Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study

Milos Jovanovic, Milan Vukicevic, Milos Milovanovic, Miroslav Minovic
Faculty of Organizational Sciences, University of Belgrade, Jove Ilica 154
Belgrade, Serbia
E-mail: {milos.jovanovic, milan.vukicevic, milos.milovanovic, miroslav.minovic}@fon.bg.ac.rs
www.bg.ac.rs

Received 6 December 2011
Accepted 15 May 2012

Abstract

In this research we applied classification models for prediction of students' performance, and cluster models for grouping students based on their cognitive styles in e-learning environment. Classification models described in this paper should help: teachers, students and business people, for early engaging with students who are likely to become excellent on a selected topic. Clustering students based on cognitive styles and their overall performance should enable better adaption of the learning materials with respect to their learning styles. The approach is tested using well-established data mining algorithms, and evaluated by several evaluation measures. Model building process included data preprocessing, parameter optimization and attribute selection steps, which enhanced the overall performance. Additionally we propose a Moodle module that allows automatic extraction of data needed for educational data mining analysis and deploys models developed in this study.

Keywords: educational data mining, prediction, students, performance, classification, clustering, Moodle.

1. Introduction

Moodle is an open source Learning Management System (LMS) that is mostly regarded as Course Management System by the open community. It is dominantly used in higher education and it has proven as a successful tool in that setting.[1,2] For that reason our faculty built a distance learning system (DLS) based on Moodle LMS. The system was built and developed as an in-house solution at University of Belgrade for the students of Information technology. One of the main requirements was to completely support distance learning process in all its aspects. The system enables dealing with advanced courses, which use multimedia lessons, advanced workshops and face to face communication through video conferencing.

Web-based learning management systems are extensively used nowadays and produce vast amounts of data that are potentially useful for improving educational process.[2,4,5] The new emerging field, called Educational Data Mining (EDM), concerns with developing methods that discover knowledge from data originating from educational (traditional or distance learning) environments.[6] Increasing research interests in using data mining in education is recorded in the last decade[7,8,9,10,11] with focus on different aspects of educational process (e.g. students, teachers, teaching materials, organization of classes etc.).

Benefits from extracting knowledge from e-learning data are expected under assumption that the trails of user actions can be used to identify specific information on users. We hope that the user behavior captured in log files and recorded in data structures can be used to create models that predict user behavior, or describe their peculiarities. There are several groups of people who can leverage this knowledge, and are potential stakeholders: Students, Teachers, e-learning system administrators, University management.

These stakeholders could use this knowledge for different goals[9]:

Published by Atlantis Press

1. Applications dealing with the assessment of students' learning performance.
2. Applications that provide course adaptation and learning recommendations based on the students' learning behavior.
3. Approaches dealing with the evaluation of learning material and educational web based courses.
4. Applications that involve feedback to both teachers and students of e-learning courses, based on the students' learning behavior.
5. Developments for the detection of atypical students' learning behavior.

These goals are achieved with help of data mining techniques such as k-nearest neighbor, naive Bayes, decision trees, artificial neural networks, support vector machines, K-means, hierarchical clustering etc.[12]

Still, learning management systems are not primarily designed with data analysis and mining in mind, because usage data is not stored in a systematic way. Its thorough analysis requires long and tedious pre-processing.[13] Furthermore, LMS systems usually produce statistic reports. These reports however do not assist instructors in drawing out useful conclusions either for the course potential or student abilities and are useful only for platform administrative purposes.[2]

This research shows how one can leverage the available data on student behavior, in order to predict success of students, as well as profile students into groups which may help improve existing learning material and collaborative learning. The study involves data from students attending online (distance learning) university courses as suggested by Romero et al.,[6] and extends available data with students cognitive styles. Additionally we propose Moodle module that allows automatic extraction of data needed for EDM analysis and deploys models evolved in this study.

The paper is structured as follows: Section 2 introduces related work on using e-learning data and applying data mining models. Architectural design of the decision-support system is given in Section 3, with experimental results in using data mining models presented in Section 4. Potential ways of using knowledge gained by data mining models is described in Section 5, and Section 6 discusses open issues and related problems for these types of applications.

2. Background

Romero and Ventura gave a systematic survey about EDM from 1995 to 2005.[10] Because of increasing popularity and number of researches in this area, the same authors gave an extensive overview about the state of the art in this area until 2011 with over 300 references.[12] In this paper we will focus on researches that are closest to our work. Study by Wang and Liao was performed in order to investigate how Data Mining techniques can be successfully used for adaptive learning.[14] In academic institutions, Moodle platform is often utilized as a significant part of e-learning systems. Romero et al. described how different data mining techniques can be used in that setting to improve the course and the students' learning.[6]

Applications or tasks that have been resolved through data mining techniques are classified by Romero and Ventura in twelve categories: Analysis and visualization of data, Providing feedback for supporting instructors, Recommendations for students, Predict students' performance, Student modeling, Detecting undesirable student behaviors, Grouping students Social network analysis, Developing concept maps, Constructing courseware, Planning and scheduling.[6]

One of the most frequent research topics in the area of EDM (also investigated in this research) is the prediction of student performance.[6,14,16] The main idea behind this research direction is that based on student activity one can predict the future outcome of student performance. For the purpose of predicting students' final outcome on a course, researchers used various techniques and algorithms. Kotsiantis et al., proposed an incremental ensemble of classifiers as a technique for predicting students' performance in distance education.[17] Neuro-fuzzy system rules are used for student knowledge diagnosis through game learning environment.[18] Kotsiantis also proposed a prototype version of decision support system for prediction of students' performance based on students' demographic characteristics and their marks in a small number of written assignments.[19] Myller et al. used neural networks (multilayer perceptron),[20] and Traynor and Gibson used combination of Artificial Neural Networks and Evolutionary Computation models to predict students' performance.[21] Similarly, Minaei-Bidgoli et al. used genetic algorithm for optimization of multiple classifier performance.[22] Delgado et al. used neural networks to predict success of the students defined with binary classes (pass or fail).[23]

Grouping students is another important research task in educational environments. Tang and McCalla, suggested data clustering as a basic resource to promote group-based collaborative learning and to provide incremental student diagnosis.[14] Student grouping by neural network based on affective factors in learning English was proposed by Bachtiar et al.[23] The clustering technique based on the implementation of the Bisection K-Means algorithm and Kohonen's SOM algorithm was applied in several researches.[25,26,27] These algorithms were used to group similar course materials with the aim of helping users to find and organize distributed course resources in process of online learning. Also, the use of k-means clustering algorithm for predicting student's learning activities was described in Ayesha's work, where the information generated after the implementation of data mining technique may be helpful for instructor as well as for students.[28] Ayers et al. used K-means and model based algorithm to group students with similar skill profiles on artificial data.[29] Zakrzewska used hierarchical clustering for grouping students based on their learning styles,[30] defined with Felder and Sylverman model,[31] in order to build individual models of learners and adjust teaching paths and materials to their needs. Usefulness of combining cognitive and emotional aspects for investigations of students' learning was emphasized in Heikkila et al.[32] Perera et al. used a technique of grouping both similar teams and similar individual members, and sequential pattern mining was used to extract sequences of frequent events.[33]

Cognitive style approach for Mining students' learning patterns and performance in Web-based environment was proposed by Chen and Liu.[34]

Adán-Coello et al., included learning styles for forming groups for collaborative learning of introductory computer programming.[35] Learning styles are also included for the implementation of the adaptive system in electronic learning.[36]

Tang and Mccalla proposed a clustering algorithm based on large generalized sequences to find groups of students with similar learning characteristics based on their traversal path patterns and the content of each page they have visited.[14] Chen et al. used K-means clustering algorithm for effectively grouping students who demonstrate similar behavior in e-learning environment.

