Courseware assessment through data mining techniques

Ioannis Kazanidis[1], Stavros Valsamidis[1], Sotirios Kontogiannis[2], Alexandros Karakos[2]

kazanidis@teikav.edu.gr, svalsam@teikav.edu.gr, skontog@ee.duth.gr, karakos@ee.duth.gr
[1] Kavala Institute of Technology
[2] Democritus University of Thrace

Abstract

Database files and additional log files of Learning Management Systems (LMS) contain an enormous volume of data which usually remain unexploited. A new method is proposed in order to analyze these data both on the level of the courses and the learners. A new architecture based on a 3-level schema for courseware evaluation is proposed. Six measures and three metrics are used and offer useful insights into the courses themselves based on their material and usage by the learners. Two data mining techniques, classification and association rule mining, are applied to the LMS data at the first two levels. Furthermore, regression analysis is applied to the same data at the last two levels. The proposed method was successfully tested to LMS data from a Greek University. The results confirmed the validity of the approach and showed a relationship among the components of the proposed 3-level schema.

Keywords: LMS usage, measures, data mining techniques.

Introduction

E-learning is a modern learning method, based on information and communication technologies (ICT). Its main characteristics are that overcomes time and spatial restrictions since learners can attend the course wherever they are, assuming they have adequate equipment, such as a computer connected to the Internet. The key of e-learning success is online educational content of high quality, appropriate for e-learning and able to fulfill course educational aims and objectives. In order to ensure these elements, it is necessary to apply process of continuous evaluation and optimization of the educational material. Consequently it is necessary to provide feedback to a course author in order to show the means to improve its courseware (Romero et al., 2004). The evaluation of educational material can be made either directly by taking feedback from the learners or through automated data mining techniques applied to courses log files data.

Data mining in education uses computational approaches to analyze educational data in order to analyze upcoming educational issues. According to (Romero & Ventura, 2010) "the term DM is used in a larger sense than the original/traditional DM definition". Although there is a great number of research in the field of DM in e-learning that use typical techniques, such as classification, clustering, association-rule mining, sequential mining etc., there is also a significant number of studies that use techniques belonging to the broader field of DM such as regression, correlation etc.

The approach in this paper is twofold. On one hand it goes backward to examine whether the courses usage by the learners is affected by the educational content exposed by the educators. One the other hand it examines whether the courses usage by the learners in a course is related to the mean performance of the learners in this course. It proposes some new metrics and measures, taking into account several statistics concerning the courses. These include the number of files and their sizes, the number of pages that each course has on the e-learning platform and statistics concerning the usage of the platform for each course by the learners, such as the number of sessions, the number of visits, the duration of each visit. These measures aim to help course authors and/or platform administrators review course usage and find online course weaknesses. Regression analysis is also used for the identification of possible dependencies.

Method

The proposed method adopts a 3-level schema for an e-learning platform. It uses six measures and three metrics for both content and usage measurement. Finally classification, association rule mining and regression analysis are applied to the e-learning data. More specifically the values of the measures and metrics and the mean marks at the corresponding courses are investigated for possible dependencies.

The 3-level dependencies

A view of the proposed schema of the approach is depicted in Figure 1.

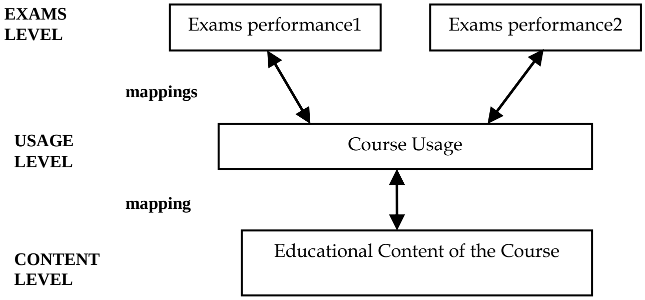


Figure 1. The 3-level approach

The Content Level (CL) includes the educational material that is exposed to the learners by the educators. It can be assessed with the use of measures. The Usage Level (UL) includes the usage of the educational material by the learners. Other measures and new metrics are also used to assess the usage. The Exams Level (EL) includes mean marks, the learners performed for each course. The mapping between CL and UL is one to one. Although the values of the measures may change every time (affected by educators' actions and learners ' usage), the assessment takes place at the end of the semester. The mapping between UL and EL is one to two, since there are two opportunities for every learner to participate at the exams.

