2003.

[7] G.J. Tsai, S.S. Tseng, and C.Y. Lin. “A Two Phase Fuzzy Mining and Learning Algorithm for Adaptive Learning Environment”. In proceedings of the Alexandrov, V.N. et al. (eds.)

[8] R. Damasevicius. “Analysis of Academic Results for Informatics Course Improvement using Association Rule Mining”. Information Systems Development towards a Service Provision Society. ISBN 978-0-387-84810-5 (print), pp 357 – 363, published by Springer US, 2009.

[9] Talavera, L., Gaudioso, E. “Mining Students Data to Characterize Similar Behavior Groups in Unstructured Collaboration Spaces”. In proceedings of the Artificial Intelligence in Computer Supported Collaborative Learning Workshop at the ECAI , Valencia, Spain, 2004.

[10] S. Z. Erdogan, m. Timor. “A Data Mining Application in a Student Database”. Journal of Aeronautics and Space Technologies, Vol. 2, Number 2., pp 53 – 57, 2005.

[11] Anna Katerina Domminguez, Kalina Yasef, and James R. Curran. “Data Mining for Individualized Hints in eLearning”, in proceedings of EDM Educational Data Mining Conference, Pittsburg PA, USA, 2010.

[12] David H. Shanabrook, David G. Cooper, Beverly Park Woolf, and Ivan Arroyo. “Identifying High-Level Student Behavior Using Sequence-based Motif Discovery”, in proceedings of EDM Educational Data Mining Conference, Pittsburg PA, USA, 2010.

[13] Anna Katerina Domminguez, Kalina Yasef, and James R. Curran. “Data Mining for Individualized Hints in eLearning”, in proceedings of EDM Educational Data Mining Conference, Pittsburg PA, USA, 2010.

[14] Jiawei Han and Micheline Kamber. Data mining: concepts and techniques, Morgan Kaufmann, 2006

[15] Fayyad, U., Piatetsky-Shapiro, G. and P. Smyth. “The KDD process for extracting useful knowledge from volumes of data”, CACM 39 (11), pp. 27-34, 1996.

[16] Han, J., Kamber, M., Pei, J., “Data Mining: Concepts and Techniques” The Morgan Kaufmann Series in Data Management Systems, (ISBN: 9781558609013) 2nd Edition, 2005

[17] L. Talavera and E. Gaudioso. “Mining student data to characterize similar behavior groups in unstructured collaboration spaces”, Workshop on Artificial Intelligence in CSCL, 16th European Conference on Artificial Intelligence, ECAI, Valencia, Spain, 2004.