In this paper we utilized K-means algorithm on the data concerning students' cognitive styles (gathered through questionnaire), and so models that are generated could be applied in e-learning as well as traditional teaching environments.[37]

3. University of Belgrade case study

In this section, we describe the data that are needed for EDM and propose automatic procedure for data extraction and preparation. Further we build and evaluate data mining models. Finally Moodle module for deployment of models is described.

3.1. *Automatic data extraction*

For evaluation of our approach, we used the data from the Moodle system as recommended by Romero et al.[6] Moodle CMS is specific when database model is in question. Since Moodle is an open source solution, currently there is a great community of developers built around it. There are many people constantly developing this system and adding functionalities, and model of data management based on modules was used in order to enable easy expansion. What does this model imply? When a developer wants to add some functionality to existing Moodle version he develops adequate PHP pages and adds tables to the database model that will be used to manage data for the new set of functionalities. New data is connected to user information through including relations towards new sets of data. This aspect of the model complicates future extraction of information on students' activities for each new module. Romero et al. gave directions for extraction and preparation of data for EDM analysis based on series of queries defined in MySQL database.[6] Here we describe in more detail problems of data extraction and preparation. Additionally, we propose a procedure for automatic extraction and preparation of Moodle data for data mining analyses (see Figure 1) implemented in RapidMiner.[38]

There are two alternative sources on student activity data available in Moodle database structure. First source is an activity log that Moodle system uses to track the activities of each student. Moodle is a Web based system and it is not able to track continuous usage of the system since it is based on a HTTP request/reply model. It is very difficult to determine any time spent in some activity since activity is only listed if a user performs click action on some link.

Second source of data can be a set of tables that are created for each individual module. They keep track of major students' activities regarding that specific module. For instance when a student performs an *Assignment,* that module keeps track when a student reads the assignment, when he/she submits it, edits it, etc. Unfortunately, this data, as opposed to the first source, provides less information on each action student performs, during each activity. During our analysis we decided to combine data from both sources since this combination gives good foundation for EDM analysis.[6]

Figure 1 shows the stream for automatic extraction and aggregation of Moodle data for EDM analyses. In order to prepare the data in format described in Table 1 we first extracted the data about students, courses and the grades achieved on every course. Second, the data from different modules is extracted, namely *Assignment*, *Quiz* and *Forum*. In case of the *Assignment* module, tracking students time spent on a specific assignment was a trivial issue. Data on starting time and submission time were provided by that module. Also in case of a *Quiz* module, time on each test is measured and recorded in the adequate data structure.

Table 1: Description of data used in experiments for each user per course.

Name	Description
course	Identification number of the course.
n_assigment	Number of assignments done.
n_quiz	Number of quizzes taken.
n_quiz_a	Number of quizzes passed.
n_quiz_s	Number of quizzes failed.
n_posts	Number of messages sent to the forum.
n_read	Number or messages read on the forum.
total_time_assignment	Total time spent on assignments.
total_time_quiz	Total time spent on quizzes.
total_time_forum	Total time spent on forum.
mark	Final mark obtained by the student in the course.

Since Moodle is primarily an LMS system, it faces common issues of such systems. LMS systems use resources from many alternative sources. Additionally, Moodle is an Open Source system which implies many developers working on the same task. This sometimes leads to data heterogeneity, especially syntactic heterogeneity. This means that information sources may use different representations and encodings for data. Syntactic interoperability can be achieved when compatible forms of encoding and access protocols are used to allow information systems to communicate. [39]

For our analysis we used data extracted from standard Moodle modules that have a long history of development. They are an integrated part of the production issue of Moodle, and this helped in data standardization and prevention of data heterogeneity. This is why our work wasn't struck by this particular problem. On the other hand data inconsistency is usually generated by improper use of the system and almost every system suffers from this problem. Problems with data inconsistency are expected with open source systems. Regardless, Moodle is rather consistent with the data it uses. For instance time values are always represented in a time stamp format that significantly simplifies preprocessing steps. This also enables easy data manipulation such as retrieving the duration time by subtracting the beginning time from end time for each activity. Most issues are generated by improper use of the system. For instance if a student does not complete the quiz and simply close the browser, the quiz will be regarded as still open, and end

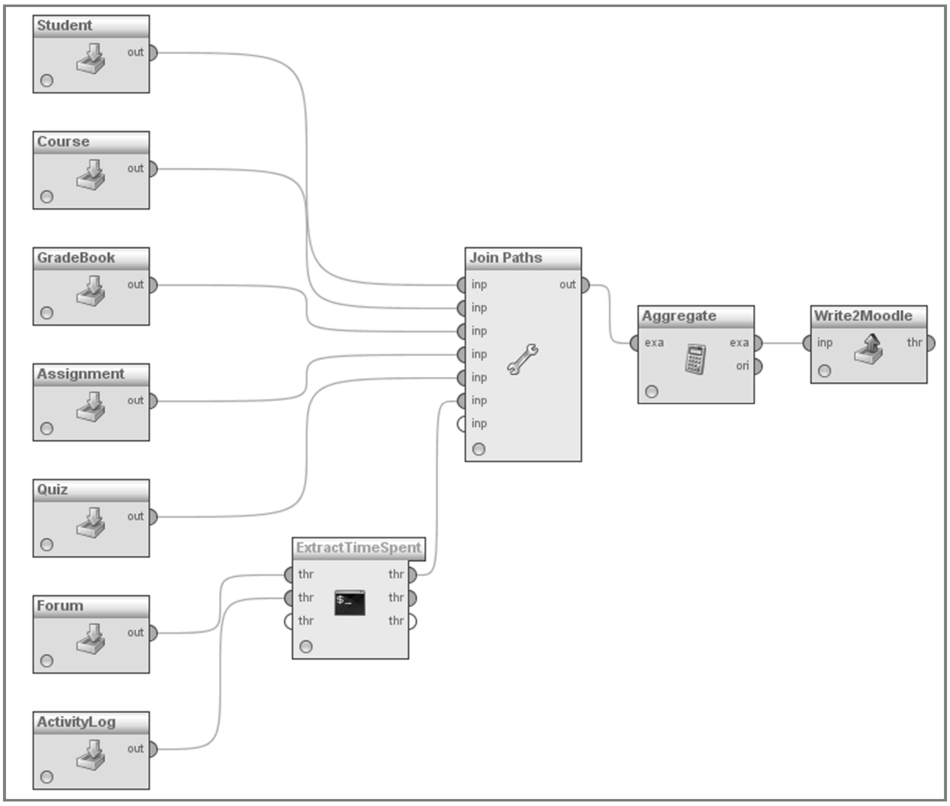


Fig. 1. Automatic data extraction model

time will be set to zero value since it is in a timestamp format. This is especially the case if the quiz is left open by the educator as is in the case of self evaluation tests that are open during the entire semester. By subtracting beginning time from end time one would get a negative time value. Such cases we had to exclude as if they did not even attempt the quiz. The same issue occurred with the assignments that students did not finish and upload.

