A Data Mining Approach to the Analysis of Students' Learning Styles in an e-Learning Community: A Case Study

Valentina Efrati[2], Carla Limongelli[1], and Filippo Sciarrone[1]

[1] Roma Tre University - Engineering Department
Via della Vasca Navale, 76
00146 Rome, Italy
{limongel,sciarro}@dia.uniroma3.it
[2] Roma Tre University - Fil.Co.Spe Department - Filosofia, Comunicazione e Spettacolo Via Ostiense, 234, 00144 Rome, Italy
valentina.efrati@uniroma3.it

Abstract. In recent years, there has been a radical change in the world of education and training that is causing that many schools, universities and companies are adopting the most modern technologies, mainly based on Web architectures and Web 2.0 instruments and tools, for learning, managing and sharing of knowledge. In this context, an e-Learning system can reach its maximum potential and effectiveness if it could take advantage of the information in its possession and process it in an intelligent and personalized way. The Educational Data Mining is an emergent field of research where the approach to personalization makes use of the log data generated by learners during their training process, to dynamically update users learning profiles such as skills and learning styles and identify students behavioral patterns. In this paper we present a case study of a data mining approach, based on cluster analysis, in order to support the detection of learning styles in a community of learners, following the Grasha-Riechmann learning styles model. As an e-learning framework we used the Moodle LMS platform and studied the log files generated by a course taken by a community of learners. The first experimental results suggest a connection between clusters and learning styles, reinforcing the use of this approach.

1 Introduction

Nowadays with the exponential growth of the Internet and the use of the Web 2.0 instruments and tools, distance education is more and more adopted by educational institutions and companies, producing a lot of data concerning learners behavioral patterns. Educational Data Mining (EDM) is an emerging discipline, concerned with the developments of methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in[1][24]. In particular,

[1] http://www.educationaldatamining.org/

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2014, Part II, LNCS 8514, pp. 289–300, 2014.

this approach is useful for personalized educational environments where one can use the data generated by learners during their training process to dynamically update their learning profiles, such as skills and learning styles. In this paper we present a case study of a data mining approach to the detection of the learning styles in a community of learners, following the Grasha-Riechmann Model (GRM) [11]. To this aim we based our study on an e-learning course delivered in by a Governative Italian institution [2] for training teachers to teach in high schools. The course was delivered using the Moodle e-learning environment [3], which is one of the most used open source e-learning platforms. We studied the navigational behaviors of learners through the analysis of the log files generated by their actions during the learning process, with the specific goal to predict the learning styles of each student, using a data-driven approach by means of the use of clustering algorithms. The goal of our work aimed to find a matching between the groups of students, i.e., the clusters, and the three dimensions of the GRM. A classic data mining process has been performed, using the Weka machine learning framework[4] while the log data files contained almost $1,500.000$ record of actions. Firstly we obtained 6 clusters of students and secondly we performed a mapping between the average values obtained for each attribute of each cluster and the values proposed by Grasha and Riechmann The experiment was conducted through the application of the clustering algorithms known as *Expectation Maximization* [12] (EM) applied to all the attributes identified in the Feature Selection phase. The mapping between the clusters and the learning styles was built taking into account, for each analyzed attribute, its average value in the cluster, as calculated by the EM algorithm. The effect of this experiment is that the learning characteristics that students belonging to the same cluster seem to possess, fit well with learning styles. The paper is structured as follows. In Section 2 some related literature and state of the art is shown while in Section 3 the pedagogical background is reported. In Section 4 we show the framework and the context of the work while in Section 5 the evaluation of the approach is reported and finally in Section 6 conclusions are drawn.

