

Automatic Extraction of Pedagogic Metadata from Learning Content

Devshri Roy, Sudeshna Sarkar, Sujoy Ghose, *Computer Science & Engineering Department, Indian Institute of Technology, Kharagpur, India*
droy@cse.iitkgp.ernet.in, sudeshna@cse.iitkgp.ernet.in, sujoy@cse.iitkgp.ernet.in

Abstract. Annotating learning material with metadata allows easy reusability by different learning/tutoring systems. Several metadata standards have been developed to represent learning objects and courses. A learning system needs to use pedagogic attributes including *document type*, *topic*, *coverage of concepts*, and for each concept the *significance* and the *role*. Moreover, in order to have a flexible and reusable repository of e-learning materials, it is necessary that the annotation of the documents with such metadata be done in an automatic fashion as far as possible. This paper describes the attributes that represent some important pedagogic characteristics of learning materials. To reduce the overhead of manual annotation we have explored the feasibility of automatic annotation of learning materials with metadata. This facilitates the creation of an e-learning open repository for storing these annotated learning materials, which can be used by learning systems. The automatic annotation is based on a domain knowledge base and a number of algorithms like standard classification algorithms, parsing and analysis of documents have been used for this purpose. The results show a fair degree of accuracy, which may be improved in future using more sophisticated algorithms.

Keywords. Learning object metadata, automatic metadata extraction, open repository

INTRODUCTION

The wide availability of content in the electronic media has given rise to new paradigms of learning and knowledge delivery. E-learning has emerged as a very promising application. However, the development of e-learning systems is expensive in terms of the time and effort required. A lot of effort is required by the content author to develop high quality learning materials. But in the age of the Internet, a lot of the content pertinent to a course may be available from the web and other sources e.g. as parts of other courses. The challenge is to make such materials usable to satisfy the specific requirements of a given learner pursuing a given course.

Metadata is an important step towards the semantic tagging of documents including learning materials. Traditionally, metadata is created by humans and it is a labour intensive activity. It is costly to create metadata, and impractical if we consider a large set of potential materials obtained from the World Wide Web and other sources. Besides when not done carefully by experts, the process may be error prone. Web search engines provide a low cost index access to a large portion of documents on the Internet, and this has proved to be of immense value. However this indexing is not suitable for more sophisticated retrieval tasks that require more detailed tagging of documents.

Existing courses developed in some contexts may contain many lessons on topics relevant to some other courses. However, some learning management systems (Weber et al., 2002; Simic, 2004) follow very specific format requirements. Therefore, the extent of tagging required varies between these learning management systems. This acts as an impediment to easy reuse or adaptation of

materials from other sources. Lessons developed for one learning management system cannot be effortlessly reused in another system. Thus there is a pressing need for standardization to facilitate reuse of learning materials. Some such standards have been developed, like IEEE LOM, SCORM etc. However even when a standard is adopted, the authors of the e-learning courses have the responsibility of manually associating metadata to learning objects. Many authors find the task of manual annotation and assigning of meta-tag uninteresting, and are reluctant to do the tagging satisfactorily. Moreover, the Internet includes a huge storehouse of information on various topics. These materials are usually not tagged in accordance with the requirements of a specific learning management system. Many people feel (Ochoa et al., 2005) that unless the process of annotating learning objects can be automated, it is difficult to create a critical mass of reusable learning objects. The survey (Friesen, 2004) confirms that metadata instantiation is a difficult task, and the correct instantiation requires educational and technical skills.

In order to overcome the problems of manual annotation of documents, researchers are working on automatic creation of learning object metadata (Downes, 2004; Duval & Hodgins, 2004; Simon et al., 2004). The *Bibliographic Control of Web Resources: A Library of Congress Action Plan* (LC Action Plan) (<http://lcweb.loc.gov/catdir/bibcontrol/actionplan.pdf>) recognizes this need and highlights *automatic metadata generation tool development* as a “near-term/high” priority. *LC Action Plan* Section 4.0 targets the development of “automatic tools... to improve bibliographic control of selected Web resources,” and Section 4.2 specifically identifies the need for a master specification to guide development of such applications. According to automatic indexing developments (Anderson & Perez-Carballo, 2001), automatic metadata generation is *more efficient, less costly, and more consistent* than human-oriented processing.