Systems that have many users can suffer from data redundancy issues. In case of Moodle the most common case of redundancy is in duplicate courses or user accounts. In our organization we minimized the occurrence of redundancy by using the centralized approach in generation of courses and user accounts. System administrator is generating courses and user accounts upon teacher's requests and it is his duty to primarily check if new addition is already entered in the system. This process practically excluded the possibility of redundancy that targets this analysis.

When it comes to the Forum module, extraction of the information about time spent is more complicated. Since student can spend undetermined amount of time reading some forum without providing feedback to the system, it is difficult to determine if the user is active in the forum or not. In this case we used the activity log that tracks each click user makes on a link in a system.

Since Moodle provides module name as one of the meta-data regarding that action we were able to track the students' movement through a forum. The time spent was determined as the addition of the times between two clicks. If a student made a last click in a forum context and then was inactive for a prolonged period of time, as a referent time we decided to use an average time between two clicks in a forum context for all users. This is caused by the fact that users often do not properly log out of the system. Usually they just close the web browser and move on to other activities. Unfortunately, this does not leave any feedback when certain activity ended. For calculation of time spent on forums for every student on every course, we designed specialized application that is integrated into stream (Figure 1) and uses extracted data from Forum module and Activity Log. Finally, extracted data is aggregated on the student-course level.

Additionally we used the data about students' cognitive styles that are gathered from a questionnaire that we administered through Moodle. We administered self-report MBTI questionnaire which is already successfully used for analysis of student's profiles.[40] The MBTI form has 95 forced-choice items that forms four bipolar scales: Extraversion-Introversion (EI), Sensing-Intuition (SN), Thinking-Feeling (TF) and Judging-Perception (JP). A combination of these dimensions builds 16 different types of cognitive functioning. Introverts who are oriented primarily to the internal cues, and extroverts, who are oriented primarily to the external events, due to the differences in focusing psychical energy, show different pattern of performing intellectual tasks. Sensing mode of perceiving world is characterized by the respect for data obtained by one of the five senses. Contrary, intuitive type is prone to lean on inner processes, perceiving the bigger picture that enables him to concentrate and to see hidden possibilities, implications of the subject in matter.

Myers and McCaulley postulate two decision-making styles when assessing the validity of perception: thinking (assessment based on logical impersonal processes) and feeling (assessment based on personal, subjective process of mental evaluation).[41] There are individual differences in preference of the quality of environment one exist (learn) in, explicitly the level of structure inherently given in it. So, there are two categories of subjects: judgers who structure and order that promote predictable surrounding where decisions could have been brought quickly, and perceivers who need to keep options open unconcerned for deadlines. As MBTI is well theoretically conceptualized, [43] and metrically evaluated instrument[44,49], we believe that it might be useful to apply it on problem of distance learning. Carlson examined great body of reliability tests for this scale and found that coefficients for split-half reliability goes from .66 to .92, and test-retest reliability shows that results are relatively stable (coefficients in different studies are ranging from .69 to .89).[44] Records about students are extended with the attributes resulting from determining their cognitive style.

3.2. *Prediction of students' success*

We defined a classification model to predict if a student would display excellent performance (i.e. highest grades) on a selected course. This problem is interesting since there are many stakeholders interested in recognizing students with excellent performance.[45,46,47,48] For the input data for the prediction, the model would use the data describing student behavior on e-learning resources (e.g. forums, discussions, quizzes, posts, assignments) as described in the previous subsection. The dataset contains 260 instances. The preparation of data included extracting more features (such as grouping courses in math-oriented and social-oriented), normalizing features, and resolving missing values within data. As opposed to Romero et al.,[50] who discretized class label in four categories (fail, pass, good and excellent) class label for our model was a new binary attribute, which separated students with highest grade (label value 1), from the rest (label value 0). This was done because research of Romero et al.,[50] showed that classification algorithms couldn't achieve accuracy over 70% if class is discretized in four categories (even with several pre-processing setting of predictors). The goal is to devise a model that would be able to predict if a student will perform with excellent results, based on the input data.

The main usage of this model would be to early detect well performing students on a course. People who could benefit this model would be:

- teachers, for distinction of students they can collaborate with;
- students, for checking if there is a need for more effort to achieve better results;
- business people, for early engaging with students who are likely to become excellent on a selected topic.

We decided to utilize eight different state-of-the-art algorithms for classification, which often showed good results in the area of EDM.[6] The algorithms used are:

- AdaBoost (with C4.5 algorithm) (abbr: "Boost")
- Bagging (with C4.5 algorithm) (abbr: "Bag")
- J4.8 (a C4.5 implementation in Java)
- Linear Discriminant Analysis (abbr: "LDA")
- Logistic regression
- Naive Bayes
- Neural net (multi-layered perceptons) (abbr: "NN")

- Random Forests (abbr: “Forests”)

The results show expected accuracy of the models in percentage. Since this model would be used on future data, we used 10-fold cross-validation technique to prevent choosing an over-trained model, and assess models’ generalization ability. The cross-validation uses so called “stratified” sampling, which means that in each fold a similar class distribution is kept. Additionally, we measured other evaluation measures, Area Under Curve (AUC) and LIFT ratio. AUC estimate can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. LIFT ratio measures the degree to which the predictions of a classification model are better than randomly-generated predictions. It is defined as the ratio of true positives to total positives resulting from the classification process compared to the fraction of true positives in the overall population. We used both of these measures to complement the evaluation based on accuracy. This is important measure since in this research we are dealing with imbalanced data.

These measures are important because accuracy often tends to overlook the classifier inability to predict all the classes, when it is concentrating only to detect one class. Testing is done in RapidMiner data mining platform,[38] using default parameters and random seeds. The results are shown in Table 2.

Table 2: Performance results of different classification algorithms for predicting student excellence on a course.

Algorithm	Accuracy	AUC	Lift
AdaBoost (J4.8)	91.74 %	0.8256	4.1071
Bagging (J4.8 unprunned)	90.87 %	0.7504	2.0536
J4.8	93.04 %	0.5000	1
LDA	93.04 %	0.5000	1
Logistic Regression	92.17 %	0.5181	1.0575
Naive Bayes	53.48 %	0.7222	1.5375
Neural Net (Rapid default)	91.30 %	0.8346	4.7917
Random Forests	93.04 %	0.7498	7.1875

Looking at the results, several algorithms show good performance in generating the needed classification models. These are NeuralNet, AdaBoost and RandomForests, and all three are comparable with respect to accuracy. While RandomForests looks most useful by accuracy, AUC evaluation measure does not prefer this algorithm, since the AUC value is too small. The suggestion is then to avoid RandomForests algorithm, since it does not have similar power in predicting both “excellent” and “other” students, resulting in lower AUC performance. Basic, and more simple algorithms, such as Logistic regression, Naive Bayes and LDA generally performed poorly. While J4.8 and LDA seem to have good accuracy, looking at the AUC measure, we see that those classifiers have close to random performance, and the “accuracy” is due to class imbalance. Furthermore, we see poor results with all algorithms that produce linear models (LDA, Logistic regression and Naive bayes), which might suggest that the decision boundary for excelent students is non-linear. Other methods which produce more complex decision boundaries performed better. Overall, both NeuralNet and AdaBoost gave quite good results, rendering models with quality which allow further use. Since the results are taken using cross-validation, we can expect to successfully predict excellence of roughly 9 out of 10 students.

