The world of MOOCs represents a new trend for Web-based instruction. The MOOCs provide high quality disciplinary contents and new solutions to make high education accessible for masses. They have also some drawbacks causing negative phenomena like decreasing engagement and motivation and, thus, high drop-out rates. This work proposes a novel approach to mitigate the aforementioned drawbacks by means of the synergistic application of Educational Data Mining and Semantic Web techniques. The proposed approach focuses on the analysis of data related to students’ characteristics and behaviours with respect to learning object metadata and subject ontologies (conceptualizing the disciplinary domain of MOOC) in order to detect anomalies in the selection and/or definition of some characteristics (i.e. Learning Resource Type, Interactivity Type/Level, etc.) of MOOC learning contents. The detected anomalies are then used to generate feedbacks for instructors who can adapt the MOOC learning material applying a suitable adaptation strategy.

**KEYWORDS**: mass education, distance study, curriculum, new technologies, differentiated teaching, remedial instruction sciences.
Introduction and Motivations

The MOOCs (Massive Open Online Courses) implement a new model of Web-based instruction, characterized by educational settings based on high quality disciplinary contents provided by international renowned Universities (Liyanagunawardena, Adams & Williams, 2013; Mangione, 2013a). These contents are open and accessible to thousands (massive) of enrolled students (Yuan & Powell, 2013).

The rise of MOOCs seems to be the University answer to new emerging trends in online education. On the one hand, MOOCs foster online course accessibility for masses, overcoming situations where economic difficulties become a barrier for education (Thrift, 2013). On the other hand, with respect to the concept of employability, they can represent a valid tool for the development of capabilities that can be concretely exploited at workplace.

Taking into account the phenomenon of MOOCs, the Universities have to: i) define new modalities by which specialized learning curricula are delivered, ii) design suitable educational practices (Yuan & Powell, 2013) in order to valorise massive models, flexible curriculum and personalized learning paths and iii) overcome the recognized issues (above mentioned) associated with MOOC solutions. Among these issues, high students’ drop-out rate is a very important aspect. Drop-out are mostly due to a lack of constant, individualized and personalized formative feedbacks from instructors (which is impracticable due to large numbers of enrolled students). Additional issues to consider are related to motivation and engagement of students.

The themes of personalization and adaptation (Park & Lee, 2003; Mangione, 2013b) clearly emerge from our previous considerations. These themes, in massive learning environments, are even more critical when the composition of classes is highly heterogeneous and includes students with differences in terms, for instance, of age, culture and interests. These students are supposed to react in different manners to the same learning content (Glaser, 1977).

A traditional MOOC delivers a set of video lectures (typically organized in weeks), a set of quizzes and exams (Mangione, 2013a). Additionally, forum, blog, social networks and wiki are provided to enable other educational scenarios like, for instances, FAQs, Collaborative Problem Solving, Question Answering, just to mention a few (Kolowich, 2013).

Indeed, with the aim of facing personalization and adaptation issues, the concept of Adaptive MOOC has been introduced in a recent taxonomy proposed by Clark (2013). The adaptive MOOCs use specific algorithms allowing definition and delivery of personalized and adaptive learning experiences based on students’ characteristics and behaviours (mainly obtained by gathering and elaborating assessment results and other users’ actions on the learning environment and its components). Sometimes, poor students’ performances are due to a poor quality of learning contents (Anderson, 2013) or simply to a high rate of difficulty of understanding these contents with respect to either proficiency level or pre-existing knowledge of students. In these cases, it could be important to detect anomalies related to the learning contents (e.g. low quality, erroneous characteristics like, for instance, learning resource type, interactivity type and so on) in order to enable instructors to adapt or adjust the materials (Williams, 2013).

A number of Educational Data Mining (EDM) techniques like, for instance, clus-
tering, classification and association rules generation (Romero & Ventura, 2010), can be used to detect such content-related anomalies and sustain the generation of feedbacks for the instructors to foster the so called adapted guidance (Vogt & Rogalla, 2009). In order to define effective adaptation strategies, the results of the EDM algorithms have to be further interpreted. The interpretation phase cannot be performed in an automatic way and requires specific competencies (related to Data Mining algorithms, practices and tools) that, typically, are not part of the instructors expertise (except for those owning Data Mining competencies) (Mangione, 2013b).