2 Related Work

One of the aims of EDM is to create user profiles automatically in order to improve the design of a course and its customization to be more responsive to the needs of the learner. On the other hand, learning environments can take advantage of the knowledge management techniques such as data mining [13], to collect and process disseminated information from exchanged e-mails and documents, from discussion forums and so on. EDM is the field of research of this work, where we address in a suitable case study the detection of learners learning styles. In the literature there are already many attempts of using data mining techniques in the e-learning field. In fact, most of e-learning platforms have a

[2] SSIS Lazio `http://www.ssis-lazio.it`
[3] `http://www.moodle.org`
[4] `http://www.cs.waikato.ac.nz/ml/weka/`

tracing mechanism storing the interactions between the learner and the learning environment in log files that subsequently can be analyzed to study the learners behaviors [23]. In [10], the authors use the Felder and Silvermann learning styles model [5] and do not use a data mining approach. Moreover they compare their results vs. the ILS questionnaire. In the work of Gaudioso and Tavalera [8], the application of data mining techniques to the virtual learning community aims to provide support for the evaluation of the course through the characterization of patterns concerning the performance of learners who then help to determine the profiles of weak learners and to identify and improve this type of behavior for future courses. In addition, they provide support to the identification and characterization of patterns of behaviors that can help to determine the different roles among learners and to manage groups in collaborative activities. An interesting framework has been developed by Mor et al. [22]. The purpose of the framework is the study of the navigational behavior of the students of an e-learning environment integrated in a virtual campus through the extraction of information that can be used to validate several aspects related to the design and usability of a virtual campus and also to determine the optimal scheduling for each course based on the profile of the student. To all the students of the *Foundations of programming 1* and *Compilers 1* was strongly advised to carry out certain activities. In the work of Wang are presented some models and methods to analyze log files in order to build a model of user behavior when browsing, that support the applications of e- Learning [28]. The teachers can investigate the model to identify some interesting or even more unexpected learning patterns in the navigational behavior of the learners and use this information to revise and reorganize the structure of content more effectively. To do this have been developed a set of tools based on data mining techniques such as clustering and association rules combined with collaborative filtering. Tang and McCalla [27] presented a system capable of suggesting to learners interested in a specific research area of scientific papers. Such a system in the domain of e-Learning requires special requirements that are not necessary in other domains. The most important requirement is the need to consider the pedagogical aspects of the learner and the need to organize the concepts actually taking into account these pedagogical issues; the result should be to maximize the utility that the learner obtains from the system gaining maximum knowledge keeping highly motivated. Other works aim to help teachers to produce and to deliver suitable didactic material to students [19,17,9,6], to propose personalization engines embedded into the Moodle platform [21,7,20,1] and also into real-work e-learning systems [18,15,14,16]. In a perspective of technology enhanced learning, there is research work [4,3,2] aiming to integrate more traditional individualized e-learning [26].

3 The Grasha-Riechmann Learning Styles Model

The GRM defines learning as a process of social nature, so also influenced by the different individual approaches to the environment class, by the relationship with peers and with the teacher. The GRM proposes the following three dimensions, each of which based on a three values scale: low, moderate and high:

- *Intra-subjectivity Vs. Inter-subjectivity*: the categories are related to the self-perception in relation to the environment and to the culture of belonging. In the context of learning, a intra-subjective personality prefers self-analysis, self-evaluation and also in cooperative contexts tends to bring out her own personal contribution vs. the group and to search for the contribution that the group can give to her personal growth. An inter-subjective student prefers socialization, knowledge of and learn through social mediation and knowledge sharing;
- *Competitive Vs. Collaborative*: the motivation for learning can be of a competitive nature and therefore connected to the need to stand out in the class, to receive awards, and to work individually, or collaborative character, and therefore related to the need to work with others, to share experiences, knowledge and tasks;
- *Independence Vs. Dependence*: a student may express a desire for autonomy and independence in relation to the rest of the class, the teacher and the task, preferring to work alone; on the contrary a dependent student feels the teacher as an authority, so strictly following her instructions and has a few perception of her educational autonomy.

