Ontology-based semantic metadata extraction approach

Baraa JEBALI
Research Laboratory of Technologies of Information and Communication & Electrical Engineering (LaTICE)
University Of Tunis
5 Avenue Taha Hussein, B.P. 56, Bab Menara, Tunis, Tunisia
jebali.baraa@gmail.com

Ramzi FARHAT
Research Laboratory of Technologies of Information and Communication & Electrical Engineering (LaTICE)
University Of Tunis
5 Avenue Taha Hussein, B.P. 56, Bab Menara, Tunis, Tunisia
ramzi.farhat@esstt.rnu.tn

Abstract—In this paper we describe our approach for automatic generation of learning objects' semantic metadata. The extraction process is based on the OBIE (Ontology Based Information Extraction) systems’ principles. The input of our approach is a set of IEEE LOM metadata elements in conformance with two requirements. First, each data element must describe the educational content of the learning object. Second, it must be one of the data elements frequently filled by the learning objects’ authors and required by most of the LOM application profiles. Concerning the outputs, each one is a couple of a domain concept (from domain ontology) and a degree of pertinence. Moreover we present the details concerning the integration of our approach to learning objects’ repositories by taking the COLORS repository as example. In fact the ultimate goal behind our approach is the improvement of repositories’ services by offering semantic metadata.

Keywords—Semantic metadata, LOM metadata, ontology based information extraction system

I. INTRODUCTION

Nowadays the learning object paradigm is widely accepted. The proof is that numerous repositories supporting this paradigm are implemented and available online such ARIADNE1, EdNA2, etc. However, many related research issues are still open. A common research issue concerns the improvement of learning object repositories’ services especially research engines and learning objects’ auto-classification. In our opinion the main restriction is the nature of the learning object metadata used by such services. In fact, learning object suppliers uses LOM (Learning Object Metadata) standard [1] to deliver the metadata describing the learning object. Those metadata are filled by the learning object author to be used by final users which are humans (e.g. learners, teachers, etc.). However those metadata are note suitable for use by computer programs. The fact that computer programs use metadata designed for humans explain why there is a problem of effectiveness.

To overcome this common issue in information science the semantic metadata concept was introduced. In practice a semantic metadata is a metadata that is linked to a given domain ontology. Semantic metadata is characterised by the use of a common vocabulary with less risks of ambiguity. Moreover the relationships between the domain concepts (described by the common vocabulary) within the ontology allow users first to have a better understanding of the concept and second to be able to make some deduction. Another interesting characteristic of semantic metadata is the fact that they are understandable by humans as well as by computer programs.

Semantic metadata are firstly introduced by the W3C to allow the understanding of the Web content by humans and by machines. For the same reasons they have been introduced to the e-learning domain. However, they are mainly used in research works but not in practice. In our opinion this is due to two main factors. Firstly it is a tedious and a time consuming task from the learning objects’ authors point of view. Secondly, from a technical point of view, semantic metadata is depending on particular domain ontology. This aspect will be a restriction to have a reusable learning object (with no dependencies).

Unfortunately without semantic metadata many intelligent e-learning services related to the improvement of the learning objects’ management and use can’t be implemented. Advanced search engines, content personalization services, content analysis and automatic evaluation approaches, content recommendation methods are some examples of what we call intelligent e-learning services.

Our proposal consists on allowing the use of existing widely accepted learning objects metadata standard such as LOM [1] and in the same time enjoy the benefits of using semantic metadata. It consists on developing an ontology-based semantic metadata extraction approach for learning objects. Semantic metadata will be extracted by analysing LOM metadata using natural language processing technologies. The automatically generated semantic metadata will not be integrated to the learning objects itself but it will be externally associated. By this, the original learning object will not be affected and therefore it will still remain reusable. In other

---

1 http://www.ariadne-eu.org
2 http://www.edna.edu.au

U.S. Government work not protected by U.S. copyright
terms with our approach we will keep learning objects interoperable and in the same time powerful tools (needing semantic metadata) can be built.

In the first section of this paper we will present an overview of our approach and define our system architecture. We will define the semantic metadata structure in the second section. The third section will contain a description of the system inputs. We will describe a use case of our approach with COLORS platform in the final section.

II. OVERVIEW OF OUR APPROACH

Existent works confirm the importance of the IEEE-LOM standard but most of them agree that it is not enough since it does not support semantic information about the learning objects. Some proposals extend LOM metadata with semantic information. The Hypermedia Learning Objects System (hylOs) extends the relational category of LOM with a semantic network to interconnect different learning objects [2]. <E-aula> is a system that uses a restricted number of the LOM categories in order to include the learning object into its pedagogic context [3]. The <e-aula> proposes a partial view of the semantic of a learning object. In our proposal we consider that the semantic information about the learning object must be generic.