The main idea underlying this work concerns with the definition of an overall approach, to assisting the instructors with making informed decisions about adaptation in MOOCs, based on Semantic Technologies and Methodologies (with particular reference to Semantic Web languages and vocabularies) in order to break down the barriers related to the use of EDM tools. More in details, we propose to use, in synergy, EDM algorithms and Semantic Web-based knowledge representation to generate enriched feedbacks which alert the instructors on anomalies in learning contents presented in a MOOC. In this work, the proposed approach is focused on detecting if a given learning content is inadequate (e.g. erroneous interactivity type/level, learning resource type, etc.) for students with specific characteristics and suggesting the instructors to define and apply suitable adaptation actions.

1. Educational Data Mining and Semantic Web in Education

1.1 Educational Data Mining

Educational Data Mining is a promising research field focused on supporting educational processes by using information and communication technologies. The EDM can be considered as the application of Data Mining (DM) techniques on educational data (coming from educational environments like, for instance, learning management systems, adaptive learning environments, traditional classes, and so on) in order to improve learning and teaching processes (Mangione, 2013b). Authors in (Romero & Ventura, 2013), also state that the EDM is committed to carry out development, research and application of computational methods to elicit patterns from big sets of educational data. A typical EDM process is reported in Figure 1, where the main activities are the following:

![Figure 1. Traditional EDM Process](image-url)
• **Data gathering.** The availability of several and heterogeneous data sources (associated with big numbers of different educational environments possible to considered) determine a high level of heterogeneity of data formats and structures that have to be used in knowledge extraction process.

• **Pre-processing.** Data collected in the first phase, typically, are not represented in a form that is compatible with requirements of Data Mining algorithms adopt in the elaborations of the whole EDM process. Thus, a set of operations (conversion, integration, discretization, etc.) is needed in order to adjust the collected data.

• **Data Mining.** The goal of this phase is mainly to use existing Data Mining techniques (prediction, clustering, outlier detection, association rule mining, text mining, etc.) to extract implicit patterns from data.

• **Result Interpretation.** The extracted patterns, often described in terms of models, have to be interpreted in order to become useful for the decision processes.

Lastly, there are many tasks in which EDM is applied. The most important ones are analysis and visualization of data (Bunkar et al., 2012), feedback provisioning for supporting instructors (Dejaeger et al., 2012), generation of recommendations for students (Klašnja-Milićevića et al., 2011), student’s performance prediction (Frias-Martinez, Chen & Liu, 2006), student modeling, grouping students, planning and scheduling (Hershkovitz & Nachmias, 2008).

### 1.2 Semantic Web in Education

The aim of the Semantic Web is to overcome the limitations of the traditional Web, consisting in a “representation” which is only designed for humans (machine-representable). The Semantic Web determines the transition to a “machine-understandable” vision which specifies the meaning of data to ensure a correct usage by machines (Signore, 2013).

Following the Semantic Web stack proposed by W3C,1 core element of the Semantic Web is represented by ontologies. The ontologies are used to represent knowledge in specified domains and have been defined by T. Gruber (1995) as “an explicit specification of a conceptualization”.

In education, the ontologies are used to improve knowledge sharing processes and enable efficient and effective access to relevant data. Ontologies are thus exploited in modelling knowledge for Adaptive Learning Environments (Zhao & Zhang, 2009) in order to improve students’ learning by providing adaptive feedback or by adjusting the learning environment basing on students’ performances and behaviours. As found in literature, the ontologies in education have been also used to model, describe and organize students’ knowledge, skills, preferences, learning styles, learning objectives (Abdel-Rahman et al., 2012; Szilagyi & Roxin, 2012), learning content and metadata (Zouaq & Nkambou, 2009; Roy, Sarkar, & Ghose, 2009), educational strategies (Fernández-Breis et al., 2012), sub-

1 [http://www.w3.org/2001/sw/](http://www.w3.org/2001/sw/)
jects and their relationships (Gaeta, Orciuoli & Ritrovato, 2009). Moreover, Bittencourt and colleagues (Bittencourt et al., 2009) summarize the main principles of the Semantic Web-based Educational Systems. In particular, they assert that the importance of ontologies in educational systems consists in enabling browsing, sharing and reusing of learning content, possibly stored in different repositories, and in allowing interoperability and integration among them. Interoperability, reuse and integration are enabled by the main standards proposed by W3C (i.e. XML\(^2\), RDF\(^3\), RDFS\(^4\), OWL\(^5\), OWL2\(^6\) and SKOS\(^7\)).