One can envisage a flexible learning system that is able to make use of content from a wide range of sources. Our objective is to facilitate building an open repository containing tagged learning materials, so that a learning system is able to select appropriate learning materials for a learner from this repository. We wish to simplify the process of creating this learning repository so that new documents can be incorporated into this repository with very little manual effort. The course designer should be able to put together a course following a specific curriculum provided that there is a repository of content with meta information on each of the topics required in the course. Moreover, such meta information should include pedagogic attributes so that a learner/tutoring system can make the appropriate choice of learning material from the repository in a flexible manner depending on the current requirement.

The basic idea motivating our work is to separate the learning materials from the course structure in order to reuse them in new situations. We wish to work on automatic annotation of learning materials. We are mostly interested in pedagogic attributes that characterize the content of learning material from an educational point of view, such as its level and pedagogic style. These are the attributes that a learning system can use to identify relevant lessons for an individual learner.

The work presented in this paper focuses on two aspects:

1. Identification of a set of pedagogic metadata relevant for personalized learning.
2. Development of algorithms to extract such metadata automatically from documents and thus automate the creation of open repository of learning materials.

If the latter can be done satisfactorily it will be possible to index these learning materials with the metadata constituting the topic and coverage of the material, as well as with its pedagogic type. This has wide implications in developing intelligent tutoring systems. Even without a tutoring system, an

individual learner using this repository for self-study will find the annotations quite helpful. However our work is confined to metadata annotation.

The rest of the paper is organized as follows. First we discuss the different metadata standards, learning object repositories and prior work done on automatic metadata generation. Following this we briefly discuss the structure of domain knowledge that is manually constructed and used by the metadata extractor algorithms. The metadata used in our work are discussed next. Then we present different algorithms for automatic extraction of metadata and also the performance evaluation of different algorithms. Finally, the automatic annotation tool is discussed.

PRIOR WORK

Metadata Standards

Several metadata standards have emerged for the description of resources. The Dublin Core metadata initiative (<http://dublincore.org/>) is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. IEEE LOM (<http://ltsc.ieee.org/wg12/index.html>) aims to develop accredited technical standards, recommended practices, and guides for learning technology. The IMS Global Learning Consortium (<http://www.imsglobal.org/>) develops and promotes the adoption of open technical specifications for interoperable learning technology. The Advance Distributed Learning (<http://www.adlnet.org>) initiative aims to establish a distributed learning environment that facilitates the interoperability of e-learning tools and course content on a global scale. The Sharable Content Object Reference Model (<http://www.adlnet.org/scorm/index.cfm>) is a set of specs by ADL concerning developing, packaging and delivering learning objects.

Learning Object Repositories

Many groups such as Health education assets library (<http://www.healcentral.org/index.htm>), Education network Australia online (<http://www.edna.edu.au/edna/page1.html>), SMETE digital library (<http://www.smete.org/>), iLumina (<http://www.ilumina-dlib.org>), LearnAlberta online curriculum repository (<http://www.learnalberta.ca/login.aspx>), Campus Alberta repository of educational objects (<http://careo.ucalgary.ca/cgibin/WebObjects/CAREO.woa/wa/Home?theme=careo>) have worked on developing learning object repositories. Their primary goal is the creation of a searchable web-based collection of multidisciplinary teaching materials for educators and learners. In addition to a simple search, these repositories offer search by certain advanced features such as *type of learning resource*, *title*, *subject*, *grade*, *author*, *primary audience* etc. In order to search and retrieve documents, learning object repositories store learning objects and metadata. In all the above-mentioned repositories, the learning objects are annotated with metadata manually by the authors/contributors.

Automatic Generation of Metadata

ARIADNE (<http://www.ariadne-eu.org>), the European digital library project has started work on automatic annotation of documents. Kris Cardinaels et al. (2005) developed a framework for automatic metadata generation of a metadata set that contains all the mandatory elements defined in the

ARIADNE application profile. The metadata generation framework of Cardinaels automatically generates a few of the IEEE learning objects metadata, namely, *document type*, *package size*, *publication date*, *creation date*, *operating system type*, *access right*, *main discipline*, *language*, *format*, *title*, and *author's detail*.