Grasha and Riechmann have reviewed and evaluated the learning styles of college students through a social perspective in order to identify the different approaches to the environment of the classroom [11]. This learning style can be seen as a scale of social interactions because it is the way of interaction of the student with teachers and peers in a learning environment rather than the way they perceive and organize the information. Although these categories are not readily translated into learning strategies, are defined by three dimensions of class: the student's attitude towards learning, vision teacher and/or peer reactions to procedures performed in the classroom. Most individuals do not put themselves at the extremes of these bipolar styles, but indicate some degree of preference for each of these three categories, i.e., low, moderate and high. The scale intra-subjectivity Vs. inter-subjectivity measures how much an individual wants to be involved in the class, as it reacts with the procedures of the class and its attitudes towards learning. The scale collaborative/competitive measures the basic motivation of an individual's interaction with others while the third, independent/dependent, measures the attitude of the student towards teachers and how much she prefers freedom or control in the learning environment.

4 The Framework

In this Section we describe the framework used for analyzing the behavior of the learner in an environment of e-Learning to try to predict a profile for her learning styles. We used the Moodle LMS e-learning environment to log data form several learners actions and activities. After we used the Weka machine learning environment to run some clustering algorithms by which to support the inference of the learner's learning styles.

4.1 The E-learning Environment

The e-Learning environment used as a data source is the Moodle platform, the most used open source e-learning environment. Moodle is a software package for producing Internet-based courses and web sites. It is a global development project designed to support a social constructionist framework of education. In particular it is a web application, using MySql as the database where all the learners actions are stored in log files. We used a course delivered to teacher trainees by S.S.I.S. Lazio, a public training organization for teachers. In Fig. 1 the home page of the course is shown. Learners can participate to many didactic activities such as read lessons, interact with the tutor or exchange knowledge with peers in forums, chat, wiki pages and so on. All these actions are stored in some log data.

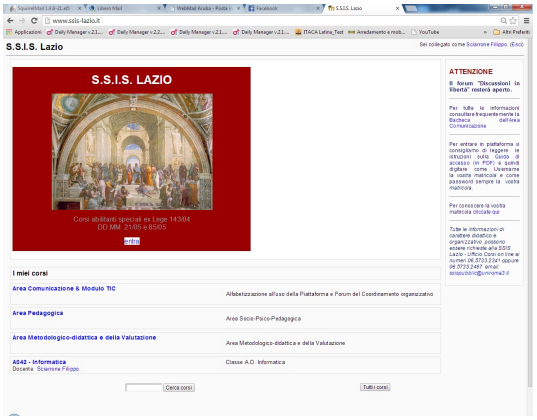

Fig. 1. The Home Page of the Moodle Environment

4.2 The Weka Platform

As a data mining environment, we selected the Weka platform, the most used open source machine learning environment. It is a collection of machine learning algorithms for data mining tasks where the algorithms can either be applied directly to a data set or launched from our own Java code. This platform contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. It is also well suited for developing new machine learning schemas and comprises many state-of-the-art algorithms for supervised and unsupervised learning, providing an easy-to-use framework for performing experimental comparisons among different machine learning outcomes. The easiest way to use Weka is through a graphical interface called *the Explorer*. There are two other graphical user interfaces in Weka: *The Knowledge Flow* interface allows for designing configurations for streamed data processing while *The*

