# Using learning analytics to identify successful learners in a blended learning course

**Article** *in* International Journal of Technology Enhanced Learning · December 2013

DOI: 10.1504/IJTEL.2013.059088



**4 authors**, including:

Sotiris Kotsiantis
University of Patras
**198** PUBLICATIONS **6,602** CITATIONS

SEE PROFILE

Nikolaos Tselios
University of Patras
**106** PUBLICATIONS **1,061** CITATIONS

SEE PROFILE

Andromahi Filippidi
University of Patras
**5** PUBLICATIONS **35** CITATIONS

SEE PROFILE

*Int. J. Technology Enhanced Learning, Vol. X, No. Y, XXXX*

# Using learning analytics to identify successful learners in a blended learning course

Sotiris Kotsiantis

Department of Mathematics,
University of Patras,
26500 Rio, Patras, Greece
Email: kotsiantis@upatras.gr

Nikolaos Tselios*, Andromahi Filippidi and Vassilis Komis

Department of Educational Sciences and Early Childhood Education,
ICT in Education Group,
University of Patras,
26500 Rio, Patras, Greece
Email: nitse@ece.upatras.gr
Email: afilippidi@upatras.gr
Email: komis@upatras.gr
*Corresponding author

**Abstract:** In this paper, students' practices while using a Learning Content Management System in a blended learning environment were examined. This is a case study involving 337 students who attended an academic course based upon a blended learning approach over three years using Moodle. Eighteen variables depicting the students' perceptions of Moodle, as well as their interaction with it, were examined using four complementary data mining and statistical analysis approaches: visualisation, decision trees, class association rules and clustering. The analysis of the collected data shows that failure in the course was associated with negative attitudes and perceptions of the students towards Moodle. On the other hand excellent grades were associated with increased use of the LCMS. Requirements elicitation of a learning analytics dashboard, are also discussed.

**Keywords:** learning analytics; blended learning; learning content management systems; case study; interaction data; perceptions; higher education; Moodle.

**Reference** to this paper should be made as follows: Kotsiantis, S., Tselios, N., Filippidi, A. and Komis, V. (XXXX) 'Using learning analytics to identify successful learners in a blended learning course', *Int. J. Technology Enhanced Learning*, Vol. X, No. Y, pp.xxx–xxx.

**Biographical notes:** Sotiris Kotsiantis is a Lecturer at the Department of Mathematics in the University of Patras, Greece. He is a mathematician and holds an MSc in computational mathematics and a PhD in machine learning from the University of Patras, Greece. His research interests are in the field of data mining, machine learning and computational intelligence.

Nikolaos Tselios is an Assistant Professor at the Department of Educational Sciences and Early Childhood Education in the University of Patras. He holds a PhD in Usability Engineering and a Diploma from the Electrical and Computer Engineering Department, University of Patras, Greece. His main research interests are educational technology, human computer interaction, e-learning, technology acceptance.

Andromahi Filippidi is a PhD Student at the Department of Educational Sciences and Early Childhood Education in the University of Patras and a Kindergarten Teacher.

Vassilis Komis is a Professor at the Department of Educational Sciences and Early Childhood Education in the University of Patras. He holds a degree in Mathematics from the University of Crete, DEA and doctoral degrees in Teaching of Computer Science (Didactique de l'Informatique) from the University of Paris 7 - Denis Diderot (Jussieu). His research interests are conception and development of digital educational environments, implementation of collaborative learning systems, didactics of computer science, teachers' training and the integration of ICT applications in education.

## 1 Introduction

Continuous web technology growth and adoption led to a variety of learning alternatives which go beyond the traditional classroom setting. Nowadays, usage of e-learning platforms such as Learning Content Management Systems (LCMS) is heavily observed, especially in the context of tertiary education (Means et al., 2010). Usually such a platform comprises digital resources related to the course in the form of presentations, pdf files, interactive demos, video lectures, etc. These resources are organised in a chronological or thematic manner. Various communication tools such as forum, chat and announcements are also provided thus enhancing the discussion opportunities and empowering a sense of community.

With the continuous improvement and maturation of available e-learning platforms, the focus should be transitioned from the technical details to the actual needs, both in terms of study settings and context but also in terms of learning scope, targets and aims. Also, control and responsibility of the learning process should be gradually shifted from the educators to the learners. Throughout this process other needs, such as ease of use and quality of feedback provided to the students as well as to the educators, will also arise which will inevitably dictate the selection of a suitable e-learning platform (Tselios et al., 2008a; Tselios et al., 2008b; Tselios et al., 2011a; Tselios et al., 2011b).

As a consequence of LCMS use throughout the educational procedure, a substantial amount of user-related data are provided and collected. However, due to the lack of established methods and tools to inspect the students' learning process, frequently monitoring and evaluation of e-learning becomes a cumbersome task. Such a volume of data can be interpreted and inform both learning design and evaluation using a variety of methods such as data mining, expert rules and predictive modelling. Since students are the key entity under consideration across all phases of an LCMS (Mandinach, 2005) rich data could be collected providing insight about their activity and informing educators about effective types of intervention to enrich the learning process.

Although LCMS platforms usually provide some functionality to log students' activity, often the abundance of the obtained data presented in a raw format provides little insight about how it might be used. Moreover, since many factors contribute to the success of e-learning systems in general and LCMS in particular, attempts to explain usage behaviour without taking into account these factors could lead to superficial explanations. In addition, the high number of participants in a typical course mediated by LCMS poses new unforeseen challenges in terms of personalised feedback and effective tutor support.

Emerging disciplines such as Educational Data Mining (EDM) and Learning Analytics (LA) (Long and Siemens, 2011; Ferguson, 2012; Siemens and Baker, 2012) focus on developing robust and holistic methods to explore such types of data. Typically, such an approach takes into account both data inherent attributes such as time and sequence of students' actions as well as some contextual factors (type of lesson, instructor's characteristics, students' attitude towards e-learning, etc.) to better understand students. Although there is clearly some overlap between the goals of the aforementioned approaches, at the same time EDM and LA have some key differences (Long and Siemens, 2011; Siemens and Baker, 2012): EDM is primarily relied upon techniques such as classification, clustering relationship mining and visualisation in order to analyse individual components and the relationship between them (Siemens and Baker, 2012). On the contrary, LA uses social network analysis, sentiment analysis, discourse analysis to build holistic sense making models (Lockyer and Dawson, 2011), attempting "to understand systems as wholes in their full complexity" (Siemens and Baker, 2012).