1.3 Synergy among Semantic Web and Educational Data Mining

In the past, Data Mining and Semantic Web have been often considered as two independent research fields where Semantic Web allows knowledge representation by means of formal languages (e.g. OWL) and Data Mining is applied to elicit non-explicitly represented knowledge from big data volumes by means of effective and efficient techniques. In recent years, more and more experiments have considered a combined application of Semantic Web and Data Mining techniques (for sake of simplicity we use the term “SW-DM” in order to address the above mentioned combined application). SW-DM has been used to build ontologies by means of automatic or semi-automatic processes (Agirre et al., 2000) (Xu et al., 2002) to find a mapping among elements of two ontologies (defining conceptualizations in the same domain) (Cañadas et al., 2004), to merge ontologies (Noy, 2004) (Bruijn et al., 2006) and to support ontology evolution (d’Aquin, Kronberger & Suárez-Figueroa, 2012) where authors propose a methodology to update ontologies with knowledge extracted from data by means of Data Mining techniques. SW-DM is also considered useful in order to support semantic annotation processes. Techniques like, for instance, Text Mining and Clustering have been exploited for annotating documents (Kiyavitskaya et al., 2006), images (Wang et al., 2006) and other digital resources. Furthermore, another field in which Semantic Web and Data Mining have been synergistically applied is the Semantic Web Mining that uses Data Mining techniques over semantic data in order to elicit knowledge that is not explicitly represented. Khan et colleagues (Khan, Verma & Jain, 2011), for instance, propose a Semantic Web Mining algorithm to improve precision and recall in Web searching processes executed in educational domains. Moreover, Mustapaşa and colleagues (Mustapaşa et al., 2011) apply techniques of Semantic Web Mining over log data in order to elicit knowledge about students’ motivation and behaviour. This knowledge is used to adapt the learning experience in order to sustain students’ engagement and selecting learning content that best fit student’s characteristics.

---

2 http://www.w3.org/standards/xml/
3 http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
4 http://www.w3.org/TR/rdf-schema/
5 http://www.w3.org/2004/OWL/
6 http://www.w3.org/TR/owl2-primer/
7 http://www.w3.org/TR/skos-primer/
In the next sections we propose a different approach to combine Semantic Web and Data Mining in Education. In particular, we combine semantic educational data and Educational Data Mining results in order to generate more detailed feedback that may support instructor’s decision making processes with respect to adequacy of learning material in MOOCs.

2. Improving Quality of Learning in MOOCs: a Proposed Approach

2.1 Overall Approach and Adaptation Strategy

The drop-out phenomenon in MOOCs is mainly due to possible students’ decreasing motivation when they feel frustrated by the difficulty to understand lectures and pass exams. In some cases, the aforementioned issues are not completely due to the inadequacy of students’ diligence. In fact, the massive nature of MOOCs imposes that the same content fits all students with no assumptions about their learning style, preferences, available tools, existing knowledge and skills. Thus, an incremental tuning of learning material is needed to adjust a specific content and make it fit for the widest and most heterogeneous possible audience.

The above mentioned tuning has to be performed by instructors who typically gather feedback by students, analyse it and adapt the produced learning content accordingly.

In traditional MOOC environments, it is difficult (although not impossible), to handle a dynamic adaptation of learning content. In this work, an adaptation scenario in which instructors improve new MOOC sessions (for instance for a new class in next semester), making use of the same material, is considered.

