A Semantic-Oriented Approach for Organizing and Developing Annotation for E-Learning

Mihaela M. Brut, Florence Sedes, *Senior Member*, *IEEE*, and Stefan Daniel Dumitrescu

Abstract—This paper presents a solution to extend the IEEE LOM standard with ontology-based semantic annotations for efficient use of learning objects outside Learning Management Systems. The data model corresponding to this approach is first presented. The proposed indexing technique for this model development in order to acquire a better annotation of learning resources is further presented. This technique extends and combines two consecrated alternative methods for structure-based indexing of textual resources: the mathematical approach of the latent semantic indexing and the linguistic-oriented WordNet-based text processing. Thus, the reason behind the good results provided by the first method becomes more transparent due to the linguistic controlled choices proposed by the second method. The paper results are important in the context of adopting semantic web technologies in the e-learning field, but also as a progress in the area of ontology-based indexing of textual resources.

Index Terms—Computer-managed instruction, ontology, semantic annotation, indexing methods, latent semantic indexing.

1 Introduction

The actual decentralization and interinstitution collaboration that characterize e-learning solutions increase the demand in making learning objects accessible not only across the e-learning platforms, but also across other Web applications. This demand could be solved with the support of semantic technologies, whose main goal is to reveal to the computer applications the resources' semantics, as information automatically processable.

The approach of extending e-learning solutions with semantic Web facilities involves mainly two aspects: e-learning data modeling and the processing technique leading to the established model.

The present paper first proposes a model that extends the IEEE LOM (Learning Object Metadata) standard with ontology-based semantic annotations for efficient use of Learning Objects (LOs) outside Learning Management Systems (LMS). This model is presented in the context of existing approaches that adopt ontologies in order to annotate e-learning resources. Its particularity consists in a simple solution for integrating semantic annotations in the structure of the IEEE LOM description of the e-learning resources, without requiring a special RDF LOM binding. Thus, the ontology considered in annotations is kept independently, while inside the IEEE LOM structure is included just information about the concepts and their weights for the current LO.

The paper also introduces a technique for automatic processing of e-learning textual resources leading to the development of the semantic annotation according to the proposed model. Traditionally, three alternative method types are used for structure-based indexing of textual resources: classification methods, linguistic methods, and mathematical methods. Instead of focusing on one such method, the proposed technique extends and combines two consacrated alternative methods: the mathematical approach of the latent semantic indexing (LSI) and the linguistic-oriented WordNet-based text processing. Thus, the reason behind the good results provided by the first method becomes more transparent due to the linguistic controlled choices proposed by the second method.

The ontology-based modeling was adopted in the e-learning field not only for the learning resources, but also for the user profile. In [12] we provided such a solution, where the user profile is focused on the user competencies, which are expressed in terms of the same ontology adopted in LOs' annotation. The ontology-based annotation uniformity of the various types of e-learning resources (such as LOs and users) could be exploited in many ways, in order to provide users with personalized functionalities:

- For selecting the suitable users for being coopted into a certain project, whose topics are also expressed through ontology concepts;
- For selecting the materials suitable for a certain user into a certain situation: when accessing a specific course site section, or when solving a certain course homework (as student), or when developing a certain course material (as teacher);
- For redirecting the student, in the case he gives a wrong answer to a test, to all the materials which provide explanations (not only to those specified by the test creator as feedback information);
- For better providing the student with the appropriate materials which are recommended or published by his collaborators in different projects or interest groups.

These facilities could be exported from an e-learning system to another if they are implemented, for example, in

- *M.M. Brut and F. Sedes are with the IRIT Laboratory, Paul Sabatier University, 118 Route de Narbonne, Toulouse 31062, Cedex 9, France. E-mail: {Mihaela.Brut, Florence.Sedes}@irit.fr.*
- *S.D. Dumitrescu is with the Politechnica University of Bucharest, Str. Ion Creanga 6A, bl. 91, ap. 14, Sector 5, Bucharest, Romania. E-mail: stefan.dumitrescu@i-neo.ro.*

Manuscript received 1 Jan. 2010; revised 27 May 2010; accepted 27 Oct. 2010; published online 10 Dec. 2010.
For information on obtaining reprints of this article, please send e-mail to: lt@computer.org, and reference IEEECS Log Number TLT-2010-01-0001.
Digital Object Identifier no. 10.1109/TLT.2010.40.

the form of web services. Thus, the user mobility across distributed e-learning communities is facilitated not only because the user competencies profile is recognized by other systems, but also because the e-learning materials are uniformly modeled.

We already worked on exploiting the semantic annotations associated with e-learning resources inside an existing tracking system that capture the user current activity, which is developed based on the Contextualized Attention Metadata (CAM)[1] [36] framework. In [7], we exposed a solution for recommending documents to students according to their current activity that is tracked in terms of semantic annotations associated with the accessed resources.

Aiming to define an ontology-based annotation model for e-learning documents, the present paper starts by presenting the context of e-learning and semantic web standards, discussing how semantic technologies could be integrated into e-learning. Considering IEEE-LOM e-learning standard, a document model solution that semantically extends it will be further presented. The solution considers an ontology-based annotation set for LOs. The document model development considers a manual and an automatic part for the annotation process. The latter concerns textual LOs and combines TFxIDF indexing with latent semantic indexing with WordNet-based processing. Before defining the solution for automatic annotations, a review of techniques for document indexing relevant to our approach is presented. Conclusions and further work directions are finally presented.

2 The E-Learning Metadata Standards and Practices

Reusability of learning content and instructional design are the main issues that various e-learning standards try to address. Two abstractions are employed to facilitate this reuse:

- *Learning Objects*. How to specify reusable chunks of learning content;
- *Learning Design (LD)*. How to define an abstract way for designing different units (e.g., courses, lessons, etc.).

The majority of the e-learning standards provide metadata specification for describing the properties of LOs (ARIADNE, DCMI, IEEE-LOM, ADL). There are also standards for Content Structure Modeling (AICC), used for describing the structure on content. As well, standards like IMS and ADL/SCORM (Sharable Content Object Reference Model) handle both metadata specification and content structure modeling [34].

Also, standard serialization formats help to improve LOs accessibiliy and reusability. For example, the XML (*eXtensible Markup Language*) and/or RDF (*Resource Description Framework*) formats facilitate Web-based resource retrieval for IEEE LOM and DCMI (educational version). Moreover, IEEE-LOM enables developers to extend and add new data elements as required by applications; for this reason, IEEE-LOM could also be used as the base standard for developing new "application profiles."

1. http://www.ariadne-eu.org/index.php?option=com_content&task=view&id=39&Itemid=55.