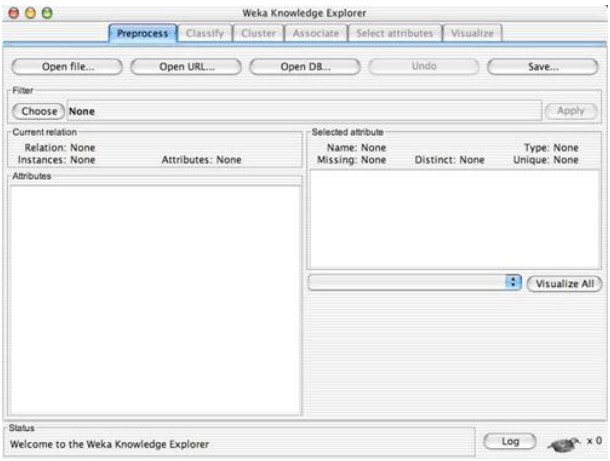

Fig. 2. The Weka Machine Learning Environment

Experimenter is designed to help to answer to basic practical questions when applying classification and regression techniques, as shown in Fig. 2. In *The Explorer* interface, there are six different panels, selected by the tabs at the top, corresponding to the various data mining tasks: the six tabs are the basic operations that *The Explorer* supports. The *Preprocess* panel is the starting point for knowledge exploration. From this panel we can load data sets, browse the characteristics of attributes and apply any combination of Weka's unsupervised filters to the data.

5 Evaluation

In this Section we show the evaluation of a data mining approach to the building of learners learning styles, according to the Grasha-Riemann learning styles model. More specifically, the research question to validate is if clustering the huge amount of source data according to some relevant features, we obtain groups of similar learners according to the Grasha-Riemann learning styles model.

5.1 The Overall Data Mining Process

When students navigate through the services provided by the Moodle environment, they leave traces that can be analyzed a posteriori with the purpose of building or integrating a student model. Most of this information is stored by the server in the form of web server log files and the analysis of these files may be of interest to obtain information about significative learning patterns. For every student action, accomplished in the learning environment a row is added to the system log file. This log file must first be pre-processed to remove all those log lines that are definitely not hits produced by the student. This step greatly

reduces the amount of lines of log. Consequently, we used a classic data mining process, according to the following schema [12,25]:

1. *Problem Definition.* Clear definition of the problem and of the goals one wants to achieve;
2. *Feature selection.* What are the relevant dimensions to select for the study?
3. *Data Gathering and Preparation.* In this phase all data are gathered and transformed in order to be elaborated by the data mining algorithms selected;
4. *Model Building and Evaluation.* Here the bottom-up models produced by the data mining algorithms are evaluated in order to asses their validity with respect to the main goals;
5. *Discussion.* Here the results of the mapping between the clusters and the GRM is discussed.

5.2 Problem Definition

One of the greatest problems related to the design of online learning environments is to extract meaningful information about the actions of the students, their behavior, their way of navigating, that is about their use of the system. It is hard to monitor what the student actually does and what it is expected to be able to do and above all to represent this in the form of behavioral and navigational patterns. Often these models are used and are very useful to determine the degree of quality of the design of the learning environment and to measure the degree of association between the requirements of usability and the navigational behavior of the students. However, finding a correlation between behavioral patterns and pedagogical issues like learning styles is a very important step to understand learner's personality. It is clear that the extraction of patterns of behavior and learning must not be used in any way with the purpose of spying the student to obtain data used for other purposes but only to ensure a satisfactory process of learning and training.

5.3 Feature Selection

The log data files generated by the learners activities contain a large set of instances and one needs to select a subset of them in order to reduce the space dimensionality from a very large set, and consequently intractable, to a reasonable tractable set. In literature this is the classic *curse of dimensionality* problem. After a deep heuristic analysis of the variables stored by Moodle in the log data, we selected the following features, as relevant for Grasha-Riemann model:

- *allresourceview.* It indicates the total number of accesses for each student to all resources both theoretical and practical. In the Moodle environment, a resource can be defined as every didactic material accessible to learners;
- *resourcetheoryview.* This feature indicates the total number of accesses of each student to theoretical didactic materials.

- *totmoveforum*. It indicates the total number of accesses for each student carried out to all kinds of forums in both reading and posting;
- *visiteforum*. It indicates the total number of accesses for each student carried out to all kinds of forum in reading only modality;
- *useforum*. It indicates the total number of messages posted by each student to all kinds of forum;
- *totmoveforumdiscipl*. It indicates the total number of accesses for each student to all disciplinary forum;
- *totmoveforumgeneral*. It indicates the total number of accesses for each student to organizational or administrative forum.
- *extra*. It indicates the total number of accesses of learners to complementary didactic resources.