Some work has been done on automatic generation of Dublin Core metadata. UK Office for Library and Information Networking (UKOLN) has developed an automatic metadata generator tool DC-dot (<http://www.ukoln.ac.uk/cgi-bin/dcdot.pl>). The set of Dublin Core metadata, which is generated by DC-dot includes *identifier*, *subject*, *title*, *keywords*, *description*, *type*, and *format*. DC-dot simply harvests the above set of metadata from the resource META tag. Similarly, in the work done by Jenkins and Inman (2000) on automatic generation of Dublin Core Metadata, *title*, *date*, and *format* are harvested from the html documents, but the keywords are extracted by parsing the actual content. Hui Han et al. (2003) proposes a machine learning method using a support vector machine for automatic extraction of Dublin Core metadata.

Context-driven and pattern based annotation through knowledge on the web (C-PANKOW) is a method for automatic semantic annotation of web content. The main idea here is to approximate semantics by considering information about the statistical distribution of certain syntactic structure over the web (Cimiano et al., 2005). It focuses on identifying domain topics in analyzed documents. KNOWITALL (Etzioni et al., 2004) is an autonomous domain independent system that automates the process of extracting large collections of fact from the web. The KIM platform and framework provides services for semantic annotation, indexing and retrieval of documents (Popov et al., 2003). The platform is based on the PROTON ontology (<http://proton.semanticweb.org>) as well as KIM system ontology and KIM lexical ontology.

Most of the learning object repositories provide the facility of advanced search on different search parameters like *subject*, *subcategories (topics)*, *learning resource types*, *content URL*, *primary audience* etc. Therefore it is important to know all this metadata information of a document and researchers are working on automatic extraction of this metadata. The framework of automatic generation given by Cardinaels (2005) generates subject (main discipline) and keywords from documents. The top-level classification gives the main discipline and the lower level classifications give keywords for the document. For example, XML is one of the topics of the course Multimedia. The documents on the topic XML are structured in folders like Multimedia/XML and the keywords of the documents are identified from this taxonomic path. Li et al. (2004) also identifies the subject/classes of the WebPages. In their approach, a term weight vector in multidimensional space represents the whole content of the resource. In the vector, each word is assigned a weight, which represents its degree of importance. The subject/class of a resource is identified from the weighted vector using the principal component analysis (PCA) technology in neural network. The ten major classes, which are identified in their work, are news, entertainment, community, sports, health, finance and economics, living information, science and technology, people and shopping.

Gelbukh et al. (1999) has given a method of document classification on a hierarchical dictionary of topics. The dictionary consists of two major parts, *vocabulary* containing keywords, and a hierarchical structure representing topics. The hierarchical links in the dictionary are supplied with the weights representing the probability for a word in a particular context to be really related to a given topic. For example, the word *Italy* belongs to the topic *Europe*, thus, the weight of this link is 1. On the other hand, the weight for the link between *Italy* and *England* is much less. To obtain the topic of the document, the keywords in a document are compared with a hierarchical dictionary of topics.

It is important to know the type of a document and in the work of Jovanovic et al. (2006), they automatically identify some of the types like definition, example and reference type. We feel that identifying the pedagogic type of the documents (whether it is exercise, explanation, experiment, etc.) is also important from the instructional design perspective.

In the educational domain, identifying the topic of a document is very important, and the challenge is to identify the topic automatically by analyzing the content of the document. It is important to identify the *subcategory/topic* of the document along with the subject of the document. In the case of identification of subcategories of a subject, there will be many common terms in more than one subcategory, which makes this task challenging.