E-learning standards provide support especially for organizing educational resources into an interoperable manner. In order to acquire interoperability with respect to the semantic description of these resources' content, some semantic metadata should be additionally defined, by using standards specific to the Semantic Web. Various standards were defined, focused on specific information type description. Among these, we mention DCMI (*Dublin Core Metadata Initiative*, extended with an educational version),[2] Really Simple Syndication (RSS)[3] and Atom[4] for the Web sites syndication, *Friend Of A Friend*(FOAF)[5] for describing people profile through their social relations and (*Description Of A Project*) DOAP,[6] aspecification for describing relations between resources in an IT/project context. It is also possible to embed semantic metadata into Web resources through existing microformats[7] or through RDFa,[8] which provide support in addition for metadata interlinking.

However, the combination between e-learning standards and Semantic Web standards is a difficult issue. Al-Khalifa and Hugh [2] provide a classification of the different manners to represent and effectively use the standard metadata in e-learning applications:

- *Standard Metadata*—some applications adopt a standard e-learning metadata scheme (e.g, IEEE LOM) to represent the educational resources through a hierarchical structure, which follows a single perspective; such examples are UK LOM Core[9] and e-mi@ge Project;[10]
- *Semisemantic Metadata*—applications that use the IEEE-LOM standard with an extended semantic component; as an example, in [26] a representation of this standard in RDF format is proposed, and the advantages and complexities in using RDF to express learning object metadata following the IEEE LOM standard are discussed;
- *Semantic Metadata*—applications that rely completely on domain ontologies to define metadata for the educational resources, which are stored independently of the LOs' repository, which is organized according to e-learning standards. As examples could be mentioned Edutella peer-to-peer network,[11] or the Lorner pan-Canadian project.[12]

The most nuanced semantic metadata to describe the resources' content could be acquired by adopting ontologies, as it will be detailed in the next sections.

3 Using Ontologies for Annotating LOs

3.1 Existing Approaches

Each metadata type provides the computer applications some information about the meaning of the data content.

2. http://dublincore.org/.
3. http://www.rssboard.org/rss-specification.
4. http://www.atomenabled.com/.
5. http://www.foaf-project.org/index.html.
6. http://usefulinc.com/doap/.
7. http://www.microformats.org.
8. http://www.w3.org/TR/xhtml-rdfa-primer/.
9. http://zope.cetis.ac.uk/profiles/uklomcore/.
10. http://www.e-miage.org/.
11. http://www.edutella.org/edutella.shtml.
12. http://lornet.iat.sfu.ca/.

But problems arise when the same metadata element is used in many contexts, and a computer application is not able to observe differences, unless we provide it with some supplementary information. For example, <dc:title> indicates only that the content is the title of the current resource, but nothing about the meaning of this title. The solution is the correlation of certain metadata with a certain ontology construct.

The goal of using ontologies is modeling the information at the semantic level. Ontology is a knowledge domain conceptualization into a computer processable format, which models entities, attributes, and axioms [35]. A lot of research has already been accomplished in the field of Semantic Web technologies integration into e-learning environments.

Brase and Nejdl [5] refer to the IEEE LOM and DCMI standards for annotating educational resources in the computer science field and also, refer to the ACM classification system (http://www.acm.org/class/1998/) and to the SWEBOK (Guide to the Software Engineering Body of Knowledge- http://www.swebok.org) ontologies—preferring the last one —for refining the annotation.

Mohan and Brooks [25] consider the ontologies useful for marking up the structure of learning objects and for describing pedagogical meaning to them (their corresponding domain concepts) so that they can be understandable by machines. The authors recommend various ontology types to be related to LOs:

- Ontologies covering domain concepts—for the intelligent discovery and thematic assembly of LOs;
- Ontologies about teaching and learning strategies—for specifying the types of techniques each LO uses to facilitate learning;
- Ontologies describing the structure of learning objects—for allowing LOs to be interpreted and rendered consistently in different learning systems.

According [28], inside a LMS, with the support of ontologies there could be specified:

- The knowledge to be learned (the common usage of ontologies for describing the domain and task knowledge).
- How the knowledge should be learned (the authors developed mechanisms, based on ontologies, to create a rich supply of feedback to learners during learning, as well as to authors during course development).

Using an ontology-based modeling approach, the *Multimedia Informaion Repository* (MIR) project [15] provides a solution of adaptive facilities inside an e-learning platform which manage the multimedia information: it includes a user modeling component (MUMS—*Massive User Modeling System*), a component for managing and annotating the learning objects (hylOs—*Hypermedia Learning Object System*), and a component for defining the adaptation model (MIRaCLE—*MIR Adaptive Linking Environment*).

The ontology-based annotating tools [27] provide visual interfaces, where an image/video fragment or a textual sequence (that belongs to a LO, for example) could be described through some instances of a concept or concept properties inside ontology; the established metadata is provided in a certain format.

In order to enhance the expressive power of annotations (which create links between LOs and ontology concepts), some techniques of associating roles and/or weights with these relations were conceived. The roles are important to distinguish among the various type of connections between concepts and documents. The most common role is "prerequisite," which indicates that a certain document is a prerequisite for understanding the corresponding concept [9]. In [8], 30 roles are defined and used in order to identify the context within which a certain concept appears. In the *KBS-HyperBook* system, the documents could be marked as "problem statement," "example," "theory" for a certain concept [18].

The e-learning standards provide support for organizing the e-learning materials into a structured manner (e.g., on modules, courses, chapters, sections, lessons, learning objects). When a new material is added, its relations with the existing materials should be specified. Usually, the relation set defined by the IEEE-LOM Relation category is used. The semantics of these relations is refined or modified in [15], and also some new relations are introduced for the purpose of relating various LOs in a more semantic manner: isNarrowerThan/isBroaderThan, isAlternativeTo, Illustrates/isIllustratedBy, isLessSpecificThan/isMoreSpecificThan. These relations are inspired from the *Simple Knowledge Organization System Reference* (SKOS)[13] vocabulary for describing thesauri, glossaries, taxonomies, and terminologies. As a remark, instead of using these relations between different LOs, they could be used for relating the concepts of an ontology which models the domain knowledge. Indirectly, the relations between concepts illustrate the relations between the corresponding annotated LOs. The annotations are obtained as result of a *document indexing* process. Three attributes are important to distinguish different indexing approaches [10]:

- *Cardinality* involves two different cases: single concept indexing, where each page is related to one and only one domain model concept; and multiconcept indexing, where each page can be related to many concepts.
- *Navigation* involves also two different cases: the link between a concept and a page exists only on a conceptual level (used only by internal adaptation mechanisms of the system) and the cases where each link also defines a navigation path.
- *Expressive power* concerns the amount of information that the authors can associate with every link between a concept and a page. Of course, the most important information is the very presence of the link. This case is called flat indexing and is used in the majority of existing systems. Still, some systems with a large hyperspace and advanced adaptation techniques may want to associate more information with every link by using roles and/or weights. Assigning a role to a link helps distinguish several types of connections between concepts and pages. For example, the prerequisite role marks the case when the concept is not presented on a page, but instead the page is a required prerequisite for understanding the concept [9]. The weight of the

13. http://www.w3.org/TR/skos-reference/.

link between a concept and a page may specify, for example, the percentage of knowledge about a concept presented on this page [13].