Figure 2. Overall Approach
Figure 2 provides a high level description of the proposed approach. More in details, in step 1 (steps are reported in black circle in Figure 2), students access the MOOC and its learning contents. Students’ actions, on the learning contents, are traced and stored in a specific database in step 2. In step 3, students’ tracks and profiles are used to build a model based on a decision tree which classifies students with respect to their ability to successfully acquire knowledge on concepts explained by a specific learning content. The constructed model and semantic data associated to the learning content (which we are considering) are used in step 4 to generate feedbacks on possible anomalies in the learning content. Anomalies can range from uncorrected metadata (associated to learning material) to uncorrected organization of learning material and are forwarded to instructors in step 5. In step 6, the instructor analyses the feedbacks received, reflects on and decides about a suitable intervention to accomplish. Lastly, in step 7, the instructor possibly takes her actions to improve the learning material and consequently the MOOC quality. Among all possible types of anomalies, this work focuses on erroneous selections of learning contents with respect to students’ characteristics. For example, let us consider an instructor, who is preparing a MOOC, and chooses a specific learning object, having a narrative text as a resource type, in order to explain a given subject. Moreover, let us assume that many students (a significant number of students with respect to the total number) achieve low scores when they are called to execute an assessment test for the above subject. In this case, it is useful for the instructor to analyse the situation and understand if poor performances are due to students or to some anomalies in the used learning object. Lastly, let us also assume that the selection of that learning object was wrong owing to the inadequacy of the narrative texts to opportunely support the learning process for the desired subject to be explained and the common characteristics of most students. The proposed approach aims at providing instructors with a useful support to detect these anomalies.

The detection of this type of anomalies allows instructors to provide alternative learning contents (explaining the same subjects but having different characteristics in terms of learning resource type, interactivity types, etc.) or, for instance, to replace the anomalous learning content with a new one (providing more adequate characteristics) in new MOOC sessions. However, the adaptation strategy of the proposed approach foresees a human intervention (instructor) and follows the flow reported in Figure 3.

Figure 3. Adaptation Process
In Figure 3 the whole adaptation process is depicted. In particular, adaptation is applied across two sessions (possibly in different semesters) of the same MOOC (for two different classes). The idea is to take advantage from the experience done in the previous session in order to improve the quality of learning in a MOOC for the next sessions. The outcome of the proposed approach is a set of feedbacks generated to support the adaptation actions the instructors could apply.

In the next sections we will give further details on feedback generation and model construction (see Figure 2) but only after we have given a clear picture of the semantic representation of knowledge in MOOCs.

2.2 A Model for Semantic Web-based MOOCs

In order to provide MOOCs with the benefits of Semantic Technologies, a model of Semantic Web-based MOOC has to be defined. MOMAMOOC8 is a MOOC Platform based on (Intelligent Web Teacher) (Capuano, Miranda & Orciuoli, 2010), an e-Learning Platform that lays on Semantic Web and other Semantic Technologies in order to provide adaptive and personalized learning experiences.

In MOMAMOOC, the learning material is decomposed and packaged in atomic units called learning objects (Sampson, Karagiannidis & Kinshuk, 2010). Video Lectures, short quizzes, texts and assessments (exams) are all packaged as learning objects. In order to describe and express the characteristics of a learning object, MOMAMOOC applies a strategy based on Metadata and makes references to IEEE LOM9 (IEEE Learning Object Metadata) and IMS LRM10 (IMS Learning Resource Metadata). The most important elements of the adopted metadata schemas concern with information about general (e.g. language), technical (e.g. format), educational (e.g. interactivity type, learning context), classification (e.g. discipline, learning objectives) characteristics. MOMAMOOC organizes learning objects by means of conceptual structures, namely Subject Ontologies that are, de facto, lightweight ontologies formally describing the knowledge of a disciplinary domain of interest by defining its relevant concepts (corresponding to subjects or topics) and relationships among them.

In particular, the adopted notation foresees a partonomy relationship, namely Has Part and two order relationships, namely Is Required By and Suggested Order. Furthermore, MOMAMOOC describes students’ profiles by taking into account IMS LIP11 (IMS Learner Information Profiles) and considering learning styles, preferences and cognitive states. MOMAMOOC also supports non-traditional forms of MOOCs (i.e. Group MOOCs and Connectivist MOOCs) (Clarke, 2013) that are mostly based on collaboration and knowledge creation.