Nevertheless, the importance of such approaches is very high if not self-evident. Close monitoring of students' learning is an essential part of quality education, since it provides the means both for personalised feedback and support as well as for curriculum improvement. Moreover, evidence-based analysis of students' performance using log files allows informed decisions about possible problems and suitable ways to address them effectively. Not surprisingly, during the last years a variety of approaches applied to identify various issues of technology enhanced learning is observed. A brief overview is presented in the following section.

## 2 Related work

In recent years, researchers have begun to investigate various data mining methods in order to help teachers improve the efficiency of e-learning systems (Baker and Yacef, 2009; Romero and Ventura, 2010). Romero et al. (2008) demonstrated the usefulness of the data mining application in course management systems. Instructors can use visualisation techniques to find some groups of students which share common characteristics and then they can apply clustering techniques in order to obtain the exact group boundaries. These groups can also be used to create a classifier in order to classify new students. The instructors can also use association rule mining techniques to find if there is any relation between these students' characteristics and other attributes.

Castro et al. (2007) also provide a taxonomy of e-learning problems to which data mining techniques have been applied, such as: Students' classification based on their learning performance; detection of irregular learning behaviours; e-learning system navigation; clustering according to similar e-learning system usage and systems'

adaptability to students' requirements. Jovanovica et al. (2012) applied classification models for prediction of students' performance, and cluster models for grouping students based on their cognitive styles in an e-learning environment.

Log analysis (Psaromiligkos et al., 2011) and visualisation for the analysis of students' behaviour have been one of the main research topics in the research communities of Artificial Intelligence in Education (AIED) and EDM. The outcome is of significant usefulness (Jong et al., 2007), especially in the domain of generating student-centred feedback by leveraging user tracking data from learning systems (Dominguez et al., 2010).

Online collaborative learning is a widespread educational form. In this context, the goal of Perera et al. (2009) was to allow the groups and their facilitators to inspect relevant aspects of the group's operation and provide feedback if these are more probably linked with positive or negative outcomes as well as to indicate where the problems are. Garcia et al. (2011) describe a collaborative educational data mining tool based on association rule mining for the ongoing improvement of e-learning courses and allowing tutors with similar course profiles to share and score the discovered information.

Romero et al. (2013a, 2013b) demonstrate how web usage mining can be applied in e-learning systems in order to accurately predict the student grades in the final exam of a course. Moreover, analysis of data from a Blackboard Vista-supported course identified 15 variables demonstrating a simple but significant correlation with student final grade (Macfadyen and Dawson, 2010).

Delavaria et al. (2008) present the capabilities of data mining in the context of higher education by proposing guidelines for institutions to improve their decision processes. Baepler and Murdoch (2010) link the concepts of academic analytics, data mining in higher education and course management system audits and suggest how these techniques and the data they produce could be useful. Ali et al. (2012) analysed the results of two qualitative studies to evaluate the two versions of LOCO-Analyst, a learning analytics tool. The results showed that multiple ways of visualising data increase the value of the feedback types.

As derived by the discussion above, the sheer amount of data collected by a LCMS combined with learning analytics approaches could efficiently inform the educational process, in a formative as well as a summative manner. Teachers could study such metrics while the course is delivered to efficiently engage isolated students. In addition, after course completion, learning analytics can be perceived as an evidence-based approach to inform course improvement and redesign. However, despite the already published promising efforts, the effect of students' attitudes towards e-learning often are not taken into account, thus possibly leading to misinterpretations and superficial results.

Towards this end, a longitudinal study involving design and delivery of a blended learning compulsory course in the context of higher education was realised. The experimental design, the data obtained and the analysis using appropriate data mining and statistical techniques are discussed in the following.

## 3 Method

### *3.1 Research goal*

The goal of this study was threefold. Firstly, to study whether, and to what extent, the students' learning performance is related with their activity in a LCMS, namely Moodle,

which mediated a blended learning academic course (Castro et al., 2007). Secondly, to identify what types of collected data are the best predictors. To this end, both interaction data as well students' perceptions about the usefulness of the adopted approach and the Moodle ease of use were used. Thirdly, if the knowledge obtained while in the process of investigating the effectiveness of various data mining and statistical analysis methods in the context of students' performance prediction could inform the design of a Learning Analytics dashboard, a system to provide real time feedback both to the students as well as to the educators.

### *3.2 Procedure*

The compulsory course entitled 'Information and Communication Technologies in Education' was held in the spring semester of 2007–2008, 2008–2009 and 2009–2010. 337 students, eight male, 329 female, aged 18–35 (mean = 20.44, sd = 3.06) attended the course over the three years (2007–2008: 127 students, three male, 124 female, 2008–2009: 93 students, three male, 90 female, 2009–2010: 117 students, two male, 115 male). One student abandoned the course, thus she was excluded from the analysis. The students attended a two hour compulsory laboratory session for 11 consecutive weeks. Each session dealt with a particular topic, directly related to the course goals. For each laboratory session, except the first two introductory sessions, the students had to deliver a personal report related to a given problem based assignment. During each lab, the tutors provided information about the topic and the goals of the session and subsequently explained each assignment given to the students. The materials provided to the students were organised according to each subject and were available to the students until the end of the semester.