3.2 The Proposed Annotation Model of LOs

An important demand for e-learning applications is the adoption of standards: in case of annotating LOs, the reference to the ontology concepts should be integrated in the structure of the e-learning standard used. To illustrate this, we take as example the IEEE-LOM standard [19], designed for schema-based LO classification, considered as the most enabling for semantic extensions [2].

Like other e-learning standards, the IEEE-LOM standard provides means for classifying LOs based on a classification scheme, through its *Classification* category. Our approach is to adopt domain ontology as such classification scheme: it is possible to specify the identification information (ideally the URI) of the domain concept we want to use in order to specify the LO is annotated with this concept.

For example, the following IEEE-LOM metadata expresses that the current LO is related to the Digital Libraries concept, which belongs to the ACM classification system, having the H.3.3 identification:

```
<Classification>
  <Purpose> discipline </Purpose>
  <TaxonPath>
    <Source> ("en",
"http://www. rkbexplorer.com/ontologies/acm")
  </Source>
    <Taxon> <id>#H.3.3 </id>
    <entry> ("en", "Information Search and
                    Retrieval")</entry>
    </Taxon>
  </TaxonPath>
</Classification>
```

The <Purpose> subelement expresses the purpose of classifying the current learning object, and we provided the pre-defined value "discipline" to suggest that we want to define the topic of the LO through the ontology reference.

The advantage of this kind of ontology-based annotation is that it allows for easy identification and interlinking of LOs dealing with the same or related topics (where relatedness between topics is inferred from the domain ontology).

Using such an approach, a certain learning object could be related to a single concept (*single-concept indexing*) or to multiple concepts (*multiconcepts indexing*). Also, a learning object could be related to single ontology concepts, or to concepts belonging to multiple ontologies, for example, through the Classification category. In the last case, some semantic overlapping problems could appear regarding concepts with similar semantics, belonging to different ontologies: for example, "knowledge management" and "knowledge databases" from two different ontologies could express the same concept. Our approach does not consider this case: the problem is quite complex and its general solution concerns different ontology alignment and merging operations [21].

In order to differentiate the importance of a concept-based LO annotation, we propose three new attributes that could be integrated in the *Classification* IEEE-LOM category. Their goal is to express the relevance degree of the referred concept for the current LO, through three possible values, expressing this relevance into a decreasing order:

- *isOnTopic*—for a LO which is especially destined to a certain topic;
- *usesTheConcept*—expressing the ordinary concepts encountered into LO;
- *makesReferenceTo*—for designating the other concepts from ontology encountered inside LO as hyperlinks or explicit references.

In [11], we proposed a process of semantic annotation of the e-learning materials that have two components:

- A *manual annotation component*. used especially for the LOs in the image/audio/video format. The annotation is accomplished mainly through the existing multimedia ontology-based annotation tools, such as PhotoStuff, AKTive Media, Vannotea, M-OntoMat-Annotizer, Semantic Web Advanced Development (SWAD) [27]. The problem in this case is that the usability of these tools for educators is very low.
- A *semiautomatically annotation component*. in the case of textual LO, the content edited by the teacher is parsed in real time in order to localize concepts existent in the selected ontologies, using the automatic annotation technique which will be exposed further in Section 4.2: when a match is retrieved, a notification is sent to the teacher, asking him/her to certify whether the match is correct or not.

Considering the example above, the annotation process will produce also the supplementary attribute:

```
<Classification>
  <Purpose> discipline </Purpose>
  <TaxonPath isOnTopic = "http://acm.rkbexplorer.
    com/ontologies/acm#H.3.3">
      <Source> ("en",
"http://acm.rkbexplorer.com/ontologies/acm")
  </Source>
      <Taxon> <id>#K.3.1</id>
        <entry> ("en", "Information Search and
                        Retrieval") </entry>
      </Taxon>
  </TaxonPath>
</Classification>
```

As could be noticed, our document model involves a multiconcept indexing, with conceptual level relations between document and concept, differentiated by the three roles: isOnTopic, usesTheConcept, makesReferenceTo. However, a single isOnTopic relation is associated with a document, representing the most expressive indexing relation.

The structure of ontology classes and relations providing support for ontology-based annotations is illustrated in Fig. 1.

4 The Document Annotation Model Representation and Development

Our proposed document indexing technique leading to automatically obtain semantic annotations (in the form

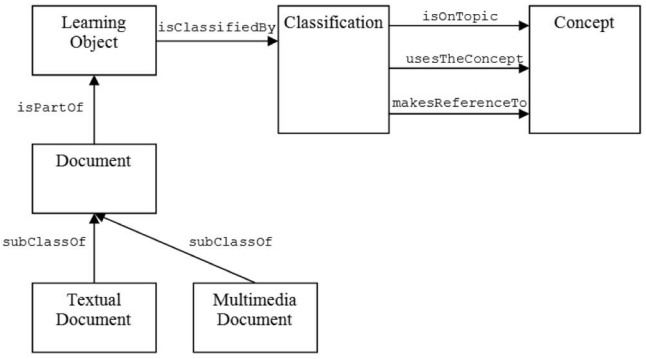

Fig. 1. The class hierarchy for ontology based LOs annotations.

presented in Section 3.2) in case of textual LOs will be exposed in Section 4.2. We will present further, in Section 4.1, the context and the main issues of the document indexing techniques that we combined and adapted in order to develop our indexing solution.

4.1 Existing Approaches in Textual Document Indexing Techniques

For textual documents, some indexing techniques are inspired by classic Information Retrieval [32], or by Web Information Retrieval, exploiting the hypertext features, such as page hyperlinks [1] and HTML general tags [6]. The progress from a term-based to a concept-based document indexation was possible due to the latent semantic indexing technique [33] or to some knowledge representation models and methods that are typical to artificial intelligence domain (such as neural networks, semantic networks, bayesian networks) [22].

The **classic IR** [32], [30] provide support to the document indexing domain for a two-step document processing:

- A document **preprocessing** phase, including. *Tokenization* (splitting sentences into words of word groups—tokens), *Stopword Removal* (the most common words, without a special meaning by themselves, are removed from the document), *Stemming* (through a morphologic analysis specific for the current language, each term is reduced to its morphologic root, in order to be recognized over its morphologic variations) and *Part-of-Speech Tagging*.
- **Term Weighting** phase. A weight is associated with each stemmed term, usually calculated by taking into account the whole collection D of documents where the current document d belongs.