Basically, the various clusters computed basing on the above features should represent different groups of students with different ways of approaching the study, i.e., learning styles. From the log data, we produced a text file containing all those user records extracted from the original log data.

5.4 Data Gathering and Preparation

Our data source was composed by 1.500.000 records. In Fig. 3 a screenshot of it is shown. The learners that took the course were 1854 over one year. Starting from this large file, we preprocessed it, forming a text file containing features instances only. It was obtained through a multiple join among all the tables created for each feature of interest, and stored using a vector of student objects.

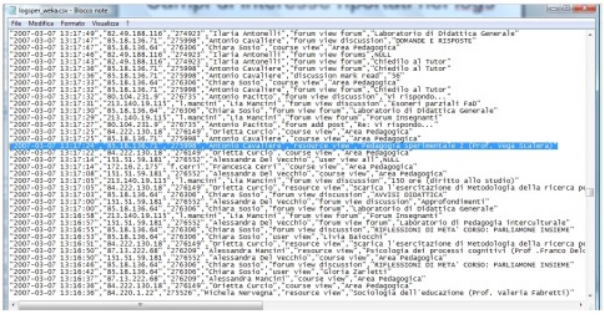

Fig. 3. A Screenshot of an Excerpt of the Log File

5.5 Model Building and Evaluation

As data mining model, we used the clustering unsupervised learning, a very useful technique to find out the distribution of the input data, allowing for a meaningful partitioning of the input data set. To perform such an analysis we

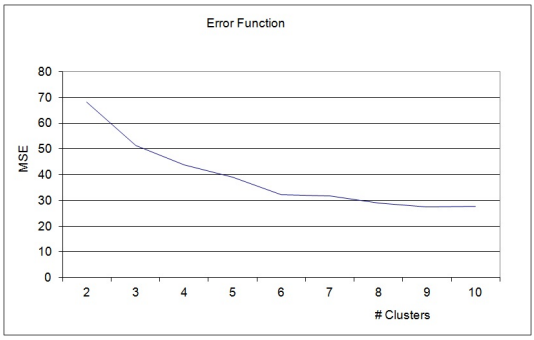

Fig. 4. The Clustering Process

used the Weka open source tool which has been fully reported in section 4. The experiment was conducted through the application of the EM algorithm of clustering applied to all the attributes identified in the previous preparation phase and stored in the user records. To estimate the right number of clusters, we used the Cross Validation technique, which consists in partitioning the training data into K distinct sub-sets, the training is done using $K-1$ subset and the network test on the remaining subsets, i.e., K-fold cross-validation. The process of training and test is repeated for each of the K possible choices of the subsets omitted from the training. The average performance on the K subsets is omitted in the estimation of the generalization performance. This procedure has the advantage of using a large portion of the data available for training and all data to estimate the error of generalization. The disadvantage is the need to train the network K times. In fact, the processing time of the EM algorithm with large data sets is quite high. From Fig. 4 we see the MSE error in the training phase. We stopped the algorithm at six clusters because this number corresponds to the six Grasha-Riemann dimensions and also because the error trend after six clusters did not present relevant changes.

5.6 Results and Discussion

In Tab. 2 the results are shown. In particular, the first three columns report, for each cluster C_i the number of students both in absolute values and in percentage with respect to the overall sample. The last three columns report the correspondent Grasha-Riechmann learning styles. This mapping has been performed at hand, reasoning on the characteristics of the features belonging to a cluster compared with those belonging to the GRM. For the sake of brevity we report here the mapping process for cluster C_3, i.e., for the cluster with the great number of learners. The characteristics of this cluster are shown in Tab. 1. This is the case of a student that has visited general forums (totmoveforum and visiteforum) moderately, has visited moderately practical and theoretical resources (allresourceview, resourcepracticalview, resourcetheoryview), has used very few times the administrative forum (useforum) and moderately disciplinary