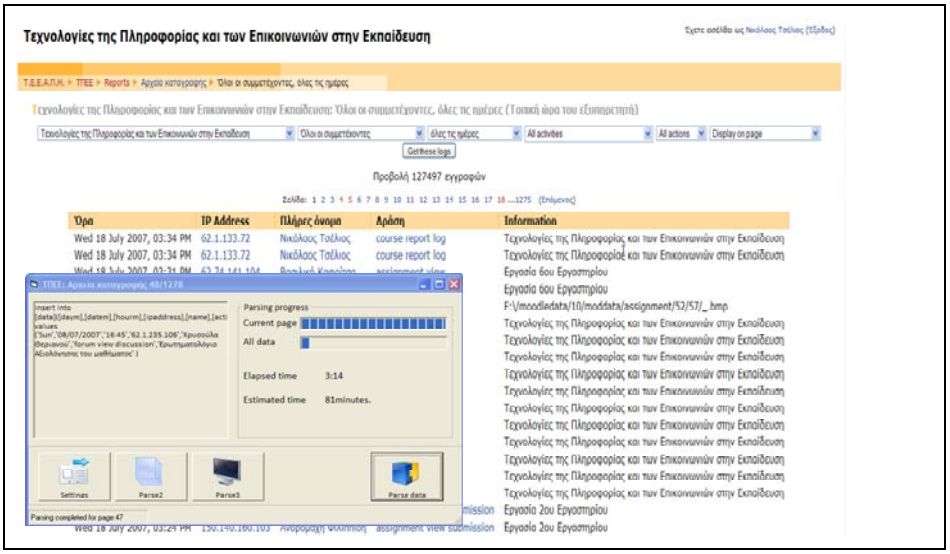
Design and delivery of the course was based on a social constructivist pedagogical framework. A number of blended learning principles referring to the goals and the context of the course by Garrison and Kanuka (2004) were adopted. In addition, face to face learning methods and online technologies (Moodle) were utilised. The adopted pedagogical model was based on the problem based learning approach proposed by Duffy and Kirkley (2004).

### *3.3 Data collection*

The procedure used for the data collection and the analysis comprised 4 phases (Romero et al., 2008): (a) data collection; (b) preliminary data pre-processing using appropriate algorithms with respect to the research questions; (c) application of the method (implementation of algorithms for carrying out the results) and (d) interpretation of results and conclusions.

A sheer amount of learners' low level interaction log files were extracted from the Moodle for this goal, combined with each student's grade. The process of data collection and pre-processing of the learners' log files was carried out using a tool entitled Moodle Parser (MP, Figure 1) developed by one of the authors. In specific, MP automatically parses the data reported by Moodle related to participants' interaction and stores them into an appropriately designed relational database.

**Figure 1** Screenshot of Moodle Parser while accessing, processing and indexing Moodle interaction data into a relational database (see online version for colours)




More than 250.000 log events were parsed, indexed and stored. Subsequently, appropriate SQL queries were applied to identify the total actions carried out by each student. Finally, the obtained pre-processed data were inputted in R and Weka for further analysis. Other useful tools realised to mine Moodle data are nowadays present (Pedraza et al., 2011). However, the predefined database schema adopted to accelerate data analysis and log overview, greatly facilitates the pre-processing stage. In subsequent versions of MP, a visualisation module will be implemented, to enable various visualisations of students' interaction.

### *3.4 Data analysis*

Data related to 19 different variables were collected and they are presented in Table 1. The first two were related to whether the student owed a computer and internet connection at home. We were not able to examine additional demographic values due to the similar characteristics of the participants: 329/337 were females and the actual differences in their age was negligible. Variables 3–7 were related to the students' perceptions about the Moodle LMS and their opinions about its educational value and usefulness. Their opinions were collected after lab completion (and prior to the course final examination) using an appropriate questionnaire delivered to the students via the surveymonkey service. Variables 8–18 were related to the students' activity. All types of data captured by Moodle were used in order to examine the relative utility of each attribute towards predicting students' performance. Variable Final Note indicates student's grade. Our goal was to examine not only variables illustrating interaction with the system, but to combine them with the students' perceptions of Moodle usefulness and usability.

**Table 1** Variables extracted from students' Moodle use

| *Variable* | *Type* | *Description* |
|---|---|---|
| Computer_at_home | Nominal | Whether the student owes a computer |
| Internet_at_home | Nominal | Whether the student has internet connectivity at home |
| Computer_use_per_week | Ordinal (1–5) | Frequency of student's weekly computer use |
| Ease_of_Moodle_use_ perceptions | Interval (1–5) | Students' perception of Moodle usability |
| Moodle_use_capability_ perceptions | Interval (1–5) | How capable did the students thought they were, while using Moodle during the semester |
| Attitude_about_Moodle | Interval (1–5) | Students attitudes towards Moodle |
| Perceived_Moodle_Usefulness_ lesson | Interval (1–5) | Students' perception about the usefuleness of the material delivered via Moodle |
| Perceived_Usefulness_ assignment | Interval (1–5) | Students' perception about the usefulness of the assignments were given via Moodle |
| Total_Of_ id | Ratio (0–X) | Total number of actions per student |
| assignment_view | Ratio (0–X) | Number of actions in the Moodle assignment section |
| course_view | Ratio (0–X) | How many times the student accessed the description and the basic material of each weekly laboratory session |
| forum_add_post | Ratio (0–X) | Number of posts on the forum per student |
| forum_view | Ratio (0–X) | How many times the student accessed the forum section |
| glossary_view | Ratio (0–X) | How many times the student accessed the glossary section |
| questionnaire_view | Ratio (0–X) | How many times the student accessed the questionnaire section which contained assessment rubrics referring to educational software |
| resource_view | Ratio (0–X) | How many times the student accessed the service which contained supplementary material and additional learning resources |
| user_view | Ratio (0–X) | How many times the student accessed the service which contained each user's overview profile |
| user_view_all | Ratio (0–X) | How many times the student accessed the service which contained all users' overview profiles |
| Final note | Ratio (0–10) | Student's final course grade |

## 4 Results

Four different methods were applied to explore the data illustrating students' activity in the Moodle environment: (a) visualisation of each variable distribution using R version 3.0.1 (Graham, 2011); (b) C4.5 decision tree algorithm (Quinlan, 1993) to identify which variables predict students' pass or fail (whether they passed the lesson or not); (c) class association rules (Bing et al., 1998) indicating which variables were associated with their grade and (d) clustering using *k*-means implementation of Weka version 3.7.8 (Hall et al., 2009). Methods (b) and (c) due to their comprehensibility are considered suitable for decision making and quick inspection. In the following, each of the aforementioned methods is presented and discussed extensively.

### *4.1 Visualisation of each variable distribution in relation to the obtained final note*

In Figure 2, variable values in relation to the obtained final note are graphically presented using R, for a total of 18 charts. Students were graded from 0 to 10. Grade which at least equals to 5 indicates a passing grade. Grades from 7 to 8.5 and higher than 8.5 indicate a very good and an excellent performance, respectively. Such a visual representation greatly facilitates tutors to make sense of data and should be incorporated in a Learning Analytics Dashboard. Moreover, students with excellent grades browsed the glossary section the most (Figure 2, seventh row, second graph: glossary_view). However, such data eyeballing approaches rarely can lead to conclusive results. As one may observe in Figure 2, few variables are characterised by extreme differences, an ascertainment which stress the complex and interrelated phenomena in such a blended learning course.