Measures and metrics

Some measures are used in the content level and some others in the usage level of the courses. With the measures of the table 1, we quantify the offered educational material to the learners by the educators in terms of input variables in a course.

Table 1. Content Measures

Measure	Description of the Measure
Pages (P)	The total number of (different) pages existing in the course
Files (F)	The total number of files in the course
Size (S)	The total size of the existing files in the course

With the measures of the table 2 (Valsamidis et al., 2012), we quantify the usage of the offered educational material by the learners in terms of usage variables in a course.

Table 2. Usage Measures

Measure	Description of the Measure
Sessions (E)	The total number of sessions per course viewed by all users
Visits (V)	The total number of visits per course by all users
Duration (D)	The duration of (total) visits per course by all users

The next step is to define some metrics in order to qualify these quantity measures. In table 3, we calculate the quality of the offered educational material in a course, in terms of usage by the learners of output variables.

Table 3. Quality Metrics

Metric	Description (Acronym) of the measure
VPS	Visits Per Session
VPD	Visits Per Duration
CUP	Course Utilization and user Perception

$$\text{VPS} = \text{Visits} / \text{Sessions} \quad (V/E) \qquad (1)$$
$$\text{VPD} = \text{Visits} / \text{Duration} \quad (V/D) \qquad (2)$$

These two metrics reflect users' behaviour related to the educational material.

We define a new course quality metric called smooth or distinct Course Utilization and user Perception metric (CUP). This metric expresses how smooth or selective or even randomly visit time of users per course is distributed over the academic semester. That is, for n LMS courses a constant histogram break is used equal to 1/2n (break=10). Then the CUP metric value is calculated as follows:

$$\text{CUP}=n*(\max(h(x))-\min(h(x))) \qquad (3)$$

where h(x) the histogram density estimate:

$$h(x)=k/n\ 1/w \qquad (4)$$

x= cell centered at x with width w that contains k data points and $n*h(x)=(2/n)k$

if CUP->0 then we have smooth course utilization over time while if CUP->1 we have only distinct weeks of course utilization. In cases of CUP>1 we assume that such courses maintain either a one time utilization or abnormal utilization of very low and very high. We consider such courses as flapping or abandoned courses.

Data mining

Data mining techniques have been applied to E-learning systems data by many researchers. Apart from the analytical review by Romero and Ventura (2010), there is some more domain specific. Castro, et al. (2007), among others, deal with the assessment of the students' learning performance, provide course adaptation and learning recommendations based on the students' learning behaviour, deal with the evaluation of learning material and educational web-based courses, provide feedback to both teachers and students of e-learning courses, and detect a typical student's learning behaviour. A survey by Koutri, et al. (2005) provides an overview of the state of the art in research of web usage mining, while discussing the most relevant criteria for deciding on the suitability of these techniques for building an adaptive web site. One relevant study (Kotsiantis et al, 2004) predicts the students' performance as well as to assess the relevance of the attributes involved.

Students are assessed in the final exams of the courses, and they are assigned a mark according to their performance on the course. Having measured the students' activity of the E-learning system according to the measures and metrics, it is possible to investigate whether there is a relationship between student activity in the platform of the E-learning system and the marks of the students in the final exams.

In the classification step, the algorithm 1R (Witten & Frank, 2005) may be applied. It uses the minimum-error attribute for prediction, discretizing numeric attributes (Holte, 1993). The attribute Mark has to be used as class since it describes the education outcome. In this step the attribute/s which best describe the classification will be discovered.

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 (Agarwal et al., 1993). An association rule X → Y expresses a close correlation among items in a database, in which transactions in the database where X occurs, there is a high probability of having Y as well. In an association rule X and Y are called respectively the antecedent and consequent of the rule. The strength of such a rule is measured by values of its support and confidence. The confidence of the rule is the percentage of transactions with antecedent X in the database that also contain the consequent Y. The support of the rule is the percentage of transactions in the database that contain both the antecedent X and the consequent Y in all transactions in the database.

The Weka system has several association rule-discovering algorithms available. The Apriori algorithm will be used for finding association rules over discretized e-Learning platform data.

Regression Analysis

Several regression techniques have been used to predict student's academic performance (using stepwise linear regression) (Golding & Donalson, 2006), to identify variables that could predict success in colleges courses (using multiple regression), to predict university students' satisfaction (using regression and decision trees analysis), to predict high school students' probabilities of success in university (Mcdonald, 2004), to predict a student's test score (using stepwise regression) (Feng et al., 2005) and to predict the probability a student has of giving the correct answer to a problem in an ITS (using a robust ridge regression regression algorithm) (Cetintas et al., 2009).