STRUCTURE OF DOMAIN KNOWLEDGE

The success of any e-learning system depends on the organization of learning objects with respect to a domain knowledge structure, and the facility to retrieve relevant learning materials. According to many researchers (Song et al., 2005; Tan & Goh, 2004), the domain ontology plays a crucial role for the development of a flexible educational system. Dicheva et al. (Dicheva et al., 2004; Dicheva & Dichev, 2004; Dicheva et al., 2005) proposed a framework for building a concept-based digital course library where the subject domain ontology is used for classification of course library content. They proposed a layered architecture of the repository, with each layer capturing different aspects of the information space such as conceptual, resource related and contextual. The semantic (ontology) layer contains a conceptual model of the knowledge domain in terms of key concepts and the relationship among them. Hoermann et al. (2003) proposed an approach of using learning object metadata together with a well-defined knowledge base in order to create adaptive and modularized courses. The knowledge base stores the keywords of the domain and semantic relations between these keywords. In the work of Gasevic et al. (2005) and Jovanovic et al. (2006), the domain ontology stores concepts describing the documents and their relationships and is used to semantically mark up the content of learning objects. Aitken and Reid (2000) have used the domain ontology in an information retrieval tool. To improve the management of information, Lauser et al. (2002) have created a prototype biosecurity ontology on food safety, animal and plant health domain. The categories are the generic concepts and are connected to the specific instances.

In our system, the knowledge base is organized into a three-layered hierarchical structure as shown in Figure 1. The three layers are the *term layer*, the *concept ontology* and the *topic taxonomy*. The *term layer* of our knowledge base stores lexical terms. Lexical terms are the raw terms or representative keywords that occur in documents. It may be noted that a lexical term can be polysemous and it may have different meanings in different contexts. While a term can have different meanings, a domain specific concept is unambiguous and can be useful for retrieving domain specific documents. Therefore, the second layer of the knowledge base i.e. the *concept ontology* contains domain specific concepts of various subject domains and relationships among concepts similar to the ontology used by Dicheva and Dichev (2004), Hoermann et al. (2003), Gasevic et al. (2005) and Jovanovic et al. (2006). These concepts are used for textual content analysis of learning resources useful for the automatic extraction of metadata from them.

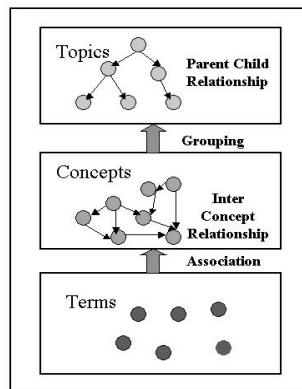


Fig.1. Knowledge Base: Three level hierarchical structure.

We have developed the concept ontology for three subjects, namely physics, biology and geography. They cover a total of around 3400 concepts. In addition to the above two layers, we have added another layer on the top i.e. the *topic taxonomy*. The motivation for adding *topic taxonomy* is that in many learning situations, especially in formal school and college education systems, the learning requirements are often mentioned in terms of topics. A topic may introduce or discuss a single concept, but it is not always synonymous with a single concept. Often a topic discusses several concepts. To make this distinction, we have added a different layer called *topic taxonomy*. The *topic taxonomy* of our knowledge base contains three high-level categories physics, biology and geography corresponding to a broad division of the domains covered by us. Each category is divided into subcategories. The total number of topics covered in the topic taxonomy is around 220. Each topic of the topic taxonomy keeps a list of concepts that are discussed in that topic. This is similar to the approach adopted in the works of (Biemann, 2005) and (Lauser et al., 2002).

In our domain knowledge, we map the entities of one layer to other layers by keeping relationships between entities of different layers. We keep the relationship between topics and concepts and also between concepts and keywords.

METADATA SCHEMA

As discussed earlier, there are several metadata standards. The IEEE learning object metadata provides a comprehensive description of learning resources. In IEEE LOM, an elaborate hierarchical scheme has been developed that includes the categories of *general*, *lifecycle*, *metametadata*, *technical*, *educational*, *rights*, *relations*, *annotation*, and *classification*. What we find especially relevant is the educational category, which includes elements such as *Interactivity type*, *Learning resource type*, *Interactivity level*, *Intended end user role*, *Context*, *Typical age range*, *Difficulty*, *Typical learning time*, *Language of the typical intended user* and *Description*.

We have adapted many attributes from the IEEE LOM, which are relevant for finding the suitability of a document to a particular learner. Even though we would like to include many of the attributes from IEEE LOM, currently our system deals with a subset of the IEEE LOM attributes. In addition to the subset of the IEEE LOM attributes, we also suggest some minor enhancement to the set

of metadata, which appears to be useful. The metadata attributes that we have currently worked on are given below in Tables 1 and 2.