**Figure 2** The variable values in relation to the students' grade (Final Note)

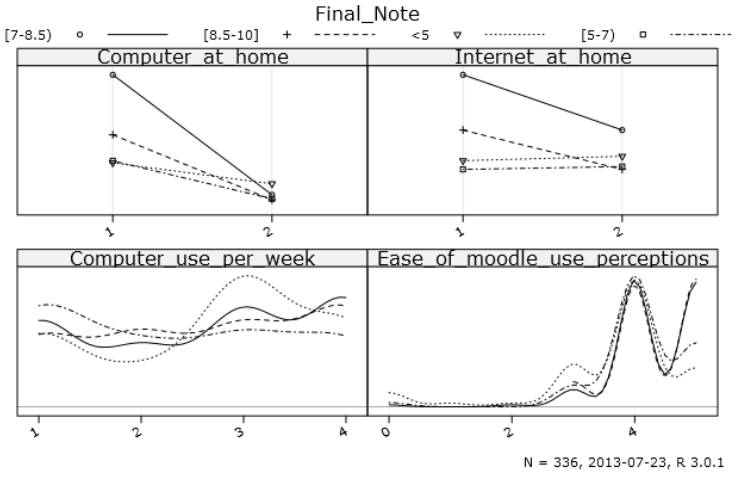
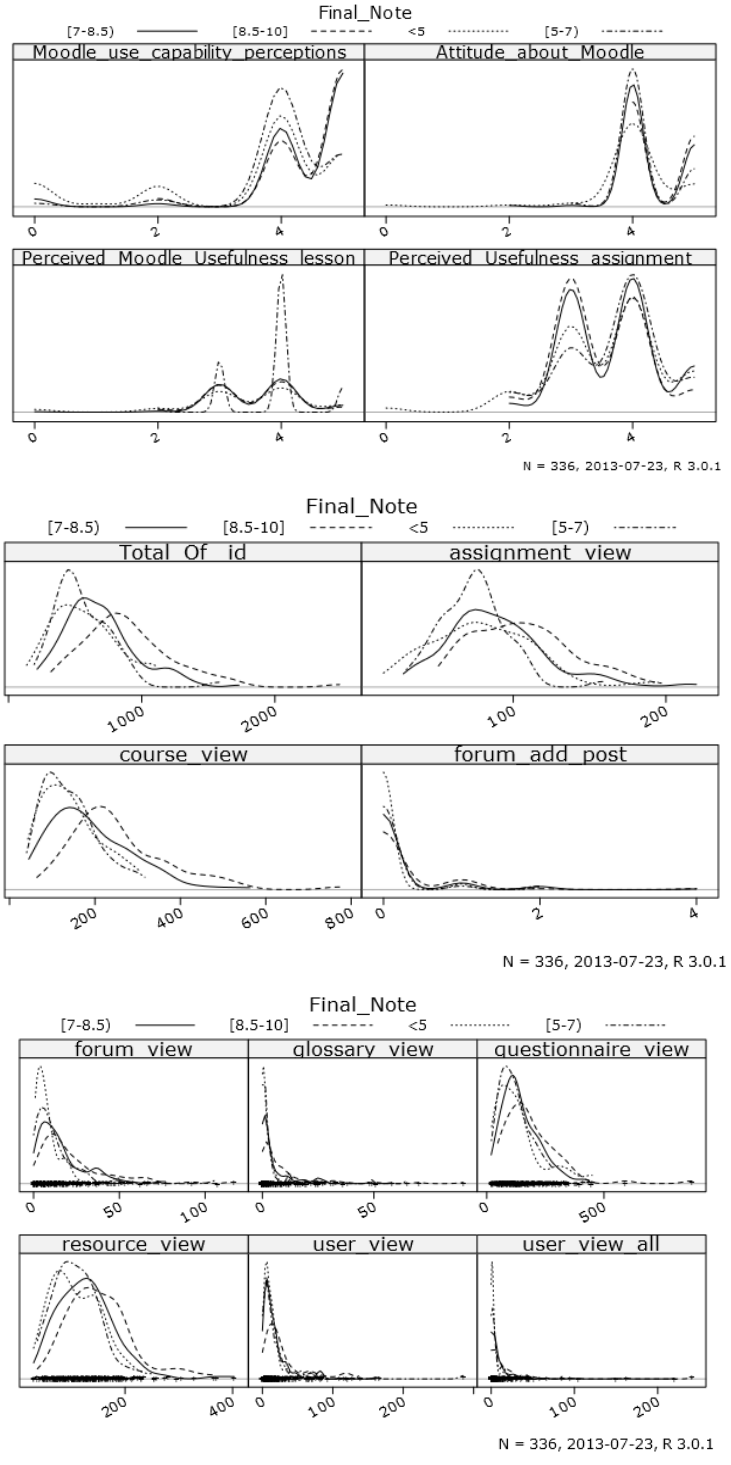
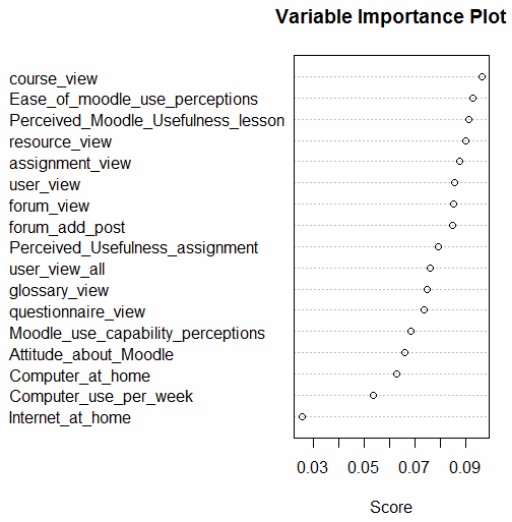


**Figure 2** The variable values in relation to the students' grade (Final Note) (continued)



Such an approach could be substantially enhanced by using a variable importance plot (Figure 3). The varplot command in R produces such plot using the improve criteria reported by Hastie et al. (2001). This is a rather standard measure for determining variable importance. In our case, course and resource view as well as students' perceptions related to Moodle usefulness and ease of use were found to be of significant importance. On the contrary, the variable depicting whether a student had an internet connection at home found to be of negligible importance.