Linear regression analysis is applied on the level of the courses. The classifier builds linear logistic regression models. In our case, the way the mean mark is affected by the metrics VPS, CUP and VPD of each course is examined. The regression coefficients show the marginal value of input (VPS, CUP and VPD) required measuring Mark.

Results

Study population and context

The recording of specific data from the e-Learning platform is the first step. The dataset was collected at from the Open eClass e-learning platform (GUNet, 2012) which is used in Kavala 's Technological Education Institute (TEI),. The data are from the spring semester of 2011 and involve 1534 students and 34 different courses and are obtained from the server log file.

Case study

A view of the collected data is shown in table 4.

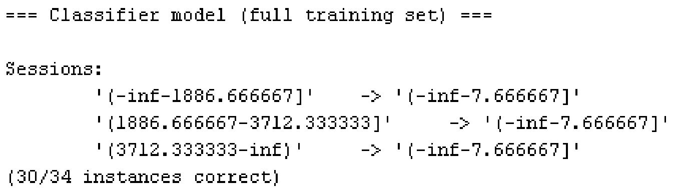
Table 4. Tracked data, measures, metrics and marks

CID	Sessions	Visits	Duration	Pages	Files	Size	VPS	VPD	CUP	Mark
AD5104	480	1567	2891	6	785	785	3,26	0,542	0,213	5,78
AD5103	477	1853	3635	6	3165	3165	3,89	0,510	0,16	6,02
AD2104	61	93	116	5	2065	2065	1,51	0,802	2,844	4,42
AD2103	3237	9756	10864	6	5135	5135	3,01	0,898	0,0462	7,26
AD6102	3734	6585	6571	6	5696	5696	1,76	1,002	0,0977	6,2
AD6114	337	938	1688	6	0	0	2,79	0,556	0,2844	6,48
AD6106	144	271	709	6	0	0	1,89	0,382	0,888	4,49
AD6105	910	2340	3586	6	7198	7198	2,57	0,653	0,1244	5,88
AD7107	1115	4334	6778	6	42277	42277	3,89	0,639	0,0799	6,3
AD4108	378	1627	3185	6	100	100	4,30	0,511	0,1777	5,56
AD7105	728	2184	3760	6	2471	2471	3,00	0,581	0,1244	5,89
AD4101	709	1501	3284	6	250	250	2,12	0,457	0,1955	5,65
AD5102	539	921	1912	6	0	0	1,71	0,482	0,355	6,23
AD5101	414	746	1731	6	0	0	1,80	0,431	0,444	6,11
AD6112	256	593	1321	6	410	410	2,31	0,449	1,4222	6,08
AD6111	383	1352	2719	6	1355	1355	3,53	0,497	0,071	6,34
AD2100	2063	8430	15915	6	358	358	4,09	0,530	0,0319	7,67
AD3107	632	1713	3662	6	6899	6899	2,71	0,468	0,1777	6,66
AD5106	308	984	1994	6	121	121	3,20	0,493	0,3555	6,31
AD5105	269	799	1750	6	0	0	2,97	0,457	0,2844	6,09
AD7101	791	3206	5706	6	27	27	4,06	0,562	0,1155	6,05
AD7100	415	1677	2724	6	0	0	4,04	0,616	0,2488	6,28
AD6108	2209	4565	7633	9	12461	12461	2,07	0,598	0,0888	6,71
AD6107	970	2088	3538	10	11525	11525	2,15	0,590	0,106	6,53
AD2106	4793	10091	14551	12	5943	5943	2,11	0,693	0,0355	7,23
AD2105	5538	11832	16780	13	52318	52318	2,14	0,705	0,0266	7,81
AD2107	3726	10113	13824	6	2206	2206	2,71	0,732	0,0248	7,19
AD6100	2697	5271	8199	6	29290	29290	1,95	0,643	0,053	6,67
AD7102	3721	7780	8846	6	61213	61213	2,09	0,879	0,044	6,82
AD3106	706	2533	5115	6	108	108	3,59	0,495	0,106	5,98
AD3108	2759	4139	5330	6	4175	4175	1,50	0,777	0,106	6,02
AD4100	616	1401	2564	6	0	0	2,27	0,546	0,213	6,22
AD3102	490	621	1515	6	0	0	1,27	0,410	0,3556	4,45
AD4104	252	390	1147	6	0	0	1,55	0,340	0,444	4,210

The values of the measures of Tables 1 and 2, which express measures of Usage Level (UL) and Content Level (CL), are presented. The aforementioned measures contribute to the evaluation of courses content and usage.