The simplest weight is the *Term Frequency* TF: TF(t, d) is the number of times term t appears in the document d.

The most popular calculation method is *TFxIDF Weighting*, where the weight w_i is calculated in such a way to be proportional to the frequency of its corresponding term t_i in the document d, and inversely proportional to the number $|D|$ of documents in the collection D in which t_i appears

$$w_i = TF(t_i, d) log \frac{|D|}{DT(t_i)}, \tag{6.1}$$

where $DT(t_i)$ is the number of documents of the collection D that include the term t_i.

Other common methods are *Boolean Weighting* (the simplest), *Okapi BM25 Weighting* (probabilistic approach), *Entropy weighting* (based on ideas of the Information Theory), *Genetic Programming Weighting* (where term weighting schemes are automatically determined by genetic evolution and then tested on standard test collections) [22]. The vector (6.2) will increase since it should concern the terms that correspond to all documents.

As can be noticed, the result of the document processing is a representation of each document d as a vector of (term, weight) pairs:

$$d = \{(t_1, w_1), (t_2, w_2), \ldots, (t_n, w_n)\}. \tag{6.2}$$

The so-called problem *Curse of Dimensionality* occurs in the case of a large collection of documents D (containing, for exemple, several thousands of documents). Such collections are frequent in the case of e-learning portals or Web repositories.

For reducing the problem space, two approaches were adopted [22]:

- The dimension remains unchanged, but many zero values for weights are obtained, as in the *Boolean model* and *Probabilistic model*.
- A dimension reduction is performed, as in the *Vector Space Model*.
- The *Vector Space Model* approach adopts a matrix representation of the weights in the document collection: a row corresponds to each document d_j containing the weights w_{ij} of the all terms in the collection t_i, $i = 1$,m (not only the terms occurring in the d_j). Of course, the m dimension is quite big, and many w_{ij} have the 0 value. The dimension reduction consists in eliminating the columns corresponding to terms that occur in all the documents (the columns with all values greater than 0). The remained weights have the following property: $w_{ij} > 0 iff t_i \in d_j$ and t_i does not belong to all documents of the collection D. Such a new n $\ll$ m dimension is obtained, and each document is now represented as a row of n weights, corresponding to the n terms considered as relevant:

$$d_j = \{w_{1j}, w_{2j}, \ldots, w_{nj}\}.$$

The vector space model adopts a similar representation for each query used to retrieve relevant documents for a certain topic:

$$q = \{w_{1q}, w_{2q}, \ldots, w_{nq}\}.$$

The advantage of this representation consists in enabling to retrieve relevant documents through a very simple document-query similarity function. The most employed similarity function in the literature is the *cosine similarity*:

$$sim(d_j, q) = cos(d_j, q) = \frac{d_j \bullet q}{|d_j \parallel q|}. \tag{6.3}$$

This measure is equal to the cosine of the angle formed by the two vectors *dj* and q in the n-dimension vector space. In fact, it allows ranking of documents according to the similarity between the documents and the query.

The main drawback of this representation is the assumption that terms are independent from each other. Despite this assumption, the approach was successfully adopted in many real IR systems, as presented in [2].

In this paper, we will adopt another method of dimension reducing, more semantically oriented, namely the **Latent Semantic Indexing**. This technique illustrates a further extension of the document model, toward a representation based on concepts and semantic relations between index terms [23].

The idea of LSI [14] is to represent a document through concepts, rather than through index terms. The technique assumes there is some hidden structure in the use of the terms included in a collection of documents: the topic addressed by a text is more associated with the concepts that are used to describe it rather than with the terms actually used.

The high dimensional space formed by all the m index terms of a document collection, $ITD \equiv \{t_1, t_2, \ldots, t_m\}$, is mapped by means of Linear Algebra techniques into a lower dimensionality space S_n, with $n << m$, where every component s_j represents a *concept*.

It should be pointed out that, broadly speaking, the result of LSI techniques cannot be interpreted from a linguistic viewpoint. Such a result has a purely mathematical value [22].

Starting from the term-document matrix A_{mxn}, the most popular technique used to reveal these hidden relations between terms is the mathematical technique called *Singular Value Decomposition* (SVD) [16], [33]. Given a matrix A_{mxn}, where $m \geq n$ and $rank(A) = r$, the SVD of matrix A is defined as follows:

$$A = \mathcal{U} \cdot \Sigma \cdot V^T,$$

where $\mathcal{U}$ is a matrix mxr orthonormal ($\mathcal{U}^T \cdot \mathcal{U} = I_r$), V is a matrix nxr orthonormal ($V^T \cdot V = I_r$), and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, with $\sigma_i > 0, 1 \geq i \geq r$ and $\sigma_i > \sigma_{i+1}$.

The decomposed matrix A should be further reduced to a smaller dimension, k, through a projection technique known as *folding-in*. The reconstructed matrix Ak is a rank-k matrix that is the closest approximation to the original matrix A. In general, the choice of k is established empirically, through multiple tests, according to the specific problem.

The matrix calculus that reduces the space is represented through the following notation:

$$A_k = \mathcal{U}_k \cdot \Sigma_k \cdot V_k^T,$$

where U_k is a matrix m x k obtained by taking the first k columns from U, V_k is a matrix n x k obtained by taking the first columns from V , Σ_k is a matrix k x k obtained from the first k values of the diagonal of Σ. The reconstructed matrix $\mathrm{A_k}$ is a rank-k matrix that is the closest approximation to the original matrix A.

The rank of matrix A is thus reduced to k dimensions.

In the context of IR, a query is treated just as another document. Thus, the query vector is given by the following equation:

$$q = q^T \cdot \mathcal{U} \cdot \Sigma^1.$$

In the reduced k-dimensional space we can write

$$q = q^T \cdot \mathcal{U}_k \cdot \Sigma_k^{-1}.$$

Thus, since each row of the matrix $\mathrm{A_k}$ represents a document, the similarity between this query and the document could be evaluated based on the cosine similarity function presented in the (6.3).

Other method for document indexing is based on the **artificial intelligence techniques**: Artificial Neural Networks, Semantic Networks, or Bayesian Networks. The document is modeled through a richer and more complex knowledge representation of the domain, even though it sometimes entails a higher computational effort. Because is more related to our approach, we will briefly expose the semantic networks technique.

A *Semantic Network* (SN) is a directed graph, whose nodes are organized in hierarchic structures, while the arcs connecting them represent the binary relations between them, such as relations *is-a* and *part-of*. Domain taxonomies such as ACM or WordNet[14] could be considered SNs. WordNet's synsets are linked by different semantic relations and organized in hierarchies, as a semantic network.