After selecting the three most promising classification algorithms for the task at hand, namely AdaBoost, NeuralNet and RandomForests, we tried to improve the performance, measured by AUC, by doing different preprocessing and parameter optimization steps. The setup for this preprocessing is shown in Figure 2.

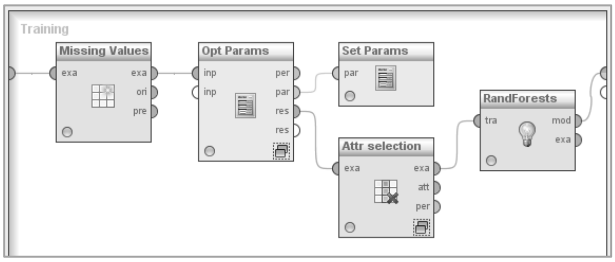


Fig. 2. Preprocessing steps in the training phase, for the selected algorithms

The results of applying these steps are given in Table 3, where the steps build upon each other in a “chain” when they bring any performance gain.

Table 3: Improvement of AUC by preprocessing, for selected algorithms

	no preprocess	handle missing	optimize parameters	attribute selection
RandForests	0.750	**0.858**	**0.890**	0.848
Adaboost	0.826	0.779	**0.839**	**0.838**
NeuralNet	0.835	0.767	**0.853**	0.812

For preprocessing, first we tried to average out the missing values present in the data, usually for attributes such as *total_time_assignment* or *total_time_quiz*. Some of the algorithms clearly needed this, since we see improvement of AUC for RandomForests. Next, we tried to optimize the algorithm learning parameters, to better fit the problem at hand. This step improved all three algorithms, and is clearly something to consider when applying these algorithms. Finally, we tried to perform attribute selection and remove attributes that act as "noise". The selection is done as backward elimination, removing one attribute at the time, until there are some improvements in the performance. Here we see that although removal of attributes did not contribute to AUC, for AdaBoost algorithm it did not render AUC any lower, even after completely removing two attributes. These attributes are *total_time_assignment* and *total_time_forum*, and AdaBoost could build a good model even without those, which is an important consideration when using this particular algorithm.

Another interesting detail to notice is the fact that RandomForests used the preprocessing to produce the best AUC score overall, even if it was least promising of the three algorithms, before preprocessing. Still, we could not strongly prefer any of these algorithms, so the recommendation for practical application is to use one of these three algorithms.

3.3. *Grouping students*

In order to provide information for better adaption of the learning material, we defined clustering models that would detect groupings of students with respect to their cognitive styles, as well as overall performance. Each student is described by cognitive styles and the score he achieved on a course. Data is separated by courses (separate cluster model is defined for each course), so student profiles can be considered for each course separately.

Using this model, one would be able to see which profile of students (defined by cognitive styles) is having difficulties, and whether that profile performed poorly in the past. This way, each teacher is guided in the way he should adapt the learning materials, to enable poor performing groups to increase performance for his course.

For this purpose we used k-means clustering algorithm, adapted for use over categorical data, since the data on cognitive styles are categorical. Results for the clustering on different courses are given in Figures 3-6.

For each course, several student profiles are found based on similarities of students by cognitive style.

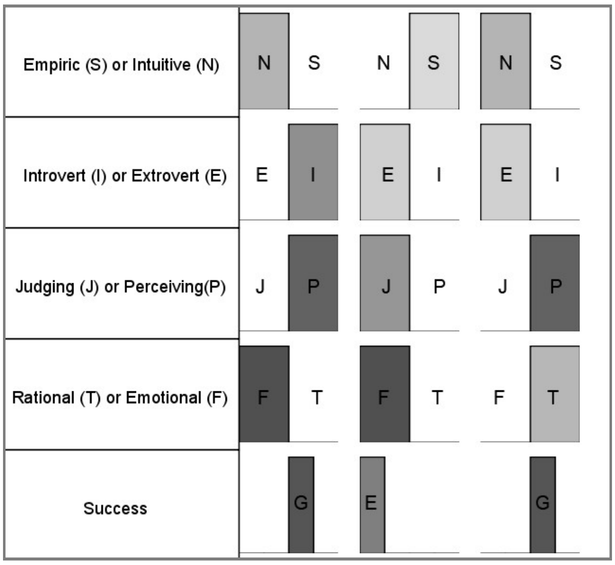


Fig. 3. Student profiles for course on Mathematics

Figures 3 and 4 show different groups of students on different courses. Each row represents different cognitive properties (described in detail in section 3.1). Each column represents one profile, which is a group (cluster) of students with similar cognitive properties. Variable "Success" describes the success of students of a particular profile, where "P" stands for poor performance, "G" for good, and "E" for excellent. Here we used three instead of two levels of success, in order to have more detailed description of found clusters. In essence, any number of success levels could be used, but empirically we detected highest clarity of cluster interpretations when using three levels. This was also true for the number of clusters, which we set to three.

We see that, for example, in course on Mathematics shown on Figure 3, students with profile "SEJF" had excellent results, while other profiles had moderate ("good") success. This means that students other than "SEJF" profile had more trouble in delivering best performance, which might be caused by many factors.

Still, since we know the cognitive profile of these students, we can direct our effort in adapting our course materials to fit that target group. For example, analysis indicates that course on Mathematics is more suitable to Empiric and Judging cognitive styles. This is probably due to the nature of the subject that gives the upper edge to Empirics that are better in deductive thinking and reasoning. Teacher can try to overcome that gap by adapting materials to the opposing cognitive styles. Intuitives might benefit through learning by doing approach, through simulations or games. Also Perceivers could find it more appealing to use interactive multimedia material. Also we could offer different examination approaches, as, for example, Introverts have more difficulty in verbal expression. Of course, to fully leverage this information, an advice from a psychologist would be very useful, but this is now possible due to this new information we have on students attending our course.

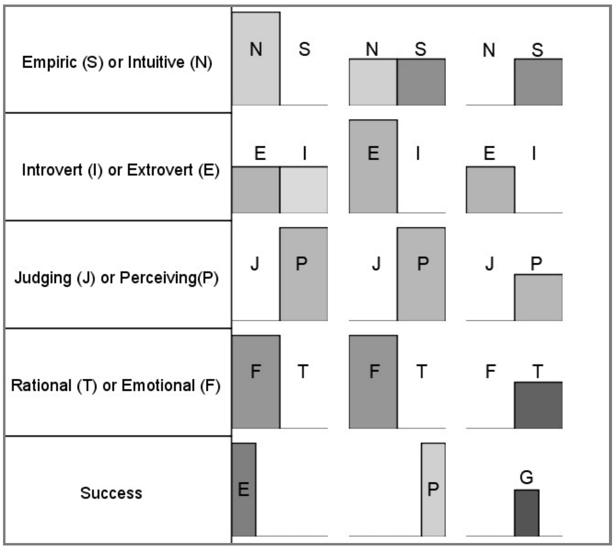


Fig. 4. Student profiles for course on Psychology

Naturally, different courses will better fit different profiles, because of the different areas of research and different materials offered by the instructors. Figure 4 shows profiles for the course on Psychology, where interestingly, profiles of “excellent” and “poor” students are similar, while “good” students are differing in cognitive styles. It is interesting though, that all the Introverts are grouped only in the “excellent” group, so it points out that Introvert students tend to understand this course material better. Also, profile with “good” students is distinct, especially for the Rational students, who always passed the Psychology course with “good”, but not with “excellent” results.