**Figure 3** Variable importance plot




We removed from the training set the variables with low score while in the attempt to realise a simpler and easier to understand decision tree, which is described in the following section.

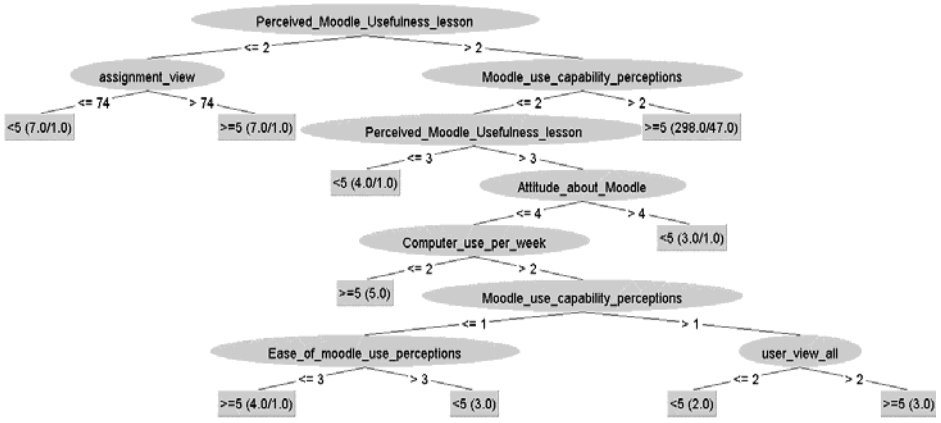
## *4.2 Decision tree to identify association between students' activity and passing grade*

Decision trees (Quinlan, 1993), are trees that classify examples by sorting them based on features values. Each node in a decision tree represents a feature in an example to be classified, and each branch represents a value that the node could have. One of the most useful characteristics of decision trees is their comprehensibility. An educator can easily understand why a decision tree classifies an instance as belonging to a specific class. Decision tree algorithms have also an embedded feature selection process to find the most important attributes which is very useful in the present study. C4.5 (Quinlan, 1993) is one of the best-known algorithms for building decision trees and has been adopted in this study as well.

In Figure 4, a decision tree representing the students' chance for acquiring a passing grade (>5) is presented. The accuracy of the C4.5 algorithm was found to be 81.84% using tenfold cross-validation. As shown in the graph, students' perception about the usefulness of the material delivered via Moodle significantly differentiates the variables influencing the students' outcome.

In the first scenario (perception value equal or less than 2) inspection of the assignments more than 74 times led to a passing grade. On the contrary, a low value of assignment_view variable is a strong predictor of failure. This is a rather straightforward result, since it is possibly related to a student's low engagement with the given assignments. Since in this area the educational software assessment rubrics were presented, which constitutes an important element to efficiently complete weekly assignments, the variable indirectly indicates a strong student engagement with the learning process. Otherwise, users who accessed user_view at least 8 times also received a passing grade. This is an indication of engagement to ideas exchange and topic discussion. If the latter is not the case, then discrimination on the passing grade across the users who accessed questionnaire view between 45 and 95 times. Fewer than 44 accesses to the questionnaire service are related to course fail.

**Figure 4** Decision tree illustrating the scenarios for acquiring a passing grade




In the second scenario (perceived usefulness of the material greater than 2/5) coupled with reported Moodle self-efficacy measurement (Moodle_use_capability_ perceptions value greater than 2) is a strong predictor of student's success with 298 correctly (and 48 wrongly) predicted cases. If the latter is not the case and the students' perceived usefulness rating of the material is 3/5 then the decision tree predicts a student's fail. If their rating is greater than 3, factors such as computer use per week, value of the variable user_view which is an indication of engagement to ideas exchange and topic discussion with other students and students' ease of Moodle use seem to influence the prediction. However, the results obtained in the before mentioned scenario should be handled with care since not many students lied in this category. As a result, possible classifications obtained could be attributed to chance, thus not depicting the real learning progress or the opposite.

The significant lesson learnt while examining the decision tree presented in Figure 4, is that students' perception constitute a remarkably strong predictor of the students' success. The result stresses the differentiations imposed by the adopted learning design, since in every laboratory session the students were asked to carry out some actions using Moodle. As a result, the differences in number of actions carried out in the system were reduced.

### 4.3 Association rules

Class association rules (Bing et al., 1998) were derived using Weka. Association rules are considered one of the most robust data mining algorithms in the research community (Wu et al., 2008).

In the present study, 26 rules were generated in total. Three rules for students with grades lower than 5 were obtained, two rules for students with a grade between 5 and 6.99, 15 for students with grades from 7 to 8.49 and six rules for students with a grade higher than 8.5. After each rule, the number of students for whom the rule is successfully applicable is presented in parenthesis, accompanied with the number of students for whom the rule did not apply successfully. As derived by the analysis, it is quite effortless to distinguish across the students with a pass or fail; Only three rules should be examined which involve six different variables strongly associated with their performance. In specific, *Lack of computer at home* (computer_at_home=2) combined with low self-efficacy was associated with student's fail (rule 1). Moreover, *perception of Moodle LMS* as a difficult to use system (ease_of_Moodle_use_ perceptions<3) combined with low frequency of access in the sections of related resources (which provides useful learning materials) is also an alarming signal for student's expected performance. Also, a low value of the assignment_view is associated with poor student performance, in concordance to the ascertainment derived while inspecting the decision tree obtained.

1 (Computer_at_home >= 2) and (Moodle_use_capability_perceptions <= 2) => Final_Note <5 (10.0/1.0).

2 (Ease_of_Moodle_use_perceptions <= 3) and (resource_view <= 117) and (user_ view_all >= 2) => Final_Note<5 (9.0/0.0).

3 (assignment_view <= 31) => Final_Note<5 (7.0/2.0).

On the contrary, inspection of the associated rules obtained for students with excellent grades provides a consistent result: Excellent grades were associated with high student activity. In specific, high total number of total actions (>793) combined with specific usage such as access to forum (>18) or user view (>29) or assignment area (>124) is related to an excellent grade. Also, even low access to the glossary seems to represent a high performing student.

1 (Total_Of_ id >= 793) and (forum_view >= 18) and (user_view >= 29) => Final_Note= [8.5–10] (31.0/10.0).