Baziz et al. [3] present a system which adopts a document conceptual indexing method based on WordNet. The document is mapped on the WordNet and converted from a set of terms to a set of concepts (*Concept Detection* phase). The extracted concepts (single or multiwords) are then weighted as in the classical index term case, using a combination between TFxIDF weighting and Okapi BM25 Weighting.

Some other systems also adopt SN for document indexing, such as *SiteIF* [22] or the *WIFS* system [22].

The above presented approaches for textual documents indexing techniques could be exploited in the case of any field for acquiring automatic generation of ontology-based topic-related metadata. In the case of e-learning there are many such applications. For example in [31], the classification method is adopted for correlating learning resources with ACM taxonomy. Jovanovic et al. [20] adopt content-mining algorithms and heuristics to determine values of certain metadata elements used to annotate content units: title, description, unique identifier, subject (based on a domain ontology), and pedagogical role (based on an ontology of pedagogical roles).

The particularity of our approach is to separate first the LO content by its IEEE LOM metadata, to process the textual content in order to identify the ontology concepts that reflect its meaning, and further to integrate in the IEEE LOM metadata the concept-based information. Thus, the method is independent by the IEEE LOM standard, while contributing to the semantic enhancement of the specific metadata.

4.2 Ontology-Based Document Indexing Proposed Technique

We further present a technique for document indexing based on an ontology that combines TFxIDF indexing with latent semantic indexing and with WordNet-based processing.

Coming back to the e-learning use case, let us consider the general case of learning objects available into PDF or

14. http://wordnet.princeton.edu/.

HTML format, in which the titles and hyperlinks could be automatically located. We expose below a solution for automatically generating the ontology-based LO annotations inside the IEEE-LOM *Classification* category (as exposed in the end of Section 3.2), by combining:

- some existing Web information retrieval techniques,
- the latent semantic indexing technique, and
- a WordNet-based disambiguation technique.

In order to generate the three *relation types* between the document and the domain ontology concepts, the document is distributed in three classes:

1. The document *title and subtitles (headings)* will be considered for the *isOnTopic* relation;
2. The *external hyperlinks* encountered in the document body and the *bibliography* will be processed for the generation of the *makesReferenceTo* relation; and
3. The *document body* (the rest of the document) will be processed in order to obtain *usesTheConcept* relation.

Each of these classes is treated as a separated document and processed in the same manner. As result, the vector of concepts' weights corresponding to each document class is obtained. The XML annotations' format presented in Section 3.2 could be easily obtained from these three vectors through a simple transformation.

We will focus further to the following key problem: given as input a set of documents and a set of ontology concepts, how could it be obtained as output, for each document, a vector where each element corresponds to an ontology concept and its value is the concept's weight in document annotation.

Our solution to this problem pursues the following steps:

Step 1. Document preprocessing. Each document is loaded into memory and split into **tokens** (individual words). A token contains the original word, but also extended information like its stem and its word type. The *stem* is obtained by running Porter's stemming algorithm. We chose *Porter Stemmer* [29] because it is a well-established stemming algorithm, with well-understood properties. Concerning the *word type*, for the purpose of this paper, we are only interested if the word is a stop word or not. We compare each token (word) with a predefined stop-word list (the *Epnet* stop list) to determine its type.

Step 2. Term Frequency matrix and TFxIDF matrix. After document preprocessing, for each of the documents in the test set, a *frequency table* is calculated that contains the document's tokens and their frequencies, sorted in descending order. Tokens that are stop words are not taken into consideration. Also, if two tokens have the same stem, they are considered as one token, with the sum of their individual frequencies.

Having the frequency tables computed, a *master token array* is created from the union of each document's token set. The union is needed to ensure that any token in any document is contained in this array. The **Term Frequency (TF) matrix** is then created, having the terms (tokens) as rows and documents as columns.

Initially, the TF matrix contains at element *(i,j)* how many times the token i appears in document j. To prevent term frequency in a single document from influencing its relative importance, *normalization* is performed for each document. For example, if a document had two words, one occurring twice and the other occurring four times, the first word would be normalized to 2/6 and the other to 4/6.

The next step is obtaining the **TFxIDF matrix A** from the TF matrix. This is done by means of Inverse Document Frequency Indexing [17], a technique that attempts to smooth out the frequency of a word across documents. This prevents words that are common across documents to be given a large importance compared to words that seldom appear.

Step 3. Latent semantic indexing. Having computed the TFxIDF matrix, we now apply the *Singular Value Decomposition* method to this matrix, in order to reduce its dimensionality and also reveal latent relationships among documents based on word cooccurrences: $\mathbf{A} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V^T} \approx \mathbf{U_k} \cdot \mathbf{\Sigma_k} \cdot \mathbf{V_k^T}$.

A note worth mentioning is that the reduction from n to k dimensions will produce quite different results depending on the choice of k. Even though k is chosen by guessing, trial-and-error, it has been shown that its value should be between 100 and 300 for best results. For the current algorithm implementation, we choose k as the number of nonzero singular values obtained through the matrix $\mathbf{\Sigma}$. In case this dimension is larger than 100, we limit k to 100. We keep in memory only the reduced $\mathbf{V_k}$ matrix (see Section 4.1 for SVD decomposition and notations).

As illustrating example, we considered $m = 10$ *documents* from the computer science area. We obtained $n = 2{,}495$ *tokens*, so a matrix TFxIFD $\mathbf{A}$ having a dimension of $10 \times 2{,}495$.

As result of the Singular Value Decomposition, we obtained 50 positive singular values. We choose $\mathrm{k} = 50$ (representing approximately the SQRT(n)—a good value for smaller token sets like the one we tested on—choice was made by trial-and-error and best results were obtained around this value). Thus, the matrix $\mathbf{V_k}$ has the 10×50 dimension.

Step 4. Concept matrix construction (for each concept). After having computed the reduced dimensionality of the TFxIDF matrix, we will detect in steps 4 and 5 the most relevant documents for each ontology concept: given a concept and the collection of m document, we want to obtain a vector with dimension m, where each element represents the weight of this concept for a specific document.

For this purpose, we apply the Latent Semantic Indexing technique in the same way as for detecting the most relevant documents for a certain query (see Section 4.1). Our solution also includes WordNet-based processing, which will lead in step 4 to a matrix, each line corresponding to a concept synonym; this matrix will be reduced in step 5 to the mentioned vector.

As illustrating example, we will consider six concepts belonging to the H.3.3 category of the ACM topic hierarchy.[15]

In this step we aim to obtain, for each ontology concept, a matrix **S** that has the original ontology concept and related concept queries as rows, and documents as columns. Element *S(i,j)* of this matrix will represent the relevancy of the concept on row i with document j.