Table 1
A subset of IEEE LOM attributes

1. <i>General category</i>	// General information describing this learning object.
1.1 <i>Identifier</i>	// A globally unique label that identifies the learning object.
1.4 <i>Description</i>	// A textual description of the content of the learning object.
3. <i>Meta Metadata category</i>	// This category describes the metadata record itself.
3.1 <i>Identifier</i>	// Unique label that identifies the learning object.
4. <i>Technical category</i>	// Describes the technical requirements and characteristics of this learning object.
4.1 <i>Format</i>	// Technical data type(s) of this learning object.
4.2 <i>Size</i>	// The size of the digital learning object in bytes.
4.3 <i>Location</i>	// A string that is used to access this document.
5. <i>Educational category</i>	// This category describes the key educational or pedagogic characteristics of the learning object.
5.2 <i>Learning Resource Type</i>	// Specific type of learning material. The types considered are // <i>Narrative text, questionnaire, and experiment.</i>
9. <i>Classification category</i>	// This category describes where this learning object falls within a // particular classification system.
9.2 <i>Taxonomic Path (topic)</i>	// Taxonomic path with respect to the topic tree in the domain // knowledge.

Table 2
Extension of IEEE LOM Specification

<i>Myvoc1.1 List of concepts</i>	// List of concepts mentioned that belong to the domain ontology // along with certain attributes for each concept.
For each concept we specify	
<i>Name</i>	// Name of the concept.
<i>Significance</i>	// Significance of the concept.
<i>Type</i>	// A concept can be one of these 2 types: Outcome or prerequisite.
<i>Myvoc1.2 List of domain terms</i>	// List of domain terms in the learning material along with their // frequency.
For each term we specify	
<i>Name</i>	// Name of the term.
<i>Frequency</i>	// Its frequency of occurrence in the document.

We now explain the importance of the concept related attributes given in Table 2. The repository is a pool of learning materials. To incorporate the facility of retrieval of learning materials, a set of document terms is extracted from the document. Since identifying the correct sense of a term is of

fundamental importance to achieve good IR performance (Baeza-Yates & Ribeiro-Neto, 1999; Salton & McGill, 1984) and concept based search gives higher precision for retrieval (Aitken & Reid, 2000; Henstock et al., 2001). Hence the *concepts* seem to be a more useful notion than the lexical *terms*. The *significance* and the *type* are two important attributes of concept, which would be very useful to a tutoring or e-learning system. We find the list of concepts from the document using the concept ontology. But all concepts that occur in a document are not equally useful for characterizing the document. We use the frequency of domain terms indicating the concept in the document as one attribute of the concept. We also keep a separate attribute for representing the *significance* of the concept that indicates whether a concept occurs along with its related concepts in the document. Further, a concept may be defined or introduced in a document, or it may be used to explain other concepts. A concept that is being defined in a document gives the *outcome* concept whereas the knowledge of the other concepts used for explaining the outcome concept are *prerequisites* for studying the document.

We now discuss the importance of the pedagogic attributes of Table 1. We have used the *classification* metadata to represent the topic taxonomy of learning materials. Corresponding to each topic, from the topic taxonomy, we can retrieve the taxonomy path, which we specify as a metadata. A learning/tutoring system especially in the context of school or college curriculum may need to identify documents belonging to a topic or a subtopic. It is also important to identify the educational category of a document, which includes the learning resource type to assess its relevance for learning in a given situation. The objective of an educational system is to provide a personalized learning experience. Different learners require different learning content depending on their learning style (Papanikolaou et al., 2002). In the context of instructional design the learning resource type (IEEE LOM's property 5.3) such as *exercise*, *simulation*, *narrative text*, *exam* and *experiment* cover the instructional type. A few more values have been proposed by Ullrich (2004) as an extension to the LOM resource type, which describes the learning resource from an instructional perspective. Many researchers (Mohan & Greer, 2003; Brooks et al., 2005; McCalla, 2004; Liu & Greer, 2004) have recommended and suggested extensions to the IEEE LOM. RDN/LTSN (<http://www.rdn.ac.uk/publications/rdn-ltsn/types/>) resource type vocabulary is very common in the UK learning and teaching community. It specifies a set of additional learning resource type that is used with a 5.2 learning resource type LOM element. Appendix 2 UK and European Type and Learning Resource Type Vocabularies of UKLOM core (<http://www.cetis.ac.uk/profiles/uklomcore>) describes a few more metadata schemas, which specify a set of learning resource type vocabulary.