Table 1. The Characteristics of the Cluster 3

Feature	Mean
allresourceview	47,391
resourcepracticalview	78,52
resourcetheoryview	29,35
extra	19,45
totmoveforum	113,47
visiteforum	118,82
useforum	1,64
totmoveforumdiscipl	15,08
totmoveforumgeneral	35,95

Table 2. The Results of the Clustering EM Algorithm

Cluster	# learners	%	Independent	Competitive	Intra-subjective
C0	218	11,76%	Low	Moderate	Low
C1	185	9,98%	Moderate	Moderate	Moderate
C2	377	20,33%	Low	Low	Low
C3	811	43,74%	Moderate	Moderate	Moderate
C4	176	9,49%	High	High	High
C5	87	4,69%	High	High	High

and general forums (totmoveforumdiscipl and totmoveforumgeneral). Finally she did some visits to extra materials. So this student is a classic moderate student, according to the GR model because she prefers a moderate social activity, a moderate competition among peers, i.e., she visits all resources and finally she prefers practical aspects.

6 Conclusions

In this paper we presented a case study of the use of a data mining approach to the detection of the learning styles of a community of learners, according to the GR model, to explore the effectiveness of such an approach. We addressed a classic unsupervised learning classification problem performed through the Weka data mining platform, running the EM clustering algorithms on our data set. After we verified a mapping among each cluster and the learning styles. Reasoning on the characteristic of each cluster, represented by its own centroid, we performed such a mapping at hand, with encouraging results. The overall process can be completely automatized in the future, e.g. giving some threshold ranges to all the means of each cluster. As a result the student model can be dynamically and automatically enriched over time. As a future work we plan first to completely automatize the overall process and second to test the method with other learning styles models.

References

1. Biancalana, C., Micarelli, A.: Social tagging in query expansion: A new way for personalized web search. CSE (4), 1060–1065 (2009)
2. De Marsico, M., Sterbini, A., Temperini, M.: The definition of a tunneling strategy between adaptive learning and reputation-based group activities. In: Proc. 11th IEEE Int. Conf. on Advanced Learning Technologies, ICALT, pp. 498–500 (2011)
3. De Marsico, M., Sterbini, A., Temperini, M.: A strategy to join adaptive and reputation-based social-collaborative e-learning, through the zone of proximal development. Int. Journal of Distance Education Technologies 19(2), 105–121 (2012)
4. De Marsico, M., Sterbini, A., Temperini, M.: A framework to support social-collaborative personalized e-learning. In: Kurosu, M. (ed.) HCII/HCI 2013, Part II. LNCS, vol. 8005, pp. 351–360. Springer, Heidelberg (2013)
5. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering education. Engineering Education 78(7), 674–681 (1988)
6. Gasparetti, F., Micarelli, A., Sansonetti, G.: Exploiting web browsing activities for user needs identification. In: Proceedings of the 2014 International Conference on Computational Science and Computational Intelligence (CSCI 2014), IEEE Computer Society, Conference Publishing Services (March 2014)
7. Gasparetti, F., Micarelli, A., Sciarrone, F.: A web-based training system for business letter writing. Knowledgel-Based Systems 22(4), 287–291 (2009)
8. Gaudioso, E., Talavera, L.: Data mining to support tutoring in virtual learning communities: experiences and challenges. In: Romero, C., Ventura, S. (eds.) Data Mining in E-Learning. ch.12, pp. 207–225. WIT Press (2006)
9. Gentili, G., Micarelli, A., Sciarrone, F.: Infoweb: An adaptive information filtering system for the cultural heritage domain. Applied Artificial Intelligence 17(8-9), 715–744 (2003)
10. Graf, S., Kinshuk, Liu, T.: Supporting teachers in identifying students' learning styles in learning management systems: An automatic student modelling approach. Educational Technology & Society 12(4), 3–14 (2009)
11. Grasha, A.: Observations on relating teaching goals to student response styles and classroom methods. American Psychologist 27, 144–147 (1972)
12. Hand, D., Manila, H., Smith, P.: Principles of Data Mining. MIT Press (2001)
13. Hanna, M.: Data mining in the e-learning domain. Campus-Wide Information Systems 21(1), 29–34 (2004)
14. Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F.: A teacher model to speed up the process of building courses. In: Human-Computer Interaction. Applications and Services - 15th International Conference, HCI International 2013, Proceedings, Part II, Las Vegas, NV, USA, July 21-26, pp. 434–443 (2013)
15. Limongelli, C., Lombardi, M., Marani, A., Sciarrone, F.: A teaching-style based social network for didactic building and sharing. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 774–777. Springer, Heidelberg (2013)
16. Limongelli, C., Mosiello, G., Panzieri, S., Sciarrone, F.: Virtual industrial training: Joining innovative interfaces with plant modeling. In: ITHET, pp. 1–6. IEEE (2012)
17. Limongelli, C., Sciarrone, F., Starace, P., Temperini, M.: An ontology-driven olap system to help teachers in the analysis of web learning object repositories. Information System Management 27(3), 198–206 (2010)