2 (Total_Of_ id >= 793) and (forum_view >= 18) and (user_view >= 37) => Final_Note= [8.5–10] (24.0/7.0).

3 (Total_Of_ id >= 793) and (assignment_view >= 124) => Final_Note=[8.5–10] (38.0/16.0).

4 (assignment_view >= 102) and (glossary_view >= 8) => Final_Note=[8.5–10] (33.0/15.0).

5 (glossary_view >= 2) and (resource_view >= 177) => Final_Note= [8.5–10] (49.0/24.0).

6 (glossary_view >= 2) and (Perceived_Usefulness_assignment <= 3) and (resource_ view >= 182) => Final_Note=[8.5–10] (20.0/7.0).

The main conclusion drawn from this method is that failure in the course is associated with the negative attitudes and perceptions of students towards Moodle. On the contrary, excellent grades are associated with increased usage. An interesting result is that apart from the total number of actions carried out, use of specific services such as forum and glossary indicate students' motivation to learn and a successful learning strategy. This is a plausible result since increased forum participation and deep study of terms as defined in the glossary barely can characterise a student with low interest and performance in the course.

### *4.4 Clustering*

The goal of a clustering approach is to distinguish a finite unlabeled data set into a finite and discrete set of hidden data structures (Xu and Wunsch, 2005). *k*-Means (Arthur and Vassilvitskii, 2007) is a rather simple but frequently used clustering algorithm. In the present study, we discretised the attribute 'grade' in four 'values' and we used $k = 4$ for the application of *k*-means. The presented results that are those exported by *k*-means with $k = 4$, using all variables as independent ones.

The result of the application of 4-means clustering on students' activity data is presented in Table 2. Computer possession was a strong indicator of students' performance. However, computer use per week is not positively correlated to the student's performance, although students with very high reported weekly computer use received excellent grades. In concordance with the methods described previously, more positive perceptions towards Moodle usability, capabilities and usefulness are generally associated with higher student grades. Also, students with total actions on the system close to, or above, the mean value (693) tend to receive a higher grade (at least 7). This is also true for the forum view, resource view and user view activity. Low access to the glossary strongly differentiates students who failed to pass the course. Finally, it is derived that students with excellent grades had significantly higher recorded number of actions related to the course_view variable.

**Table 2** The application of 4-means in our dataset using Weka

| *Variable* | *Full Data* | *Final_Note <5* | *Final_Note = [5,7)* | *Final_Note = [7,8.5)* | *Final_Note =[8.5,10]* |
|---|---|---|---|---|---|
| Computer_at_home | 1.1369 | 2 | 1 | 1 | 1 |
| Internet_at_home | 1.4107 | 2 | 2 | 1 | 1 |
| Computer_use_per_week | 2.6101 | 3.0217 | 2.2717 | 1.3837 | 3.6607 |
| Ease_of_Moodle_ use_perceptions | 4.1875 | 3.6087 | 4.0435 | 4.1279 | 4.5893 |
| Moodle_use_ capability_perceptions | 4.2411 | 3.7174 | 4.1196 | 4.1628 | 4.6161 |
| Attitude_about_Moodle | 4.2679 | 4.087 | 4.2826 | 4.2791 | 4.3214 |
| Perceived_Moodle_ Usefulness_lesson | 3.625 | 3.4348 | 3.5543 | 3.9535 | 3.5089 |
| Perceived_Usefulness_as signment | 3.6429 | 3.4348 | 3.6196 | 4.0116 | 3.4643 |

**Table 2** The application of 4-means in our dataset using Weka (continued)

| *Variable* | *Full Data* | *Final_Note <5* | *Final_Note = [5,7)* | *Final_Note = [7,8.5)* | *Final_Note =[8.5,10]* |
|---|---|---|---|---|---|
| Total_Of_ id | 693.006 | 648.8696 | 646.0217 | 688.3256 | 753.3214 |
| assignment_view | 88.9613 | 85.0435 | 83.8587 | 88.9767 | 94.75 |
| course_view | 188.7946 | 154.9348 | 159.6848 | 172.4302 | 239.1786 |
| forum_add_post | 0.1577 | 0.0652 | 0.0435 | 0.2326 | 0.2321 |
| forum_view | 15.6994 | 10.6304 | 11.4565 | 18.7674 | 18.9107 |
| glossary_view | 5.747 | 3.2174 | 6.0652 | 6.5 | 5.9464 |
| questionnaire_view | 149.3095 | 175.6304 | 146.9348 | 124.8953 | 159.1964 |
| resource_view | 127.8006 | 111.9565 | 132.8478 | 126.7093 | 131 |
| user_view | 20.8929 | 14.0217 | 20.0761 | 29.4186 | 17.8393 |
| user_view_all | 10.6875 | 4.2174 | 6.5761 | 13.907 | 14.25 |

*Source:* Hall et al. (2009)

## 5 Discussion

The aforementioned methods are of complementary utility, unveiling different aspects of students' performance in a variety of a ways. Visualisation techniques, could inform tutors without forcing them to deal with huge and unstructured volumes of interaction data. Moreover, variable importance plots could highlight important variables which require increased attention. Decision trees quickly unveil relation of attributes which differentiate a successful student, thus allowing the tutor to provide feedback in time. Association rules, with their if-then-else form, unveil concrete patterns of use which lead to increased student performance thus allowing timely and relevant feedback. Finally, clustering techniques provide structured information about students' characteristics in each performance group. Such a clustering technique could be used to identify learners with different expected performance and reorganise possible collaboration groups accordingly.

As shown previously, students' performance seems to be closely interrelated not only with their digital footprint but with their beliefs and perceptions about the Moodle usefulness and usability. In specific, failure in the course is associated with negative attitudes and perceptions of the students towards Moodle. On the other hand excellent grades are associated with increased use of the LCMS. This is an important finding, which stress the importance to provide an unobtrusive experience to the students and to persuade them that the learning platform will not require significant effort to use it appropriately. Moreover, the provided material and functions should be adequately differentiated from traditional approaches in order encourage students to invest time and effort in the LCMS.