In this paper, we have handled the automatic classification of documents into three categories, namely *narrative text*, *questionnaire* and *experiment*. Handling other categories would be an important future extension.

AUTOMATIC METADATA EXTRACTION

Different metadata extraction algorithms are implemented to generate the set of metadata given in Table 1 and Table 2 from documents. To evaluate the performance of the different metadata extractor algorithms, the results are compared with manually annotated documents by human evaluators. Documents are collected from diverse sources. The web is the main source for collection of the documents. We have also collected documents on different topics from some books written by

different authors by scanning the content of the books. Different algorithms for extracting pedagogic metadata from learning contents are discussed below.

Concept and its Significance Identification

A document contains a set of terms. Since a term may have multiple meanings while a concept is unambiguous, it is more useful to annotate a document with a list of concepts rather than terms. For this, we need to disambiguate the meaning of a term, and identify the concept it refers to. In some cases more than one term may refer to the same concept. In such cases, the frequency of a concept will include the frequencies of all synonymous terms for the concept in the document.

We maintain a dictionary of terms in the knowledge base. Corresponding to each term, we keep a link to the possible set of concepts that the term may refer to. We identify the terms and their frequencies from each document. For each term, the corresponding set of concepts is found from the dictionary. Out of these candidate concepts, we want to find the most appropriate concept. We note that concepts rarely occur in isolation. If a concept is significant for a document, the document usually contains other concepts related to it. For example if a document talks about *kinetic energy*, the document usually contains other terms like *motion*, *mass*, *velocity* etc. while in the case of *light energy*, the document may contain terms like *sun light*, *solar light* etc.

To find the most appropriate concept, we use the inter concept relationship which is captured in the concept ontology. A concept is more significant if more related concepts of that word occur in the document. The proposed algorithm takes the terms with their frequency as input, and returns a list of concepts along with the significance of each concept. The algorithm works as follows. For each term t_i in the term list, the associated concepts C_{ij} are obtained from the ontology. The significance of each associated concept C_{ij} is initially taken as the normalized frequency of the term t_i . For each associated concept C_{ij} , we look at the presence of the related concepts in the document. We then increment the significance of the associated concept C_{ij} by $w \times$ term frequency for the occurrences of the terms corresponding to the related concept, where w is the weight given to the related concepts.

$$\text{Significance } C_{ij} = \text{Normalized term frequency } t_i + w \times \text{Term frequency of the related concept's corresponding term}$$

In our experiment, we have taken $w = 1/2$. For a particular term, we choose a concept with the maximum significance value.

Algorithm : Identification of Concept and its significance

Input: t_1, t_2, \dots, t_n is the list of domain terms in the document D ;

t_i frequency is the frequency of domain term t_i ;

num is the total number of tokens in the document D

Output: list of concepts c_1, c_2, \dots, c_m and their significance c_i .significance

```
(1) for  $i \leftarrow 1$  to  $n$  {
// Normalize the frequency counts
 $t_i$  frequency  $\leftarrow t_i$  frequency /  $num$ 
}
(2) for  $i \leftarrow 1$  to  $n$ 
{
```

```

ti.concepts ← {ci1 .. cij .. cik}
// where {ci1 .. cij .. cik} be the list of associated concepts of ti
cik.significance ← ti.frequency
}
(3) for i ← 1 to n
{
for j ← 1 to k
{
Let RC be the set of related concepts of cij in D.
for every term t corresponding to a concept in RC
cij.significance ← cij.significance + w × t.frequency // we take w = 1/2
}
}
(4) // Select the final concept
for i ← 1 to n
{
find the concept x in ti.concepts which has the highest significance score.
if x.significance > threshold
return x
else
return null
}

```

The algorithm returns the list of selected concepts and their significance scores. The algorithm is linear in the length of the document and relevant portion of the ontology.