Indeed, our aim is to propose an approach to enrich the description of the existing learning objects by semantic metadata generated automatically. However to reach our goal we must overcome three challenges: (i) specifying which kinds of semantic metadata we must extract, (ii) Identifying the most suitable inputs to understand the semantic of a learning object, (iii) and defining how the process can be automated.

We suppose that we have learning objects with LOM like metadata. They are given in natural language. Some of them describe the content of the learning object and can be used to understand the semantic of its content. By using an ontology based information extraction system we must be able to understand the semantic of the learning object’s content. Finally to be useful, the extracted information must be structured and stored as semantic metadata. To recapitulate we can say that our approach uses an ontology based information extraction system taking LOM metadata as inputs and providing semantic metadata as outputs.

An Ontology-Based Information Extraction (OBIE) System is a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies [3]. We proposed a customized ontology semantic metadata extraction system (OBSME) architecture presented in Figure.1 inspired from the common architecture proposed in [3]. In fact in our case the ontology already exists (defined by the domain experts) so the task of its creation is not accorded to the system. Thus our system doesn’t contain any ontology creator or editor. That is why in our proposal the ontology is a part of the system and not an input as the case of the general architecture.

In our system the preprocessor contains two parts. The first part is a tokenizer\(^3\) that takes as input a set of LOM data elements. This module deals with the inputs and gives a set of separated terms as output. The second part is a vectorization module that converts each term to a vector using a vector space model\(^4\). Indeed we convert each term to tf-idf weights.

As an example of a classic vector space model\(^4\) we suppose that we have four documents containing LOM metadata. Every document contains two data elements (DE). We will vectorize every term in these LOM documents. We will take the example of the term “algo” as showed in Figure 2.

![Figure 1. Customized OBIE system architecture](image)

![Figure 2. Exemple of a term vectorization](image)

\(^3\) http://en.wikipedia.org/wiki/Tokenization.
number of LOM documents in the repository. \( IDF \) presents the inverse document frequency and \( w_i \) presents the weight of the term \( i \).

\[
\begin{align*}
\text{IDF} &= \log(D/df_i) \\
\text{w}_i &= tf_i \cdot idf_i \\
\text{df}_i &= 3 \text{ and } D = 3 \\
\text{IDF} &= \log(4/3) = 0.124
\end{align*}
\]

The information extraction module uses the hierarchical perceptron algorithm named Hieron [4]. The Hieron algorithm for information extraction is based on hierarchical classification that exploits the structure of concepts [5]. Existent experimentation demonstrated promising results by applying this algorithm to ontology based metadata extraction [6]. Hieron exploits the relationships among classes in the ontology; the expectation is that it would perform better on OBIE [6] [7].

The Hieron algorithms takes a set of instances \( X \). These instances in our case are the output of the vectorization module. To each instance this algorithm sets a label \( Y \). To each label it sets a table \( W \) containing its weights. The ontology is the set of concepts \( Y \). Finally Hieron predicts the most suitable label to the instance. Using these predictions, our system provides semantic metadata as output based on the prediction made by the Hieron algorithm.

III. SYSTEM OUTPUTS: SEMANTIC METADATA

The main goal behind the use of semantic metadata is to allow the implementation of intelligent services. Those services, such as advanced research engines or learning object recommendation and personalization, need to recognize the semantic of the educational content through the identification of the domain model that the concepts are included in it. The domain model (which is in general represented by an ontology) offers additional semantic information thanks to the description of the relationships between concepts.

Semantic metadata is the process of attaching semantic descriptions to Web resources by linking them to a number of classes and properties defined in Ontologies [9]. On the scope of this definition, we propose to represent the semantic of a learning object as a set of couples \( \{C_i, DP_i\} \) with \( C_i \) is a concept and \( DP_i \) is its degree of pertinence. The concept \( C_i \) is one of the concepts covered by the educational content belonging to the domain ontology. The \( DP_i \) is the degree of pertinence of the concept \( C_i \) having a value varying from zero to one. Concepts with high values are more representative of the learning object educational content.

For example, if we consider a learning object having as content a relational database course, our system can provide as outputs the following semantic metadata:

\{Relational Database, 1\}

\{Relational Model, 0.8\}

The degree of pertinence indicates that the educational content is mainly about “relational databases” (which have the highest degree of pertinence). It indicates too that the educational content covers the “relational model” concept as well as “SQL” concept (the degree of pertinence is equal to 0.8). Finally it provides some educational content about “normal forms” (the degree of pertinence is equal to 0.5).

IV. SYSTEM INPUTS: LOM METADATA

The LOM standard [1] contains about seventy fields organized in nine categories: General, Life Cycle, Metadata, Technical, Educational, Rights, Relation, Annotation and Classification. In our approach we must identify which data elements are relevant to be used as input to calculate semantic metadata (describing the learning object’s content). We have fixed two requirements that must be verified by a metadata data element to be considered as relevant. First, the data element must describe the educational content of the learning object. Second, it must be one of the data elements frequently filled by the learning objects’ authors and required by most of the LOM application profiles. In fact these requirements makes our approach coherent with the existing application profiles and compatible with the majority of the learning objects already stored in repositories.