1st mapping dependencies

All the DM techniques were performed using the open source Weka. The data mining methods are applied to the measures of the table 4. The results of the classification based on the OneR algorithm show that the measure Pages is better classified (described) by the measure Sessions as it is depicted in Figure 2.



```
=== Classifier model (full training set) ===

Sessions:
        '(-inf-1886.666667]'    -> '(-inf-7.666667]'
        '(1886.666667-3712.333333]'     -> '(-inf-7.666667]'
        '(3712.333333-inf)'     -> '(-inf-7.666667]'
(30/34 instances correct)
```

Figure 2. Classification results

The results of the association rule mining based on the Apriori algorithm (Agrawal & Srikant, 1994) show 10 rules, as it is depicted in Table 5.

Table 5. Apriori algorithm based on confidence metric

Best rules found			
1.	Visits='(-inf-4006]' 23	==>	Sessions='(-inf-1886.666667]' 23 conf:(1)
2.	Visits='(-inf-4006]' Duration='(-inf-5670.666667]' 22	==>	Sessions='(-inf-1886.666667]' 22 conf:(1)
3.	Sessions='(-inf-1886.666667]' Duration='(-inf-5670.666667]' 22	==>	Visits='(-inf-4006]' 22 conf:(1)
4.	Visits='(-inf-4006]' Pages='(-inf-7.666667]' 22	==>	Sessions='(-inf-1886.666667]' 22 conf:(1)
5.	Sessions='(-inf-1886.666667]' 24	==>	Visits='(-inf-4006]' 23 conf:(0.96)
6.	Sessions='(-inf-1886.666667]' 24	==>	Pages='(-inf-7.666667]' 23 conf:(0.96)
7.	Duration='(-inf-5670.666667]' 23	==>	Sessions='(-inf-1886.666667]' 22 conf:(0.96)
8.	Duration='(-inf-5670.666667]' 23	==>	Visits='(-inf-4006]' 22 conf:(0.96)
9.	Visits='(-inf-4006]' 23	==>	Duration='(-inf-5670.666667]' 22 conf:(0.96)
10.	Visits='(-inf-4006]' 23	==>	Pages='(-inf-7.666667]' 22 conf:(0.96)

Table 5 shows how a large number of association rules can be discovered. There are some uninteresting rules such as rules 8 and 9, since there is obvious dependency between Duration and Visits. There are also redundant rules, rules with a generalization of relationships of several rules, like rule 2 with rules 1 and 7, 3 with rules 5 and 9 and 4 with rules 6 and 10. There are some similar rules, rules with the same element in antecedent and consequent but interchanged, such as rules 1, 2, 8 and rules 5, 3, 9 respectively. But there are also rules that show relevant information for educational purposes, like those that show conforming relationships such as rules 1, 2, 3, 5 and 7. And there are also rules that show interesting relationships such as rules 4, 6 and 10, which can be very useful for the educator in decision making about the activities of their courses.

2nd mapping dependencies

Simple linear regression analysis was applied to the metrics and marks of Table 4. The courses were classified to 3 classes according to their mean marks.

The first class which corresponds to courses with low marks is described by the equation

$$\text{Mark}= -1.46 + \text{VPS}*0.9+\text{CUP}* 2.73 \qquad (1)$$

The second class which corresponds to courses with mid marks is described by the equation

$$\text{Mark}=1.21+ \text{VPS}*0.55 \qquad (2)$$

The third class which corresponds to courses with high marks is described by the equation

$$\text{Mark}=0.27+\text{VPS}*0.61-0.981.21*\text{VPD} \qquad (3)$$

An increase in the VPS metric leads to higher mean marks for each course. Also increase in the VPS metric leads to higher mean marks for each course.

Discussion and Conclusions

This study proposes an approach for discovering dependencies based on histories e-learning data. It tackles the problem of analysing these data in three levels. Initially it examines whether the courses usage by the learners is affected by the educational content exposed by the educators. Then, it examines whether the courses usage by the learners in a course is related to the mean performance of the learners in this course. It proposes some measures such as the number of files and their sizes, the number of pages that each course has on the e-learning platform, the number of sessions, the number of visits, and the duration of each visit. New metrics are also proposed to assess the courses usage. Two data mining techniques, classification and association rule mining, were applied to the e-Learning data at the first two levels. Furthermore, regression analysis was applied to the same data at the last two levels.