18. Limongelli, C., Sciarrone, F., Temperini, M., Vaste, G.: Lecomps5: A web-based learning system for course personalization and adaptation. In: Proceedings of IADIS 2008, Proceedings, Amsterdam, The Netherlands, July 22-25, pp. 325–332 (2008)
19. Limongelli, C., Sciarrone, F., Temperini, M., Vaste, G.: The lecomps5 framework for personalized web-based learning: A teacher's satisfaction perspective. Computers in Human Behavior 27(4), 1310–1320 (2011)
20. Limongelli, C., Sciarrone, F., Vaste, G.: LS-PLAN: An effective combination of dynamic courseware generation and learning styles in web-based education. In: Nejdl, W., Kay, J., Pu, P., Herder, E. (eds.) AH 2008. LNCS, vol. 5149, pp. 133–142. Springer, Heidelberg (2008)
21. Limongelli, C., Sciarrone, F., Vaste, G.: Personalized e-learning in moodle: The moodle-ls system. Journal of E-Learning and Knowledge Society 7(1), 49–58 (2011)
22. More, E., Minguillón, J., Carbó, J.M.: Analysis of user navigational behavior for e-learning personalization. In: Romero, C., Ventura, S. (eds.) Data Mining in E-Learning. ch.13, pp. 227–243. WIT Press (2006)
23. Pahl, C.: Data mining for the analysis of content interaction in web-based learning and training systems. In: Romero, C., Ventura, S. (eds.) Data Mining in E-Learning. ch. 3, pp. 41–56. WIT Press (2006)
24. Romero, C., Ventura, S.: Educational data mining: A survey from 1995 to 2005. Expert Syst. Appl. 33(1), 135–146 (2007)
25. Sciarrone, F.: An extension of the q diversity metric for information processing in multiple classifier systems: a field evaluation. International Journal of Wavelets, Multiresolution and Information Processing IJWMIP 11(6) (2013)
26. Sterbini, A., Temperini, M.: Selection and sequencing constraints for personalized courses. In: Proc. IEEE Frontiers in Education, FIE, pp. T2C1–T2C6 (2010)
27. Tang, T.Y., McCalla, G.: Active, context-dependent, data centered techniques for e-learning: a case study of a research paper recommender system. In: Romero, C., Ventura, S. (eds.) Data Mining in E-Learning. ch. 5, pp. 207–225. WIT Press (2006)
28. Wang, F.: On using data mining for browsing log analysis in learning environments. In: Romero, C., Ventura, S. (eds.) Data Mining in E-Learning. ch. 4, pp. 57–73. WIT Press (2006)