It is argued that the aforementioned approaches could be effectively combined to provide a more concise view of students' learning progress. Such visual representations could be implemented in real time and while the course is delivered, thus helping teachers to adapt their teaching in a more objective and informed manner. For instance, a tutor can briefly inspect visualised data to quickly identify isolated students, outliers or useful materials which were not drawn the attention of the participants. In addition,

implementation of real-time visual feedback mechanisms related to each student activity in comparison to the whole class could leverage intrinsic motivation and re-align student effort even at the early stages of the course's delivery. Subsequently, he/she can closely observe and confirm his initial hypothesis using the patterns extracted by the approaches discussed previously and/or the visualised data obtained.

Towards this goal, a visualisation module will be implemented in forthcoming versions of MP, to enable various real-time visualisations of students' interaction. At the moment, the tool stores the pre-processed data into a relational database, thus an educator even with minimum technical knowledge can inspect the data and identify various interest interaction patterns. However, development of a usable dashboard in the form of a web-based application to study visualisations, make comparisons and allow inspection of predefined alarms (for instance, low student participation), could substantially reduce the effort required to deploy Learning Analytics techniques in real-world educational settings, thus increasing probability of adoption. Realisation of variable importance plot to quickly identify the critical attributes which substantially influence the learning outcome is considered a significant feature of such a LA Dashboard.

Nevertheless, the methods reported in this work were also used in other research efforts (Romero et al., 2008; Romero et al. 2013a; Romero et al. 2013b). However, to the best of our knowledge, we are not aware of similar studies conducted in a rigorously designed blended learning setting. For instance, Romero et al. (2013b) clearly state that their online environment was not a mandatory module of the course. In our case, the students were requested to retrieve specific elements of educational material, download and submit all required assignments via the LCMS system. Moreover, the lesson has been designed, builds and delivered around Moodle which is considered an indispensable part of the course. Also, in our study we do not report application of the method for a course delivered of a single year, but for 3 consecutive years. A more exhaustive study of the year-by-year differences constitutes a future research goal.

Due to the aforementioned ascertainment (i.e. strong differences in the learning design), different variables emerge as significant. For instance, in the course studied by Romero et al. (2008) the significance of time is much higher compared to the blended learning setting studied in our work. In addition, in our work we combine various types of data, such as students' perceptions, with interaction data. In the presented case, as derived while applying a decision tree approach, students' success in the course is heavily influenced by such perceptions. Therefore, it is argued that predictive techniques and Learning Analytics Dashboard should incorporate such data in order to achieve better accuracy. Also, an interesting finding in our study is that the variables related to 'non-obligatory' activity seems to constitute a better predictor of student's performance. Since students were asked to carry out specific tasks to acquire a passing grade, it is expected that their digital footprint in the variables expressing such behaviour will not be dramatically different. On the contrary, browsing sections such as glossary and using functionalities such as forum, which are not prerequisites actions, possibly indicate students with strong motivation to learn. Therefore, high activity in these variables is expected to predict more efficiently the student's learning outcome.

## 6 Conclusions

In this paper, students' practices while using a Learning Content Management System in a blended learning environment were examined. Using four complementary techniques,

namely data visualisation coupled with variable importance plot, decision trees, class association rules, and clustering it was derived that student's recorded interaction practices coupled with their perceptions towards LCMS adoption are a significant predictor of their performance. It is argued that presentation of lessons learned, contrasted with results already collected in other studies could inspire and encourage researchers to identify possible differences and investigate possible causal relationships with characteristics of the educational setting, the learners' attitudes towards e-learning adoption, or the topic taught, thus providing a fertile ground to implement more efficient learning analytic approaches and tools.

However, the study is not without limitations. The participants were students from a Department of Social Sciences with specific characteristics such as age, gender, computer skills and experience etc. Moreover, the results obtained do not explain how the students have benefited from their engagement with the LCMS system. Other studies using similar approaches (Romero et al., 2008; Romero and Ventura, 2010; Romero et al., 2013a; Romero et al., 2013b), share the same limitations, thus stressing the need of conducting more studies in a variety of settings.

In addition, it is not known to what extent the students were improved in other important non-cognitive aspects, such as self-organisation, collaboration, attitudes towards technology and openness (Tapscott, 2009) and in turn, how these aspects influence students' performance and consequently the prediction process. Moreover, it is argued that other approaches to study students' activity such as eye tracking sessions (Katsanos et al., 2010) and task modelling techniques (Tselios and Avouris, 2003; Tselios et al., 2008a; Tselios et al., 2008b) could enrich the effectiveness and robustness of the presented approach. Similar Learning Analytics techniques should be also applied in other educational contexts as well as in different web-based collaborative learning platforms such as wikis (Tselios et al., 2011a; Tselios et al., 2011b) to investigate the generalisability of the proposed approach.

## Acknowledgements

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program 'Education and Lifelong Learning' of the National Strategic Reference Framework (NSRF) – Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

## References

Ali, L., Hatala, M., Gašević, D. and Jovanović, J. (2012) 'A qualitative evaluation of evolution of a learning analytics tool', *Computers & Education*, Vol. 58, No. 1, pp.470–489.

Arthur, D. and Vassilvitskii, S. (2007) 'k-means++: the advantages of careful seeding', in Society for Industrial and Applied Mathematics Philadelphia (Eds): *SODA'07 Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 7–9 January, New Orleans, Louisiana, USA, pp.1027–1035.

Baepler, P. and Murdoch, C.J. (2010) 'Academic analytics and data mining in higher education', *International Journal for the Scholarship of Teaching & Learning*, Vol. 4, No. 2, pp.1–9.

Baker, R. and Yacef, K. (2009) 'The state of educational data mining in 2009: a review and future visions 2009', *Journal of Education Data Mining*, Vol. 1, No. 1, pp.3–17.

Bing, L., Wynne, H. and Yiming, M. (1998) 'Integrating classification and association rule mining', in Agrawal, R., Stolorz, P.E. and Piatetsky-Shapiro, G. (Eds): *4th International Conference on Knowledge Discovery and Data Mining*, 27–31 August, New York, NY, USA, pp.80–86.

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

Delavaria, N., Phon-Amnuaisuka, S. and Beikzadehb, M.R. (2008) 'Data mining application in higher learning institutions', *Informatics in Education*, Vol. 7, No. 1, pp.31–54.

Dominguez, A.K., Yacef, K. and Curran, J. (2010) 'Data mining to generate individualized feedback', *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, 14–18 June, Pittsburgh, PA, USA, pp.303–305.