MOMAMOOC represents knowledge about learning material and students by using Semantic Web-based ontologies and vocabularies in RDFS and OWL. More in details, learning object metadata, subject ontologies and students’ profiles

---

8 http://www.momamooc.com
10 http://www.imsglobal.org/metadata/
11 http://www.imsglobal.org/profiles/
are written respectively by using and extending Dublin Core\textsuperscript{12}, SKOS and FOAF\textsuperscript{13}. Moreover, the SIOC\textsuperscript{14} ontology is exploited to semantically describe and organize collaborative activities performed by students on Web 2.0 tools (e.g. discussion forums, wikis, blogs, social networks, instant messaging software, and so on). Figure 4 illustrates how learning objects in MOMAMOOC are semantically described and organized by means of Semantic Web vocabularies and ontologies. In particular, this organization allows interoperability, reuse and other valuable characteristics across different MOOCs also deployed on different Platforms. Firstly, RDFS and OWL are standard languages enabling interoperability, integration and extensibility of schemas. The availability of robust and high performing \textit{Semantic Storage and Inference Systems} (which are able to manage semantic data in RDFS and OWL) guarantees scalability and distribution over the network. These aspects allow the system to evolve. Secondly, the standard metadata layer enables portability of learning objects across multiple and heterogeneous systems and a richer filtering capability with respect to different qualities of a learning content. Thirdly, the adoption of subject ontologies enable learning object searching, sequencing, browsing and filtering at the conceptual level. This aspect is important in order to allow rich and pedagogy-oriented course authoring tools for instructors. Subject ontologies play an important role also in scenarios where cMOOC (Clow, 2013) are considered. In fact, the conceptual representation of the disciplinary domain allows user-generated content (UGC), produced by Web 2.0 tools, to be correlated to existing content (e.g. UGC and learning objects) dealing with the same or similar subjects.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.png}
\caption{Semantic Web-based MOOC structure}
\end{figure}

\textsuperscript{12} http://dublincore.org/
\textsuperscript{13} http://www.foaf-project.org/
\textsuperscript{14} http://sioc-project.org/
Lastly, MOMAMOOC added values regard its capability to allow adaptation and personalization of learning experiences by exploiting subject ontologies, learning resource metadata and students’ profiles. More in details:

- already acquired subjects (this information comes from the cognitive status part / side stored in students’ profiles) are not further proposed in new learning paths;
- remedial works (additional sequences of learning objects) are provided in response to poor scores obtained by students in assessment phases;
- some learning objects are preferred to others (dealing with the same subjects) if their characteristics (considering their learning resource metadata) better fit the considered student’s profile.

All these distinctive characteristics provide support to students, in terms of formative feedback, automatic tutoring and personalization, and offer plausible solutions for students’ disengagement and frustration. The original contribution of this paper concerns with the enhancement of MOMAMOOC with Educational Data Mining techniques applied in synergy with its existing Semantic Web-based architecture, in order to face the problem related to students’ motivation, engagement and (consequently) drop-out from the perspective of a learning object quality.

2.3 Finding candidate anomalies in learning objects by using Educational Data Mining

The aim of this step is to detect a set of possible anomalies related to the learning objects in a MOOC and, in particular, in the choice to provide a learning object with some specific characteristics (expressed in the associated metadata) to explain a given subject in a course.

The idea is the following. During the execution of a MOOC, some students perform actions involving the use of learning objects included in the course. Given a specific learning object, the students produce a set of actions like, for instance, giving answers to quizzes about subjects explained in the learning object and producing a number of information like, for instance, time spent studying the learning object, number of times the LO has been visualised, etc.. These data and additional information about students’ profiles can be associated and elaborated by means of Data Mining algorithms in order to perform a classification with respect to three categories (i.e. low score, medium score, high score) related to the students’ learning performance. The classification algorithm identified for the aim of this work is named C4.8 (Quinlan, 1993) (Hall et al., 2009) and builds a decision tree starting from a dataset previously pre-processed. The adoption of this algorithm is due to its capability to extract knowledge, from data, that is interpretable and representable by means a set of IF-THEN-ELSE rules. In particular, the algorithm is used to understand what students’ characteristics and behaviours are really relevant to obtain low, medium or high score for each subject in the course.

The dataset structure is represented by a set of records including the fields illustrated in Table 1.
The dataset has to be preliminary pre-processed (second step of the process in Figure 1) by means of three phases: data reduction (some existing attributes have to be eliminated, in this work we only focused on the attributes reported in Table 1), data integration (data from different sources must be combined in order to obtain a uniform representation) and data cleaning (to improve dataset quality it needs to clean up data from outliers, noisy and incomplete values).