Its originality lies in the different use of existing techniques. It builds on existing work, but also extends it in a different way for the e-learning field. It has the following advantages: (1) It is independent of a specific e-Learning platform, since it is based on the Apache log files and not the e-Learning platform itself. Thus, it can be easily implemented for every e-Learning platform. (2) It uses measures and metrics in order to facilitate the evaluation of each course in the e-Learning platform and the instructors to make proper adjustments to their course educational material. (3) It uses classification, association rule mining and regression analysis in order discover possible dependencies of the e-Learning data.

The results disclose dependencies between content of the course and its usage by the learners. There is dependency of the number of modules (Pages) of the platform with the number of sessions and the number of visits. The results also confirm the assertion that there is dependency between the students' usage in an e-Learning platform with their corresponding performance in the exams.

The fact that only 34 courses in one platform were investigated is a limitation to the study. Especially for the data mining techniques which demand large datasets. However, this was ineluctable since the case study department implements this number of online courses. But the proposed approach seems to be quite reliable if the inspection takes place over a long time period.

The results of this research are remarkable from pedagogical point of view. On the one hand this approach contributes to the improvement of courseware content quality since the proposed measures and metrics and their correlations to students' marks provide to the authors a feedback about their courses efficiency. Improvement of course quality provides to students the opportunity of asynchronous study of courses with actualized and optimal educational material. On the other hand, since students usage results are correlated with students' grade, online platform may provide educators with notifications about students' online actions. For example a learning platform could record student actions and after the application of specific algorithms to classify them into predefined groups according their estimated performance. Educators could study these groups of students and try to help and motivate weak students. They could provide with more educational content to the advanced students in order to achieve a more depth learning. Thus, learning performance of all students could be improved.

References

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of 20th International Conference on Very Large Data Bases* (pp. 487-499).

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). Applying data mining techniques to e-learning problems, in L. C. Jain, R. Tedman, and D. Tedman, Eds. *Evolution of Teaching and Learning Paradigms in Intelligent Environment (Studies in Computational Intelligence)*, 62, (pp. 183–221). New York: Springer-Verlag.

Cetintas A.,,Si, L., Xin, Y. P., & Hord, C. (2009). Predicting correctness of problem solving from low-level log data in intelligent tutoring systems, *Proceedings of International Conference on Educational Data Mining*, (pp. 230-238). Cordoba, Spain.

Feng, M. Heffernan, N., & Koedinger, K. (2005). Looking for sources of error in predicting student's knowledge, *Proceedings of AAAI Workshop on Educational Data Mining* (pp. 1–8).

Golding, P. & Donalson, O. (2006). Predicting academic performance. *Proceedings of Frontiers Educational Conference* (pp. 21–26). San Diego, CA.

GUNet. (2012). OPEN eClass, Retrieved January 30, 2012 from http://eclass.gunet.gr/.

Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*. 11, 63-91.

Kotsiantis S., Pierrakeas C., & Pintelas P. (2004). Predicting Students' Performance in Distance Learning Using Machine Learning Techniques, *Applied Artificial Intelligence*, 18(5), 411 – 426.

Koutri, M., Avouris, N., & Daskalaki, S. (2005). A survey on web usage mining techniques for web-based adaptive hypermedia systems. In S. Y. Chen & G. D. Magoulas (Eds.), *Adaptable and adaptive hypermedia systems* (pp. 125–149). IRM Press.

Mcdonald, B. (2004). *Predicting student success, Journal for Mathematics Teaching Learning*, 1, 1–14.

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art, *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 40(6), 601-618.

Romero, C., Ventura, S., & De Bra, P. (2004). Knowledge discovery with genetic programming for providing feedback to courseware author, *User Model. User-Adapted* Interaction: J. Personalization Res., 14 (5), 425-464.

Valsamidis, S., Kontogiannis, S., Kazanidis, I., Theodosiou, T., Karakos, A. (2012). A clustering methodology of web log data for Learning Management Systems, *Journal of Educational Technology and Society*, 15(2), 154-167.

Witten, I., & Frank, E. (2005*). Data Mining Practical Machine Learning Tools and Techniques*, San Francisco: Morgan Kaufmann.