According to the first requirement we made the choice of three categories containing data elements about the content: General, Educational and Classification. From the General category we have selected “title”, “description”, “keywords” and “coverage” data elements. The “educational” category offers information about the educational and pedagogical use of the learning object. The “learning resource type” in the educational category is the unique useful data. It allows the extraction of the learning object educational role. The “classification” category gives us the opportunity to identify the real use context of the learning object. But the given classification of the different taxons in this category is defined by the author. Thus it is probably not in conformance with the vocabulary defined in the ontology. This classification is helpful to understand the author point of view concerning his learning object.

To satisfy the second requirement we must verify that those LOM data elements are in the most cases filled by authors. To verify this requirement we have investigated two aspects. First, we have analyzed the following widely used LOM application profiles to check the data elements statute (i.e. mandatory, recommended or optional) in order to ensure that our data elements are not optional: UK Learning Object Metadata Core[10], Normetic [11], Celebrate [12], ScoILOMfr [13], ManUel [14], LOM-FR [15], Canadian Core (CanCore) [16], CG LOM Core [17], FAILTE [18], BECTA NLN [19]. This first step shows that we must exclude the data element “coverage” when it is considered by all those application profiles as optional (see the first column of Tab. 1.).

The second aspect investigated is the filling of LOM data elements in existing wide used repositories. We have used for
this purpose the results of an existing study on eight repositories [14]. It shows the percentages of the filled LOM data elements in the following repositories: Learning Resource Exchange (LRE), Community on Learning Objects (LACLO), OER Commons (OER), Korean OCW (KOCW), LO Repository Network (LORNET), Open University Japan (OUJ). This study confirms that the retained different data elements except the “coverage” data elements contain in the most cases information. Thus the input of our system will not be empty. This input is composed by the following data elements: “title”, “description”, “keywords”, “learning resource types”, “taxon path” and “taxon”.

<table>
<thead>
<tr>
<th>Percentage of application Profiles setting the data element as mandatory or recommended (%)</th>
<th>Percentage of the stored learning objects filling the data element (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>90</td>
</tr>
<tr>
<td>Description</td>
<td>80</td>
</tr>
<tr>
<td>Keywords</td>
<td>50</td>
</tr>
<tr>
<td>Coverage</td>
<td>0</td>
</tr>
<tr>
<td>Learning resource type</td>
<td>70</td>
</tr>
<tr>
<td>Taxon path</td>
<td>50</td>
</tr>
<tr>
<td>Taxon</td>
<td>50</td>
</tr>
</tbody>
</table>

* The Description and the Keywords are parts of the General Category

V. INTEGRATION TO COLORS

As mentioned in the introduction, the extraction of learning objects’ semantic metadata is not the finality of our work but a step forward to support the implementation of advanced services in learning objects repositories. To attend this goal it is crucial to demonstrate that our approach can be integrated in classic repositories. In this context we took the example of COLORS (COoperative Learning Object Repository) [19] that is a repository supporting SCORM learning objects. COLORS offer many services: exploration of learning objects throw a concept tree, identification of relevant learning objects throw the use of a search engine and a learning object analyzer providing feedbacks to the authors. Each one of those services can be improved if the learning objects’ semantic metadata are provided.

As shown in Figure 3, the specific LOM data elements are read from the COLORS database by the preprocessor. Then the information extraction module uses the ontology and applies the Hieron algorithm to produce semantic metadata. The extracted semantic metadata will be stored and used by COLORS services. As mentioned above, semantic metadata will not be packaged within the learning object when it is downloaded by users since the semantic metadata depends on ontology. In fact, in practice different communities may define for same domain different ontologies (this is due to some parameters like the fixed granularity of concepts, the nature of relationships considered, the experts’ divergence, etc.).

VI. CONCLUSION AND ONGOING WORKS

The main issue of our approach is to extract automatically semantic metadata describing learning objects’ content. With the automatic extraction of this kind of metadata it will be possible to implement advanced semantic oriented services and tools.

In our work we took advantage from existent works. We used LOM metadata as inputs of our system since there are widely accepted and already integrated in learning objects. We modeled the semantic context by a domain ontology. In fact, semantic information in general and semantic metadata in particular can never be defined outside of a context. The semantic metadata extraction process is based on an implementation of Hieron algorithm.

Our first experimentations are done using GATE libraries to deal with the ontologies and Java programming language to implement the preprocessor and the information extraction module.

In the future works we will make experimentations on an instance of COLORS repository in order to evaluate the quality of the semantic metadata automatically generated and their relevance based on experts’ opinion.

REFERENCES