Once the dataset has been pre-processed, the decision tree can be constructed by means of the above mentioned algorithm (Frank et al, 2009). The algorithm allows to divide the set of students into three subsets corresponding respectively to students who achieve a high score (first subset), students who achieve a medium score (second subset) and students who achieve a low score (third subset) for a given learning object (related to a specific subject). The C4.8 algorithm characterises each subset by means of rules determining the single students who belong to the subset.

It is important to note that the decision tree can generate different branches of classification rules for each considered category (i.e. low score, medium score and high score). Thus, it is necessary to select only that branch characterized by a greater number of instances correctly classified over a given threshold (experimentally calculated). This choice allows to improve precision and relevance of the generated feedbacks.

The next step is to consider the information (coming from the decision tree we have built) concerning the set of students who obtained low scores for a specific subject explained by means of the learning object. Let us call the aforementioned set and the set of relevant metadata fields (taken from the specifications introduced in the previous sections)\(^\text{15}\): \textit{LearningResourceType}, \textit{TypicalAgeRange},

\(^\text{15}\) http://www.imsglobal.org/metadata/imsmdv1p2p1/imsmd_bestv1p2p1.html
Difficulty, TypicalLearningTime, InteractivityType, InteractivityLevel and SemanticDensity. Lastly, and are two 7x2 matrices representing learning object metadata. For each matrix, the first column (and) is filled with element in and the second column (and) with values associated to the metadata field inserted in the first column of the same row (for example and contains the difficulty value of the field). (build experimentally) provides metadata values representing characteristics that the learning object should have to be suitable for students in. Moreover, is the metadata instance provided by the instructor who has authored the MOOC. The idea is that comparing and it is possible to identify metadata fields that can be really relevant for the success of the learning process. The candidate fields are those having different values for and.

More in details, is constructed by considering properties and values returned by the execution of algorithm C4.8 and their matches with learning object metadata fields and values. For instance, in Table 1, the learningStyle (one of the student’s characteristics) is considered. We have no corresponding values in the set for learningStyle, but it is possible to relate this characteristic to LearningResourceType, InteractivityType and InteractivityLevel by considering scientific results (Franzoni et al., 2008) (Baniulis, et al., 2009). Moreover, it is possible to note a direct correspondence for TypicalAgeRange (it can be acquired by considering the age value in the dataset structure) and TypicalLearningTime (it can be acquired by considering the time spent by students in). At the moment, SemanticDensity and Difficulty are not managed, but they will be considered in future works.

This phase ends with the comparison between and and produces a list of candidate’s anomalies (and their corresponding values), identified by considering metadata fields (titled by the label and) that have values such that.

2.4 Generating feedback by reasoning on Semantic Web-based structures

Feedback generation represents a phase in which candidate’s anomalies detected by applying the previous steps are refined and transformed into feedback for instructors who can start their decision process and possibly adapt the MOOC for next sessions by modifying content, metadata and ontologies related to used and analysed learning material.

The goal is to define a set of rules testing conditions over the available Subject Ontologies in order to refine (i.e. accept, order, etc.) these candidate’s anomalies. For sake of simplicity, we only focus on the relationships of the MOMAMOOC Subject Ontologies. In particular, let us consider subjects and in the sample Subject Ontology in Figure 5 (note that the set Q contains all Learning Objects on which an anomaly has been found). and are related through a relationship. Let us also assume that candidate’s anomalies are detected for and, that are learning objects explaining subjects and respectively. The idea is that low scores for subject could be influenced by an erroneous selection/definition of the content in (in terms of its characteristics) but they could be also influenced by anomalies in explaining subject.
Thus, Subject Ontologies are used to better specify feedback that can be generated from candidate’s anomalies obtained by applying Educational Data Mining techniques. In particular, the semantics of their relationships can be used to provide a richer feedback messages also organized in coherent sequences as reported in Table 2.
3. First Experimentation and Evaluation

Early experiments were conducted to divide the students into three subsets (low score, medium score, high score) and to derive the characteristic parameters (age, learning time, learning resource type, learning style and interactivity level) for each of them.

For a first test of our approach we used a dataset related to a MOMAMOOC platform adopted for supporting an online course of “Geometry” deployed at the University of Salerno. The course was accessible for a limited number of students. A pre-processing step was executed to improve the quality of the dataset:

- Data integration was used for combining together tuples coming from almost 330 tables of the MOMAMOOC databases containing log file, students’ profiles, metadata, ontologies, etc.;
- Data cleaning, instead, was used to clean up data from outliers, noisy and incomplete values.