Also, there are occasions when we cannot isolate a complete cognitive profile of successful students, as in Figure 5. However, partial information could be observed, for example, looking at only first two cognitive attributes. Here, Empiric, as well as Introvert, are part of only the first cluster of students (first column), which all turned out excellent by the end of the course. This links only those attributes (Empiric and Introvert) to the success of students.

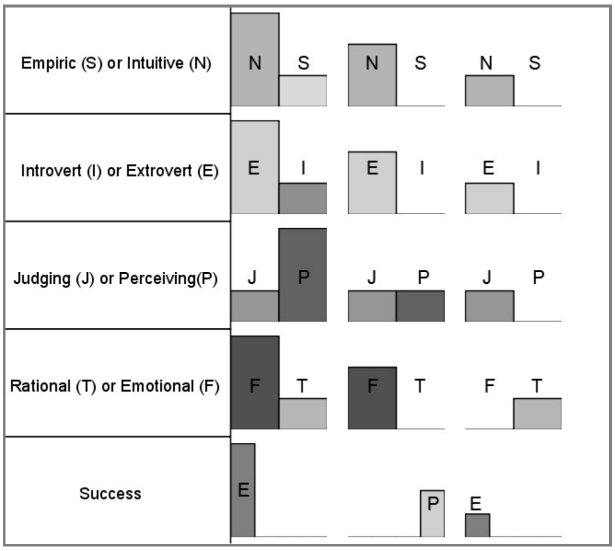


Fig. 5. Student profiles for course on Management

For English language (Figure 6), it is interesting that in first two clusters there were students with different success.

First cluster mainly contains students with “excellent” and “poor” success. Second cluster mainly contain students with “good” and “poor” success. Both of these clusters have an overlap between cognitive styles. Third cluster resolved this confusion by identifying that students with “poor” success are Empiric and Judging, in contrast to Mathematics where deductive thinking and reasoning (that characterize this combination of cognitive styles).

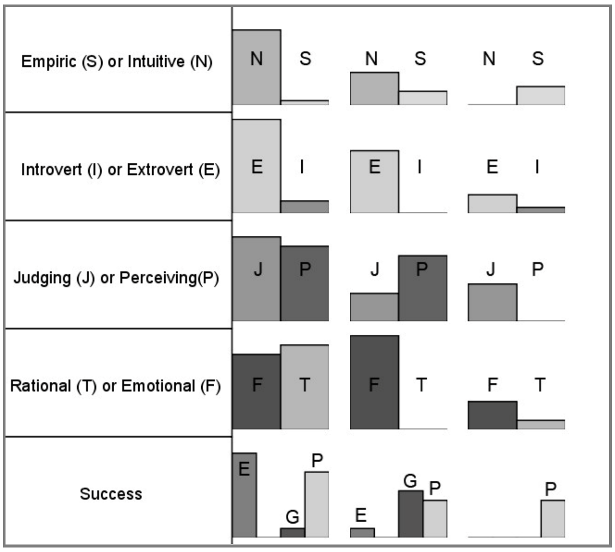


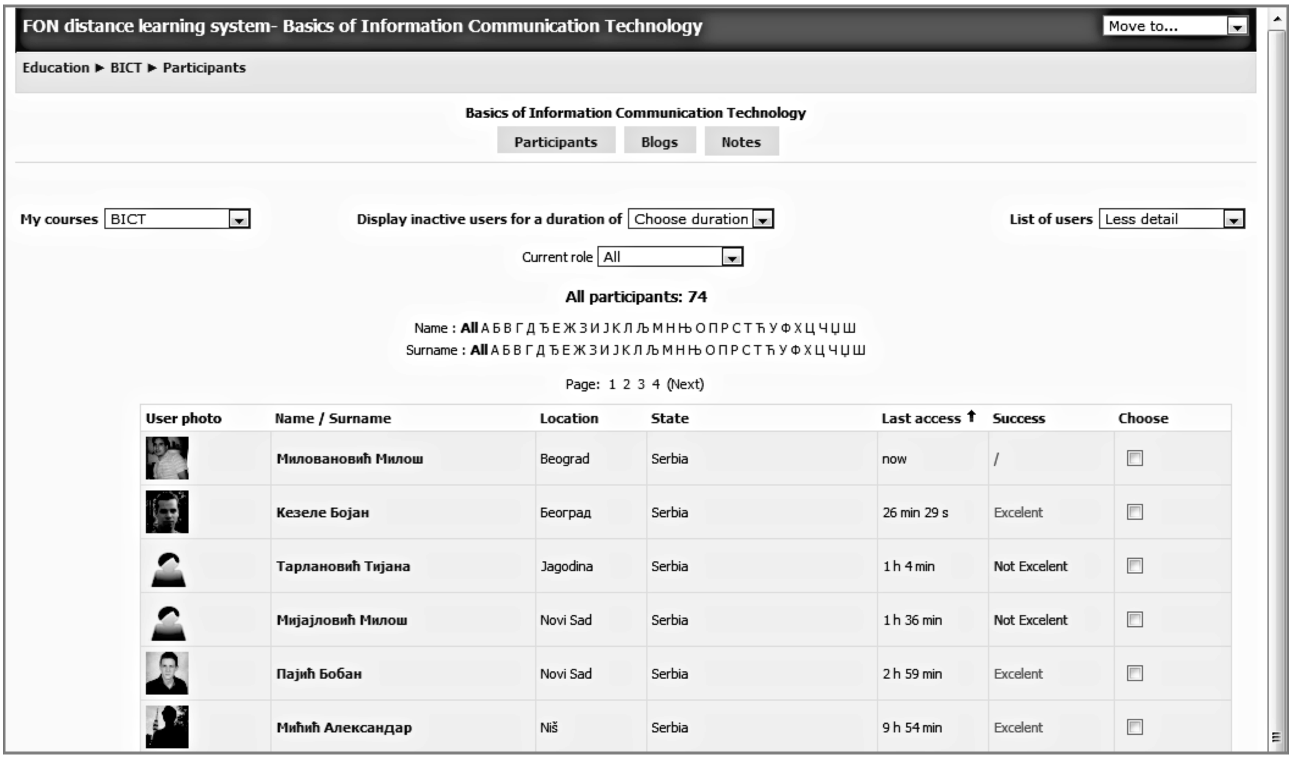
Fig. 6. Student profiles for course on English language

These students should work on developing a “sense” for language by more involving in listening conversations between native speakers or reading and understanding books written by native speakers. On the other hand, intuitive and perceiving students already have developed sense for the language and they don’t need additional activities.