As exposed in Section 4.1, a user query over the document collection can be viewed from the Latent

15. http://www.acm.org/about/class/. This topic hierarchy received OWL representations: http://www.rkbexplorer.com/ontologies/acm.

Semantic Indexing technique's point of view as a document that contains only the query's words. We adopt the same approach considering each ontology concept as a user query, so as a document containing only the concept's words. In addition, we use *WordNet* to derive additional similar concepts to this master concept. Each similar concept is treated as a document also.

To do this, we take each word of the concept and build a synonym array. The synonyms are taken from WordNet's first synset of the concept word, because this synset usually contains the word's primary sense (the problem of choosing the correct synset falls into the fine-grained word-sense-disambiguation problem for which the high baseline is set by always choosing the first sense for any word from WordNet, a baseline that provides decent results). A similar concept is obtained by combining synonyms of each initial concept word. For example, in the case of the concept "selection process" from the ACM topic hierarchy, four synonyms for "selection" and two synonyms for "process" were detected

$$(0.50 \quad 0.78 \quad 0.87 \quad 0.28 \quad 0.51 \quad 0.59 \quad 0.52 \quad 0.77 \quad 0.18 \quad 0.33).$$

The above example shows the relevancy of "selection process" for every one of the 10 documents in a test collection. From their combination, eight similar concepts for the initial concept were built: [choice, procedure], [choice, process], [selection, procedure], [selection, process], [option, procedure], [option, process], [pick, procedure], [pick, process], [selection, process] (the concept itself is kept on the last supplementary, position). We search for expressions built using these words; we do not search for individual synonyms.

Each such concept is represented first as a sparse large vector (the size n of the master token array), with value 1 in the places where the concept's words match the master token array's words. The vector is reduced to k dimensions (LSI), and then compared to each of the documents in the test set. This gives us a row of our similarity matrix.

The distance between each LSI-reduced vector and each line of the TFxIDF matrix A that corresponds to a document, is calculated using the standard cosine metric. The mentioned matrix S is obtained, which is, in essence, a similarity matrix between documents.

For the considered example of the 10 documents collection and the particular concept [selection, process], a matrix with 9×10 is obtained, where the first eight lines correspond to the concept synonyms, while the last line includes the weights of the concept itself for each document.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
C1	0.70	0.80	0.94	0.28	0.70	0.77	0.49	0.88	0.39	0.34
C2	0.51	0.91	0.98	0.22	0.52	0.59	0.51	0.79	0.18	0.27
C3	0.41	0.94	0.97	0.21	0.41	0.50	0.54	0.75	0.06	0.25
C4	0.39	0.94	0.98	0.20	0.40	0.49	0.53	0.74	0.04	0. 24
C5	0.68	-0.1	0.21	0.88	0.72	0.83	0.71	0.85	0.40	0.91
C6	0.53	0.83	0.92	0.41	0.54	0.65	0.67	0.88	0.14	0.46
C7	0.93	-0.1	0.16	0.41	0.94	0.91	0.16	0.65	0.87	0.47
C8	0.76	0.74	0.90	0.34	0.77	0.83	0.50	0.91	0.45	0.40
C9	0.39	0.94	0.98	0.20	0.40	0.49	0.53	0.74	0.05	0.24

For each concept from the ontology we obtain one such similarity matrix, with m columns corresponding to document collection, and $s+1$ lines, corresponding to the similar concepts to the currents concept (represented on the last line).

Step 5. Concept matrix reduction. After we have built the concept matrices, we will reduce them to only one result matrix, showing the similarity between every concept from the ontology and every document in the test-set. To do this, we take each concept matrix and reduce it to one row. We apply the following formula for each column j (meaning for each document):

$$Weight[j] = \frac{1}{2} S[s+1, j] + \frac{1}{2} \sum_{i=1}^{s} \frac{s[i,j]}{s}.$$

We choose to give the main concept half the importance in the final weight, while all the concept's synonyms are averaged and given an equal share of the remaining half of the weight. Thus, we obtain a balance between the main concept and its synonyms.

For the considered concept "selection process" the above matrix is condensed to the following vector:

Step 6. Result matrix. In the previous step, we obtained a matrix row for each of the concepts in the ontology. We take each row and place it in the final result matrix R, which will be a similarity matrix of pxm dimensions, where p is the number of concepts in the ontology and m is the number of documents in our document set. The value R [i, j] represents the weight of concept i for the document j.

Entirely, a column j includes the weights of all the considered concepts for the document j. From this information, the XML annotation for the document j could be obtained through a simple transformation process. The relationType attribute (isOnTopic, usesTheConcept, makesReferenceTo) is completed according to the processed document class (titles, hyperlinks, body).

The implementation of this technique made use of the Apache Commons Mathematics Library[16] and of the JWI—the MIT Java WordNet Interface.[17]

Results evaluation. We performed an expert-based evaluation of the exposed indexing technique. The test was focused on seven concepts from the H.3.3 section of the ACM topic hierarchy (Information Search and Retrieval): [clustering], [information, filtering], [query, formulation], [relevance, feedback], [retrieval, models], [search, process], [selection, process]. The 10 test documents were selected from the computer science literature such as to be on the topic of the considered concepts: two documents for each of the first three concept, and one documents per each of the remaining concepts. The resulted matrix for this test is

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
C1	0.74	0.04	0.26	0.18	0.74	0.70	0.05	0.50	0.70	0.23
C2	0.42	0.17	0.32	0.95	0.47	0.64	0.93	0.85	0.03	0.97
C3	0.26	0.88	0.87	0.01	0.26	0.31	0.34	0.53	0.01	0.04
C4	0.84	-0.1	0.10	-0.1	0.82	0.69	-0.4	0.28	0.99	-0.1
C5	0.86	-0.1	0.09	-0.1	0.85	0.72	-0.3	0.33	0.99	-0.1
C6	0.61	0.49	0.65	0.69	0.63	0.76	0.77	0.93	0.23	0.73
C7	0.51	0.79	0.87	0.28	0.51	0.59	0.52	0.77	0.18	0.33

For example, in this matrix, the last line (C7) corresponds to the "selection process" concept, while the eight column (D8) corresponds to a document that was selected by a human expert as being on "selection process" topic. We could notice that on the concept line, the eight value 0.77 is

16. http://commons.apache.org/math/.
17. http://projects.csail.mit.edu/jwi/.

indeed the biggest (the document D8 is the most related to this concept), but on the D8 column, the biggest value 0.93 is on the sixth line, corresponding to the "retrieval models" concept, while the second value 0.85 (corresponding to "information filtering" concept C2) is still superior to 0.77. Indeed, these two concepts are important for the document D7; however, "selection process" could be considered its dominant concept. For bigger document collections and a bigger set of concepts the results are similar. Different results are obtained by varying the value of k in the SVD decomposition: the previous results are obtained in the test with 2,495 tokens from 10 documents, for $k = 50$. For bigger values of k (300, 500, 1,000), the results are less pertinent for our data set.