Duffy, T.M. and Kirkley, J.R. (2004) 'Learning theory and pedagogy applied in distance learning: the case of Cardean University', in Duffy, T. and Kirkley, J. (Eds): *Learner-Centered Theory and Practice in Distance Education: Cases from Higher Education*, Lawrence Erlbaum, Mahwah, NJ, USA, pp.107–143.

Ferguson, R. (2012) 'Learning analytics: drivers, developments and challenges', *International Journal of Technology Enhanced Learning*, Vol. 4, Nos. 5/6, pp.304–317.

Garcia, E., Romero, C., Ventura, S. and Castro, C. (2011) 'A collaborative educational association rule mining tool', *The Internet and Higher Education*, Vol. 14, No. 2, pp.77–88.

Garrison, D.R. and Kanuka, H. (2004) 'Blended learning: uncovering its transformative potential in higher education', *The Internet and Higher Education*, Vol. 7, No. 2, pp.95–105.

Graham, W. (2011) *Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery Series*, Springer Science + Business Media.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) 'The WEKA data mining software: an update', *ACM SIGKDD Explorations Newsletter*, Vol. 11, No. 1, pp.10–18.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York, NY, USA.

Jong, B.S., Chan, T.Y. and Wu, Y.L. (2007) 'Learning log explorer in e-learning diagnosis', *Journal of IEEE Transactions on Education*, Vol. 50, No. 3, pp.216–228.

Jovanovica, M., Vukicevica, M., Milovanovica, M. and Minovica, M. (2012) 'Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study', *International Journal of Computational Intelligence Systems*, Vol. 5, No. 3, pp.597–610.

Katsanos, C., Tselios, N. and Avouris, N. (2010) 'Evaluating web site navigability: validation of a tool-based approach through two eye-tracking studies', *New review of Hypermedia and Multimedia*, Vol. 16, Nos. 1/2, pp.195–214.

Lockyer, L. and Dawson, S. (2011) 'Learning designs and learning analytics', in ACM (Eds): *1st International Conference on Learning Analytics and Knowledge (LAK 11)*, 27 February–1 March, Banff, Alberta, Canada, pp.153–156.

Long, P. and Siemens, G. (2011) 'Penetrating the fog: analytics in learning and education', *Educause Review Online*, Vol. 46, No. 5, pp.31–40.

Macfadyen, L.P. and Dawson, S. (2010) 'Mining LMS data to develop an "early warning system" for educators: a proof of concept', *Computers & Education*, Vol. 54, No. 2, pp.588–599.

Mandinach, B.E. (2005) 'The development of effective evaluation methods for e-learning: a concept paper and action plan', *Teachers College Record*, Vol. 107, No. 8, pp.1814–1835.

Means, B., Toyama, Y., Murphy, R., Bakia, M. and Jones, K. (2010) *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies*, U.S. Department of Education, Office of Planning, Evaluation, and Policy Development Policy and Program Studies Service, Washington, DC, USA.

Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaiane, O. (2009) 'Clustering and sequential data mining of online collaborative learning data', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 6, pp.759–772.

Pedraza-Perez, R., Romero, C. and Ventura, S. (2011) 'A Java desktop tool for mining Moodle data', in Pechenizkiy, M. et al. (Eds): *3rd Conference on Educational Data Mining*, 6–8 July, Eindhoven, The Netherlands, pp.319–320.

Psaromiligkos, Y., Orfanidou, M., Kytagias, C. and Zafiri, E. (2011) 'Mining log data for the analysis of learners' behaviour in web-based learning management systems', *Journal of Operational Research*, Vol. 11, No. 2, pp.187–200.

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Francisco, CA, USA.

Romero, C., Espejo, P.G., Zafra, A., Romero, J.R. and Ventura, S. (2013a) 'Web usage mining for predicting final marks of students that use Moodle courses', *Computer Applications in Engineering Education*, Vol. 21, No. 1, pp.135–146.

Romero, C., López, M.I., Luna, J.M. and Ventura, S. (2013b) 'Predicting students' final performance from participation in on-line discussion forums', *Computers & Education,* Vol. 68, No. 9, pp.458–472.

Romero, C. and Ventura, S. (2010) 'Educational data mining: a review of the state of the art', *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 40, No.6, pp.601–618.

Romero, C., Ventura, S. and Garcia, E. (2008) 'Data mining in course management systems: Moodle case study and tutorial', *Computers & Education*, Vol. 51, No. 1, pp.368–384.

Siemens, G. and Baker, R.S.D. (2012) 'Learning analytics and educational data mining: Towards communication and collaboration', *2nd International Conference on Learning Analytics and Knowledge (LAK'12)*, 29 April–2 May, Vancouver, British Columbia, Canada, pp.252–254.

Tapscott, D. (2009) *Grown up Digital: How the Net Generation is Changing Your World*, McGraw-Hill, New York, NY, USA.

Tselios, N., Altanopoulou, P. and Komis, V. (2011a) 'Don't leave me alone: effectiveness of a framed wiki-based learning activity', *7th International Symposium on Wikis and Open Collaboration*, 3–5 October, Mountain View, CA, USA, pp.49–52.

Tselios, N. and Avouris N.M. (2003) 'Cognitive task modeling for system design and evaluation of non-routine task domains', in Hollnagel's, E. (Ed.): *Handbook of Cognitive Task Design*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, pp.307–332.

Tselios, N., Avouris, N. and Komis, V. (2008a) 'The effective combination of hybrid usability methods in evaluating educational applications of ICT: issues and challenges', *Education and Information Technologies Journal*, Vol. 13, No. 1, pp.55–76.

Tselios, N., Daskalakis, S. and Papadopoulou, M. (2011b) 'Assessing the Acceptance of a blended learning university course', *Educational Technology & Society*, Vol. 14, No. 2, pp.224–235.

Tselios, N., Katsanos, C., Kahrimanis, G. and Avouris, N. (2008b) 'Design and evaluation of web-based learning environments using information foraging models', in Pahl, C. (Ed.): *Architecture Solutions for E-Learning Systems*, Information Science Reference, Hershey, PA, USA, pp.320–339.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H. and Steinberg, D. (2008) 'Top 10 algorithms in data mining', *Knowledge and Information Systems*, Vol. 4, No. 1, pp.1–37.

Xu, R. and Wunsch, D. (2005) 'Survey of clustering algorithms', *IEEE Transactions on Neural Networks*, Vol. 16, No. 3, pp.645–678.