3.4. *Deployment of models*

In order to provide educators with information acquired by using models defined in previous sub-sections we propose a Moodle module that utilizes defined models. They can browse through list of students involved in their course and see prediction of success for each student (Figure 7) on that course.

Based on this prediction, educators can change their approach in working with students that are predicted not to be excellent or further involve with students, predicted as excellent, in extra-curricular activities.



FON distance learning system- Basics of Information Communication Technology

Move to...

Education ▶ BICT ▶ Participants

Basics of Information Communication Technology

Participants Blogs Notes

My courses BICT

Display inactive users for a duration of Choose duration

List of users Less detail

Current role All

All participants: 74

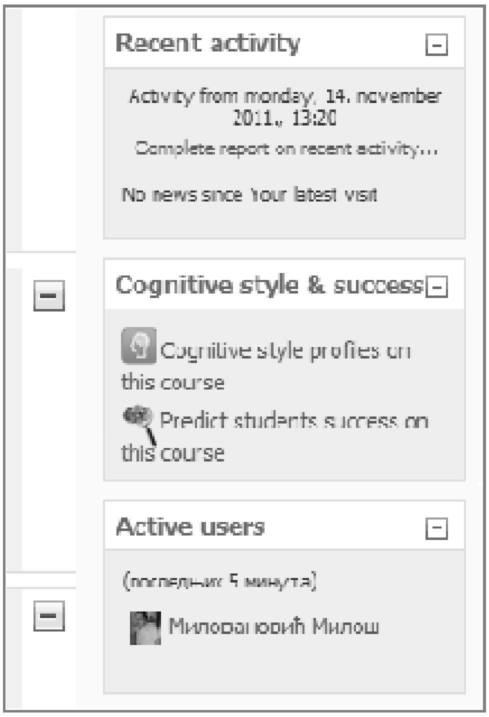
Name : **All** А Б В Г Д Ђ Е Ж З И Ј К Л Љ М Н Њ О П Р С Т Ћ У Ф Х Ц Ч Џ Ш
Surname : **All** А Б В Г Д Ђ Е Ж З И Ј К Л Љ М Н Њ О П Р С Т Ћ У Ф Х Ц Ч Џ Ш

Page: 1 2 3 4 (Next)

User photo	Name / Surname	Location	State	Last access ↑	Success	Choose
	Миловановић Милош	Beograd	Serbia	now	/	□
	Кезеле Бојан	Београд	Serbia	26 min 29 s	Excelent	□
	Тарлановић Тијана	Jagodina	Serbia	1 h 4 min	Not Excelent	□
	Мијајловић Милош	Novi Sad	Serbia	1 h 36 min	Not Excelent	□
	Пајић Бобан	Novi Sad	Serbia	2 h 59 min	Excelent	□
	Мићић Александар	Niš	Serbia	9 h 54 min	Excelent	□

Fig. 7. New Moodle report for prediction of each student

By selecting link “Cognitive style profiles on this course” (Figure 8) educator is provided with graphical presentation of student groups based on their cognitive styles and success (Figure 9).



Recent activity

Activity from monday, 14. november 2011., 13:20

Complete report on recent activity...

No news since your latest visit

Cognitive style & success

Cognitive style profiles on this course

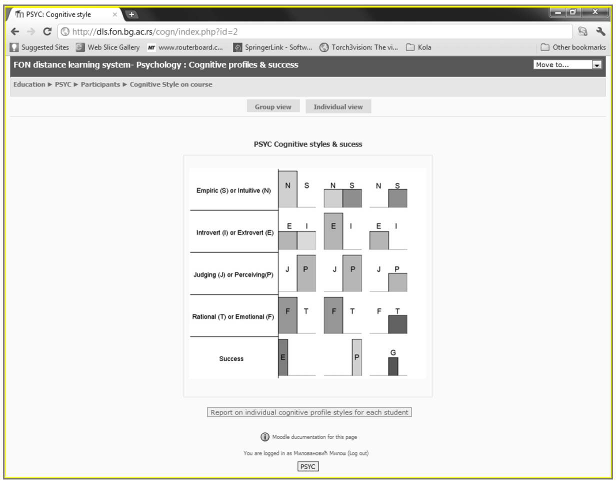
Predict students success on this course

Active users

(последних 5 минута)

Миловановић Милош

Fig. 8. Moodle block for accessing prediction



PSYC: Cognitive style
http://dls.fon.bg.ac.rs/cogn/index.php?id=2
Suggested Sites
Web Slice Gallery
www.routerboard.c...
SpringerLink - Softw...
Torch3vision: The vi...
Kola
Other bookmarks
Move to...
Group view
Individual view
Empiric (S) or Intuitive (N)
N
S
N
S
N
S
Introvert (I) or Extrovert (E)
E
I
E
I
E
I
Judging (J) or Perceiving(P)
J
P
J
P
J
P
Rational (T) or Emotional (F)
F
T
F
T
F
T
Success
E
P
G
Report on individual cognitive profile styles for each student
PSYC

FON distance learning system- Psychology : Cognitive profiles & success

Education ▶ PSYC ▶ Participants ▶ Cognitive Style on course

PSYC Cognitive styles & sucess

Moodle ducumentation for this page

You are logged in as Миловановић Милош (Log out)

Fig. 9. Moodle report for description of students’ cognitive style and the success they had on the course

Educator can then go further in analysis and track each student according to its expected success based on cognitive style. Educator can change his approach to the specific student and attempt to adapt the material to better fit that specific student and its profile.

4. Discussion and future work

Prediction of students’ success and grouping of students are common tasks in educational data mining and are valuable for educators as well for students. In this paper we presented a case study that involves these tasks for the web usage data gathered from University of Belgrade distance learning system.

We defined the automatic procedure for extraction of data from Moodle LMS and pre-processing them in to appropriate form for analysis with EDM algorithms.

Further we created classification models, accurate enough to predict if students will have an excellent performance on different courses based on web usage mining data. Valuable information can also be retrieved from students’ cognitive styles. Describing students with their cognitive styles seems natural in the educational context, and this research encourages further usage of this kind of data. So we built a clustering model that identifies the groups of students with similar cognitive styles and different success. Defined models are evaluated and used for construction of Moodle module that can help educators for two purposes: for distinction of students they can collaborate with or identification of students that need extra attention on that course, adaption of learning materials to better fit some specific cognitive styles or even recommend courses to students that better fit their cognitive style.

One problem we had to overcome in our study is the lack of data for more thorough analysis. This issue targets many distance learning institutions.[50] Usually, distance learning systems do not have many students enrolled which implies smaller number of potential participants for data mining research. As distance learning system was introduced at our University only several years ago, future usage of the system will allow more analysis, and verification of hypothesis tested in this paper. What is more, these early analysis of the e-learning system would benefit its faster advancement towards maturity, and offer all participants functionalities that make introducing such a system worthwhile.

In future work we plan to evaluate more classification and clustering algorithms in order to make even better fitting of models to web usage data.[51,52,53] For this purpose reusable component based algorithms could also be used.[54,55,56,57,58] Additionally, enriching the student data with even more descriptors (e.g. data gathered through social network analysis) of their behavior on the educational system is definitely a worthy investment. Specifically, informal learning becomes more and more important because learning can happen anywhere at any time and analysis of informal learning data in distance learning systems provides a growing research area.[59] This will open a true potential for analysis of student behavior, more than has ever been possible in the traditional learning context.