So, despite the fact our proposed technique is correct from a conceptual point of view (because it combines multiple well-established techniques, with foreseeable results), there is still place for improvements with respect to accuracy.

We intend to improve our technique through a more accurate document preparation in the first step—document preprocessing, part of speech tagging and better word sense disambiguation. Currently, all the words that aren't stop words are considered as tokens. We will filter these tokens through part-of-speech tagging process, keeping only the nouns because the concepts are composed of nouns.

As well, we will split our technique in two, one adopting latent semantic indexing and the other focused on the WordNet exploitation, and we will then compare the results obtained in the three cases against the same test collections.

The particularity of the exposed indexing technique consists in the idea of separately processing the three parts of the document in order to generate semantic differentiated annotations. Also, the idea of combining the singular value decomposition technique with a WordNet-based processing phase in order to obtain ontology-based annotations is particular to our approach.

Performance-wise, the proposed technique currently has two bottlenecks: synonym extraction and SVD decomposition. The synonym extraction speed is as fast as the current WordNet wrappers allow, given that WordNet is kept as an on-disk dictionary file. If the dictionary was kept in RAM memory, the speed would be greatly improved. The second bottleneck is the SVD matrix decomposition, part of the LSI process. We have used Apache Commons Math Library[18] for the SVD computation, providing good performance. Worst case scenario, SVD decomposition is at most $O(n^3)$ expensive; however, there are techniques like the Jacobi algorithm and smart use of the QR decomposition that provide decent performance even when computing SVD for large matrices. We have tested SVD for matrices of at most 10^5 columns, solving in less than 2 minutes, smaller tests ($<1{,}000$ columns) being computed almost instantly on a single-core 2 GHz processor. The level of performance is good, considering that computing the SVD matrices is done offline where small time spans are allowable (we estimate that a large collection of 10^6 documents having 10^3 extracted terms indexed by a 10^3 ontology can be done in a matter of a few hours). The total time depends heavily on the SVD decomposition (performance that can be greatly improved if necessary), and in a smaller measure on WordNet synonym extraction. However, the tasks that do not depend on SVD decomposition are executed very fast, like reading and parsing the documents from the disk or memory, later stages matrix reductions.

We accomplished our evaluation with existing resources, especially with a limited resources collection. We intend to extend this evaluation using the resources provided by Ariadne Finder repository.[19]

5 Conclusions and Further Work

The resources available inside an e-learning system should be accessible and connected. Moreover, the actual decentralization and interinstitution collaboration that characterize the e-learning solutions increase the demand of making learning resources accessible not only across e-learning platforms, but also across other Web applications. This demand could be solved with the support of semantic technologies, as the solution presented in this paper illustrates. It extends the IEEE LOM standard with ontology-based semantic annotations for efficient use of LOs outside Learning Management Systems. The proposed method is automatically developed in the case of textual LOs by combining semantic technologies with Natural Language Processing techniques: TFxIDF indexing, Latent Semantic Indexing, and WordNet-based processing.

The paper results are important in the context of adopting Semantic Web technologies into the e-learning field, but also as a progress in the area of ontology-based indexing of the textual resources.

We have chosen IEEE LOM standard because it was proven as permissive for semantic extensions. We adopted ACM topic hierarchy as use case for annotations because we considered the case of LOs from computer science field, and other similar approaches considered it as well. The presented approach remains however ontology independent since the considered ontology could be easily replaced when required.

The presented approach could be exploited in multiple applicative scenarios, as mentioned in Introduction.

Our future work will concern both directions considered by the paper. The acquired results for ontology-based LOs annotation and users modeling will be exploited in order to develop various personalized functionalities across the different LMS. As well, the LOs indexing technique presented in this paper will be enhanced and confronted to the existing classification-based methods, which constitute the third alternative for structure-based indexing of textual resources, alongside with the mathematical and with the linguistic-oriented approaches.

Acknowledgments

This work was supported by the SOMIR project PIEF-GA-2009-235229 and by the Sectoral Operational Programme Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/6/1.5/S/19.

18. http://commons.apache.org/math/.

19. http://ariadne.cti.espol.edu.ec/.

REFERENCES

[1] *Information Retrieval and Hypertext,* M. Agosti and A.F. Smeaton, eds. Kluwer Academic, 1997.

[2] H.S. Al-Khalifa and D. Hugh, "The Evolution of Metadata from Standards to Semantics in E-Learning Applications," *Proc. Seventh Conf. Hypertext and Hypermedia (Hypertext '06),* 2006.

[3] M. Baziz, M. Boughanem, and S. Traboulsi, "A Concept-Based Approach for Indexing Documents in IR," *Proc. Actes du XXIII-eme Congres INFORSID,* 2005.

[4] M.W. Berry, S.T. Dumais, and G.W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval," *SIAM Rev.,* vol. 37, no. 4, pp. 573-595, 1995.

[5] J. Brase and W. Nejdl, "Ontologies and Metadata for eLearning," *Handbook on Ontologies,* Springer-Verlag, 2003.

[6] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems,* vol. 30, nos. 1-7, pp. 107-117, 1998.

[7] J. Broisin, M. Brut, V. Butoianu, F. Sedes, and P. Vidal, "A Personalized Recommendation Framework Based on CAM and Document Annotations," *Proc. RecSysTel Workshop,* 2010.

[8] P. Brusilovsky and D.W. Cooper, "Domain, Task, and User Models for an Adaptive Hypermedia Performance Support System," *Proc. Int'l Conf. Intelligent User Interfaces,* Y. Gil and D.B. Leake, eds., pp. 23-30, 2002.

[9] P. Brusilovsky, J. Eklund, and E. Schwarz, "Web-Based Education for All: A Tool for Developing Adaptive Courseware," *Proc. Seventh Int'l World Wide Web Conf.,* vol. 30, H. Ashman and P. Thistewaite, eds., pp. 291-300, 1998.

[10] P. Brusilovsky and E. Millán, "User Models for Adaptive Hypermedia and Adaptive Educational Systems," *The Adaptive Web,* P. Brusilovsky, A. Kobsa, and W. Nejdl, eds., pp. 3-53, Springer, 2007.

[11] M. Brut and S. Buraga, "A Web Service-Based Solution for E-Learning Resources Annotation and Retrieval - Conceptual Architecture," *Proc. Int'l Joint Conf. Computer, Information, and Systems Sciences and Eng. (CISSE '07),* 2007.