The generation of the decision tree has been executed by means of Weka\(^\text{16}\) and, in particular, the J48 algorithm. This J48 was configured as follows: \(\text{binarySplits} = \text{false}, \text{confidenceFactor} = 0.25\) and \(\text{minNumObj} = 2\).

The whole training set was composed of 4 attributes (i.e. age, learningStyle, meanVote and percentage that are explained in Table 1) from 208 students. The result of the J48 algorithm for the learning object related to the “Matrix Definition” subject is shown in Figure 6 and commented in the following paragraphs.

\[\text{Figure 6. A fragment of the generated decision tree}\]

\(^{16}\) http://www.cs.waikato.ac.nz/ml/weka/
All attributes in the dataset appear in the resulting decision tree, therefore they can be considered meaningful to classify our students according to the proposed approach. The quality of the decision tree can be assessed through the confusion matrix shown in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>2</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Confusion Matrix

Calculating the Cohen’s Kappa associated with the confusion matrix we obtain a value of 0.6596 that can be considered a good result (Landis and Koch, 1977). The results of the constructed decision tree, expressed in terms of rules, have been then used to assess the quality of the learning process on subject “Matrix Definition” of the “Geometry” course (Figure 8 shows a part of the ontology used). The metadata associated with the resource titled “mt-defmatrici”, responsible for explaining the considered subject, is configured as follows: \textit{interactivitytype}=expositive, \textit{learningresourcetype}=narrative text, \textit{interactivitylevel}=very low, \textit{semanticdensity}=low, \textit{typicalagerange}=18-50, \textit{difficulty}=easy and \textit{typicallearningtime}=06m40s.

Coherently with the proposed model, we choose the branch, in the constructed decision tree, characterized by a \textit{meanVote} between 5.76 and 7.08, \textit{learningStyle} = analytical, \textit{percentage} > 3 and \textit{vote} = low. In this case, the number of correctly classified instances for the class “low score” is 29 (i.e. the group of students, with a low score regarding the current learning object, better classified by the J48 algorithm).

Again, according to the proposed approach we can assert that the type of learning resource provided in the MOOC is incompatible with the students’ learning style (this comes from rules generated by the J48 algorithm) and can negatively impact on the learning process quality. Thus, a feedback to the instructor has to
be generated and sent. Moreover, students with a learning style of analytic type prefer interactive learning resources than simple narrative text (Franzoni et al., 2008) (Baniulis, et al., 2009). So, the instructor has the chance to provide alternative learning objects (with a high level of interactivity) in a new “Geometry” course session. Lastly, the process is repeated for all the 40 learning objects of the “Geometry” course. 12 feedbacks have been generated. Almost 75% of feedbacks have been considered useful by the instructor.

**Final Remarks**

This work proposes an approach to improve the quality of learning processes in MOOCs by synergistically exploiting Educational Data Mining and Semantic Web. More in details, the Semantic Web-based representation of metadata associated to learning objects together with the subject ontologies applied to organise the course material are used to refine the results of a classification algorithm (C4.8) providing feedbacks on possible anomalies related to the content characteristics of given learning objects. The outcome of the proposed approach consists in sequences (the order provides information about priorities) of feedbacks for instructors who can provide several forms of off-line adaptation to MOOC contents. The proposed approach is experimented by using the MOMAMOOC installation at the University of Salerno and, in particular, for a “Geometry” course. Information about the early experimentation is provided in the paper. The experimentation results are promising, although the authors are aware that in order to definitely demonstrate the effectiveness of the approach more complete experimentation activities have to be executed.

In a future work we planned to consider: i) a propagation algorithm to manage the impact of all pre-requisites of a given subject in order to consider also indirect dependencies (pre-requisite path of length greater than 1), ii) other metadata properties like Semantic Density and Difficulty and iii) additional anomaly types like errors in metadata and in subject ontologies. These aspects, in our opinion, will improve the quality of feedbacks.

**References**


d’Aquin, M., Kronberger, G., & Suárez-Figueroa, M. (2012). Combining data mining and ontology engineering to enrich ontologies and linked data. In *Proceedings of the Knowledge Discovery and Data Mining Meet Linked Open Data (Know@ LOD) at the Extended Semantic Web Conference (ESWC).*