Acknowledgements

This research is partially funded by a grant from the Serbian Ministry of Science and Technological Development, project ID III 47003.

This research is partially funded with support from the European Commission through Life Long Learning Programme project TRAILER No. 519141-LLP-1-2011-1-ES-KA3-KA3MP.

This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

1. Y-C Lee, N. Terashima, A Distance Instructional System with Learning Performance Evaluation Mechanism: Moodle-Based Educational System Design, Distance Education Technologies 10 (2) (2012). doi: 10.4018/jdet.2012040104
2. T. Martin-Blas, A. Serano-Fernandez, The role of new technologies in the learning process: Moodle as a teaching tool in Physics, Computers & Education 52 (2009) pp. 35-44. doi:10.1016/j.compedu.2008.06.005
3. I. Kazanidis, S. Valsamidis, T. Theodosiou and S. Kontogiannis, Proposed framework for data mining in e-learning: The case of Open e-Class, in Proc. *IADIS International Conference of Applied Computing,* (Rome, Italy, 2009), pp. 254–258.
4. F. J. García-Peñalvo, M. Á.Conde, M. Alier, María J. Casany, Opening Learning Management Systems to Personal Learning Environments, *Journal of Universal Computer Science* 17(9)(2011), pp. 1222-1240.
5. A. J. Berlanga, F. J. García-Peñalvo, P. B. Sloep, Towards eLearning 2.0 University. *Interactive Learning Environments* 18 (3) (2010), pp. 199-201.
6. C. Romero, S. Ventura and E. García, Data mining in course management systems: moodle case study and tutorial, *Comput. Educ.* 51(1) (2008) 368–384.
7. V. Kumar, An Empirical Study of the Applications of Data Mining Techniques in Higher Education, *International Journal of Advanced Computer Science and Applications*, *2*(3) (2011) 80-84.
8. V.Ramesh, P.Parkavi, P.Yasodha, Performance Analysis of Data Mining Techniques for Placement Chance Prediction, International Journal of Scientific & Engineering Research 2 (8) (2011) pp. 1-7.
9. F. Castro, A. Vellido, À. Nebot and F. Mugica, Applying data mining techniques to e-learning problems. Evolution of teaching and learning paradigms in intelligent environment, 62 (2007) pp. 183–221.
10. A. C. Romero, and A. S. Ventura, Educational data mining: A survey from 1995 to 2005, *Journal of Expert Systems Applications*, 33(1) (2007) 135-146.
11. C-H Weng, Mining fuzzy specific rare itemsets for education data, *Knowledge-Based Systems* 24 (5) (2011) pp. 697-708.
12. C. Romero and S. Ventura, Educational data mining: a review of the state-of-the-art, *IEEE Trans. Syst. Man Cybernet. C Appl. Rev.,* 40(6) (2011) 601–618.
13. A. Krueger, A. Merceron and B. Wolf, A Data Model to Ease Analysis and Mining of Educational Data, in Proc. *Third International Conference on Educational Data Mining*, (USA, Pittsburgh, 2010) pp. 131-140.
14. Y-H Wang, H-C Liao, Data mining for adaptive learning in a TESL-based e-learning, Expert Systems with Applications 38 (6) (2011), pp. 6480-6485.
15. V.Ramesh, P.Parkavi, P.Yasodha, Performance Analysis of Data Mining Techniques for Placement Chance Prediction, *International Journal of Scientific & Engineering Research* 2 (8) (2011).
16. C. Vialardi, J. Chue, J.P. Peche, G. Alvarado, B. Vinatea, J. Estrella and Á. Ortigosa, A data mining approach to guide students through the enrollment process based on academic performance, User modeling and user-adapted interaction 21 (1-2) (2011), pp. 217-248. doi: 10.1007/s11257-011-9098-4.
17. S. Kotsiantis, K. Patriarcheas and M. Xenos, A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education, *Knowledge-Based Systems*, 23(6) (2010) 529-535.
18. K. Kuk, P. Spalevic, S. Ilic, M. Caric, Z. Trajcevski, A Model for Student Knowledge Diagnosis through Game Learning Environment, Technics Technologies Education Management – TTEM, 7 (1) (2012) 103-110.
19. S. Kotsiantis, Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades, *Artificial Intelligence Review*, (Online First) (2011) 1-14.