[12] M. Brut, F. Sedes, T. Jucan, R. Grigoras, and V. Charvillat, "An Ontology-Based Modeling Approach for Developing a Competencies-Oriented Collective Intelligence," *Proc. ED-L2L Conf. World Computer Congress (WCC '08),* 2008.

[13] P. De Bra and J.-P. Ruiter, "AHA! Adaptive Hypermedia for All," *Proc. World Conf. WWW and Internet (WebNet '01),* W. Fowler and J. Hasebrook, eds., pp. 262-268, 2001.

[14] S.T. Dumais, "Latent Semantic Indexing (LSI) and TREC-2," *Proc. Text REtrieval Conf. (TREC '93),* pp. 105-116, 1993.

[15] M. Engelhardt, A. Hildebrand, D. Lange, and T.C. Schmidt, "Reasoning about eLearning Multimedia Objects," *Proc. Int'l Workshop Semantic Web Annotations for Multimedia,* 2006.

[16] G.H. Golub and C.F.V. Loan, *Matrix Computations,* second ed. Johns Hopkins Univ., 1989.

[17] D.A. Grossman and O. Frieder, *Information Retrieval: Algorithms and Heuristics,* second ed. Springer, 2004.

[18] N. Henze and W. Nejdl, "Adaptation in Open Corpus Hypermedia," *Int'l J. Artificial Intelligence in Education,* vol. 12, no. 4, pp. 325-350, 2001.

[19] *IEEE 1484.12.1-2002, Draft Standard for Learning Object Metadata,* IEEE, http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf, 2002.

[20] J. Jovanovic, D. Gaševic, and V. Devedi, "Ontology-Based Automatic Annotation of Learning Content," *Int'l J. Semantic Web and Information Systems,* vol. 2, no. 2, pp. 91-119, Apr.-June 2006.

[21] C.-C. Kiu and C.-S. Lee, "Ontology Mapping and Merging through OntoDNA for Learning Object Reusability," *Educational Technology and Soc.,* vol. 9, no. 3, pp. 27-42, 2006.

[22] B. Magnini and C. Strapparava, "User Modelling for News Web Sites with Word Sense Based Techniques," *User Modeling User-Adapted Interaction,* vol. 14, nos. 2/3, pp. 239-257, 2004.

[23] A. Micarelli, F. Sciarrone, and M. Marinilli, "Web Document Modeling," *The Adaptive Web: Methods and Strategies of Web Personalization,* P. Brusilovsky, A. Kobsa, and W. Nejdl, eds., Springer, 2007.

[24] B. Mobasher, "Data Mining for Web Personalization," *The Adaptive Web: Methods and Strategies of Web Personalization,* P. Brusilovsky, A. Kobsa, and W. Nejdl, eds., Springer, 2007.

[25] P. Mohan and C. Brooks, "Learning Objects on the Semantic Web," *Proc. Int'l Conf. Advanced Learning Technologies,* 2003.

[26] M. Nilsson, M. Palmér, and J. Brase, "The LOM RDF Binding-Principles and Implementation," *Proc. Third ARIADNE Conf.,* 2003.

[27] Z. Obrenovic et al., "Multimedia Semantics: Overview of Relevant Tools and Resources," *Web Consortium,* http://www.w3.org/2005/Incubator/mmsem/wiki/Tools_and_Resources, 2007.

[28] H. Passier and J.T. Jeuring, "Ontology Based Feedback Generation in Design-Orientated E-Learning Systems," *Proc. IADIS Int'l Conf. E-Soc.,* P. Isaias, P. Kommers, and M. McPherson, eds., pp. 992-996, 2004.

[29] M.F. Porter, "An Algorithm for Suffix Stripping," *Program,* vol. 14, no. 3, pp. 130-137, 1980.

[30] C.J.V. Rijsbergen, *Information Retrieval,* second ed. Butterworths, 1979.

[31] P.S. Saini, M. Ronchetti, and D. Sona, "Automatic Generation of Metadata for Learning Objects," *Proc. Sixth IEEE Int'l Conf. Advanced Learning Technologies (ICALT '06),* 2006.

[32] G. Salton and M. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[33] B. Sarwar, G. Karypis, J.A. Konstan, and J. Riedl, "Incremental SVD-Based Algorithms for Highly Scaleable Recommender Systems," *Proc. Int'l Conf. Computer and Information Technology,* 2005.

[34] M. Stratakis, V. Christophides, K. Keenoy, and A. Magkanaraki, "E-Learning Standards," SeLeNe (Self E-Learning Networks IST-2001-39045), Project Deliverable 2.1, Greece, 2003.

[35] *Web Semantics Ontology,* D. Taniar and J.W. Rahayu, eds. Idea Group Publishing, 2006.

[36] X. Ochoa and E. Duval, "Use of Contextualized Attention Metadata for Ranking and Recommending Learning Objects," *Proc. First Int'l Workshop Contextualized Attention Metadata: Collecting, Managing and Exploiting of Rich Usage Information,* 2006.

Mihaela M. Brut received the PhD degree in computer science in 2008 and the PhD degree in humanities in 2000. She holds a lecturer position with the Faculty of Computer Science from Al. I. Cuza University of Iasi, Romania. She currently holds a Marie-Curie Intra-European Fellowship for Career Development at IRIT. Her research is focused on the areas of document and user modeling, techniques of personalized recommendations, e-learning, semantic web, semantic annotations, and multimedia information indexing and retrieval. She is/was a member in 11 research projects with European partners (one FP7, one FP6, one ITEA, three Leonardo da Vinci, and five national Romanian projects).

Florence Sedes is a full professor of computer science at the University Paul Sabatier Toulouse III and has multiple coordination duties in research: she is the head of a research group of IRIT, a member of the French National University Council for Computer Science, and an expert for the International Research Board of the French Research Ministry. She leads the national research network of the CNRS "Information-Interaction-Intelligence" (GDR I3), focusing on the computerized processing of data, information, and knowledge. She has directed multiple PhD theses and she currently coordinates the academic part of the Lindo ITEA2 project, leading the multimedia-indexing cluster. Her research interests are concentrated around indexing, querying, and managing semistructured and multimedia documents and their adaptation in different applications. She is a senior member of the IEEE and the IEEE Computer Society.

Stefan Daniel Dumitrescu graduated from the Computer Science Department, Automatic Control and Computers Faculty, Politehnica University of Bucharest, Romania. He is currently working toward the PhD degree in the field of semantic technologies at the Politehnica University of Bucharest. His interests range in informatics applications with emphasis on information extraction, knowledge representation, and human-computer interaction. He is also interested in project management and is skilled in network technologies.