20. N. Myller, J. Suhonen and E. Sutinen, Using Data Mining for Improving Web-Based Course Design, in Proc. *International Conference on Computers in Education*, (USA, Washington, 2002) pp. 959- 964.
21. D. Traynor and J.P. Gibson, Synthesis and Analysis of Automatic Assessment Methods in CS1, in Proc. *The 36th SIGCSE Technical Symposium on Computer Science Education SIGCSE'05*, (ACM Press., Louis Missouri, USA , 2005) pp. 495-499.
22. B. Minaei-bidgoli, D. A. Kashy, G. Kortmeyer and W. F. Punch, Predicting student performance: an application of data mining methods with an educational Web-based system, in Proc. *33rd International Conference on Frontiers in Education*, (Colorado, Westminister, 2003) pp. 13-18.
23. M. Delgado, E. Gibaja, M.C. Pegalajar and O. Pérez, (2006). Predicting Students' Marks from. Moodle Logs using Neural Network Models, in Proc. *International Conference on Current Developments in Technology Assisted Education*, (Sevilla, Spain, 2006) pp. 586-590.
24. F.A. Bachtiar, W.E. Cooper, K.K. Kamei, Student grouping by neural network based on affective factors in learning English in Proc. International Conference on e-Education, Entertainment and e-Management (ICEEE), 2011. doi: 10.1109/ICeEEM.2011.6137792.
25. A. Drigas, J. Vrettaros, An Intelligent Tool for Building e-Learning Contend-Material Using Natural Language in Digital Libraries. WSEAS Transactions on Information Science and Applications 5(1) (2004) 1197-1205.
26. K. Hammouda, M. Kamel, Data Mining in e-Learning. In: Pierre, S. (ed.): e-Learning Networked Environments and Architectures: A Knowledge Processing Perspective. Springer-Verlag, Berlin Heidelberg New York (2005).
27. J. Tane, C. Schmitz, G. Stumme, Semantic Resource Management for the Web: An e-Learning Application. In: Fieldman, S., Uretsky, M. (eds.): The 13th World Wide Web Conference 2004, WWW2004. ACM Press, New York (2004) pp. 1-10
28. S. Ayesha, T. Mustafa, A.R. Sattar, M.I. Khan, Data Mining Model for Higher Education System, Europen Journal of Scientific Research 43(1)(2010), pp.24-29.
29. E. Ayers, R. Nugent, and N. Dean, A Comparison of Student Skill Knowledge Estimates. in Proc. *International Conference On Educational Data Mining*, (Cordoba, Spain, 2009), pp. 1-10.
30. D. Zakrzewska, Cluster analysis for user's modeling in intelligent e-learning systems, in Proc. In International Conference on Industrial, *Engineering & Other Applications of Applied Intelligent Systems: New Frontiers in Applied Artificial Intelligence (IEA/AIE '08*, eds. N. T. Nguyen, L. Borzemski, A.Grzech, and M. Ali, (Springer-Verlag, Berlin, Heidelberg, 2008) pp. 209-214.
31. R.M. Felder and L.K. Silverman, Learning and teaching styles in engineering education, *Eng. Educ.*, 78 (7) (1988) 674–681.
32. A. Heikkila, N. Markku, J. Nieminen, K. Lonka, Interrelations among university students' approaches to learning, regulation of learning, and cognitive and attributional strategies: a person oriented approach, *High Educ* 61 (2011), pp. 513–529. doi: 10.1007/s10734-010-9346-2
33. D. Perera, J. Kay, I. Koprinska, K. Yacef and O. R. Zaïane, Clustering and Sequential Pattern Mining of Online Collaborative Learning Data, *IEEE Transaction on Knowledge and Data Engineering*, 21 (6) (2009), pp. 759-772.
34. S.Y. Chen and X. Liu, Mining students' learning patterns and performance in Web-based instruction: a cognitive style approach, *Interactive Learning Environments* 19 (2) (2011). doi:10.1080/10494820802667256
35. J.M. Adán-Coello C.M. Tobar E.S.J. de Faria, W.S de Menezes, R.L. de Freitas, Forming Groups for Collaborative Learning of Introductory Computer Programming Based on Students' Programming Skills and Learning Styles, *International Journal of Information and Communication Technology Education* 7 (4) (2011). doi: 10.4018/jicte.2011100104
36. S. Jevremovic, Implementation of the adaptive system in electronic learning, Management 14 (53) (2009), pp.57-61.
37. C-M. Chen, M-C. Chen, Mobile formative assessment tool based on data mining techniques for supporting web-based learning, *Computers & Education* 52 (2009), pp. 256–273. doi:10.1016/j.compedu.2008.08.005
38. I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz and T. Euler, YALE: Rapid prototyping for complex data mining tasks, in Proc. *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (Philadelphia, USA, ACM Press, 2006) pp. 935-940.
39. J. Cardoso, Developing Course Management Systems Using The Semantic Web, The Semantic Web, Semantic Web and Beyond, 2008, Volume 6, Part IV, 169-188. doi: 10.1007/978-0-387-48531-7_8
40. M. Minović, M. Milovanović, I. Kovačević, J. Minović and D. Starčević, Game design as a learning tool for the course of Computer Networks, *International Journal of Engineering Education*, 27(3) (2011) 498 - 508.
41. I.B. Myers and M.H. McCaulley, *Manual: a guide to the development and use of the Myers±Briggs Type Indicator*, (*Consulting Psychologists Press*, Palo Alto, CA, 1985).
42. I.B. Myers, M.H. McCaulley, N.L. Quenk and A.L. Hammer, *MBTI Manual. A guide to the Development and Use of Myers-Briggs Type Indicator*, (Consulting Psychologists Press, Palo Alto, CA, 1998).
43. C.G. Jung, *Psychological Types. The Collected Works, vol. 6.*, (Routledge and Kegan Paul, London, UK, 1971)
44. J.G. Carlson, Resent Assessments of Myers-Briggs Type Indicator, *Journal of Personality Assessment*, 49(4) (1985) 356-365.
45. R. Colomo Palacios, E. Tovar Caro, A. García Crespo, & J.M. Gómez Berbís, Identifying Technical Competences of IT Professionals: The Case of Software Engineers, *International Journal of Human Capital and Information Technology Professionals* 1(1) (2010), pp. 31-43.
46. R. Colomo-Palacios, E. Fernandes, P. Soto-Acosta & M. Sabbagh, M, Software product evolution for Intellectual

Capital Management: The case of Meta4 PeopleNet, *International Journal of Information Management* 31(4) (2011), pp. 395-399.

47. A. García-Crespo, R. Colomo-Palacios, J.M Gómez-Berbís, & M. Mencke, M,. BMR: Benchmarking Metrics Recommender for Personnel issues in Software Development Projects. *International Journal of Computational Intelligence Systems* 2(3) (2009), pp. 257-267.
48. S. Westlund, Leading Techies: Assessing Project Leadership Styles Most Significantly Related to Software Developer Job Satisfaction. *International Journal of Human Capital and Information Technology Professionals* 2(2) (2011), pp. 1-15. doi:10.4018/jhcitp.2011040101
49. O.C.S. Tzeng, S.L. Ware, J-M. Chen, Measurement and Utility of Continuous Unipolar Ratings for the Myer-Briggs Type Indicator, *Journal of Personality Assessment*, 53(4) (1989) 727-738.
50. C. Romero, S. Ventura, P. G. Espejo and C. Hervs, Data Mining Algorithms to Classify Students, in Proc. *1st International Conference on Educational Data Mining (EDM'08)*, (Montreal, Canada, 2008) pp. 8–17.
51. P. Lingras, M. Joshi, Experimental Comparison of Iterative Versus Evolutionary Crisp and Rough Clustering, *International Journal of Computational Intelligence Systems*, 4(1)(2011), pp.12-28.
52. Y.-C. Lin, T.-K. Wu, S.-C. Huang, Y.-R. Meng, W.-Y. Liang, Rough Sets as a Knowledge Discovery and Classification Tool for the Diagnosis of Students with Learning Disabilities, *International Journal of Computational Intelligence Systems*, 4(1) (2011), pp.29-43.
53. M. Matijaš, M. Vukićević, S. Krajcar, Supplier Short Term Load Forecasting Using Support Vector Regression and Exogenous Input, *Journal of Electrical Engineering* 62(5)(2011) pp. 280-285. doi:10.2478/v10187-011-0044-9
54. B. Delibašić, M. Jovanović, M. Vukićević, M. Suknović, Z. Obradović, Component-based decision trees for classification, *Intelligent Data Analysis* 15 (5) (2011) pp. 671-693. doi: 10.3233/IDA-2011-0489
55. M. Suknovic, B. Delibasic, M. Jovanovic, M. Vukicevic, D. Becajski-Vujaklija and Z. Obradovic, Reusable components in decision trees induction algorithms, Computational Statistics (2012). doi:10.1007/s00180-011-0242-8.
56. B. Delibasic, K. Kirchner, J. Ruhland, M. Jovanovic, M. Vukicevic, Reusable components for partitioning clustering algorithms. Artificial Intelligence Review 32 (1-4) (2009) pp. 59-75. doi: 10.1007/s10462-009-9133-6
57. M. Vukicevic, M. Jovanovic, B. Delibasic, S. Isljamovic, M. Suknovic, Reusable component-based architecture for decision tree algorithm design, *International Journal on Artificial Intelligence Tools* (2012). doi: 10.1142/S0218213012500224
58. B. Delibasic, M. Vukicevic, M. Jovanovic, K. Kirchner, J. Ruhland, M. Suknovic, An architecture for component-based design of representative-based clustering algorithms, Data & Knowledge Engineering (2012). doi: 10.1016/j.datak.2012.03.005
59. B. Chen and T. Bryer, Investigating Instructional Strategies for Using Social Media in Formal and Informal Learning, *The International Review of Research in Open and Distance Learning*, ISSN: 1492-3831, 13 (1) (2012).