Concept Type Identification

In order to assess whether a learning material is useful for a given student, it is useful to know not only the concepts that the learning material includes, but also the role of the concepts in the learning material. In particular, it is important to know whether a concept mentioned in a document is a prerequisite for studying that document, or it can be an outcome, learned by studying the document. In this work, we attempt to identify the role of concepts by parsing the sentences containing mentions of the concepts. In the context of individual sentences, we use the notion of two types of concepts i. e. *defined concepts* and *used concepts*. Definitions of the different types of concepts are as follows:

Outcome concept: The outcome concept is the concept that a learner learns from the document.

Prerequisite concept: The *prerequisite concepts* for the document are the concepts, which are used to explain the *outcome concept* and required to be known by a learner to understand the document.

Defined concept: A concept is said to be a *defined concept*, if it is defined in a sentence.

Used concept: A concept that has been mentioned in a sentence to explain something is said to be a used concept.

A learner usually learns the concepts, which are defined somewhere in a document. Therefore the list of *defined concepts* gives the *outcome concepts* for a document. The *prerequisite concepts* are used to explain the outcome concepts. Therefore the list of *used concepts* should give the list of *prerequisite*

concepts for the document, but all the concepts present in the *used concept* list in a document may not be the prerequisite for the document. For example, a concept *x* is defined in the first paragraph of a document, then the same concept *x* is used to define some other concept *y* in the second paragraph. Although, the concept *x* is a *used concept* for defining the concept *y*, but it is not the *prerequisite* for the document.

The *defined concepts* list gives the *outcome concepts* of the document. To find the *prerequisite concepts* list, the *used concepts* list is compared with the *defined concepts* list. The *defined concept*, which also exists in the *used concepts* list is removed.

To extract the type of concepts, our approach uses features such as verbs, cue phrases with their associated semantics in conjunction with patterns. The sentences, which state definitions usually contain verbs like “*defined*”, “*derived*”, “*called*”, “*known*”, “*states*”, and follow some pattern. Occurrences of verbs like “*deal*”, “*described*”, “*discussed*”, “*explained*” etc. indicate that some concept is being explained in a document. Sentences that contain one of these trigger verbs are further analyzed to find the role of the concepts mentioned in them.

We analyze sentences using a shallow parsing approach. Sentences with these trigger verbs are parsed using a publicly available parser called the link parser (<http://bobo.link.cs.cmu.edu/link/>) and the constituent tree is obtained with labels. Labels associated with links represent the type of dependency and represents a direct semantic relationship. A path is extracted from each sentence using link labels, which give a semantic relationship between words. Let us take a sentence “*work is defined as force acting upon an object to cause a displacement*”. In this sentence the subject *work*, which precedes the trigger verb is the defined concept, while other concepts in the sentence such as *force*, *object* and *displacement* are the used concepts. But there are many ways to write the same sentence. For example the above sentence can also be written in the following way, “*Force acting upon an object to cause a displacement is called work*”. In the above sentence, the object of the sentence *work*, which is connected with the trigger word *called*, gives the defined concept whereas the subject contains the used concepts. We use different inference rules to obtain the type of concept from sentences.

The algorithm for concept type identification was tested on 50 documents. Each document was processed to produce the list of concepts. Each document was then read manually to identify the type of concepts in the list. The document was also processed through our algorithm to produce the same categorization into defined and used concepts. It may be noted that we get a large number of *common concepts* like distance, angle etc. which have been mentioned in documents but not specifically used to define or explain the *defined concepts*. These common concepts are not counted in the list of used concepts in the manual checking. Total *used-concepts* give the total number of concepts that are specifically used for defining the *defined concepts*. The performance of the algorithm is shown in Table 3.

Table 3
Performance evaluation of concept type identification

Total Documents	Total Concepts	Manual Observation		Algorithm Output	
		Total Used Concepts	Total Defined Concepts	Total Used Concepts	Total Defined Concepts
50	952	483	220	284	144

We found that in many cases, the used concepts may have been mentioned in sentences prior to the sentence which contains the defined concept. For example, “*The lens in your eye is elastic and can change its shape to accommodate the distances of various objects. This is called accommodation*”. In some cases the used concepts are present in sentences after the defining sentence. For example “*A condition which is common is called farsightedness. This is where the eye can't focus close objects and the image forms beyond the retina*”. The use of a shallow parsing approach for identification of the type of concepts from only those sentences, which contains the trigger word, misses many of the used concepts. There are many sentences, which define some concepts, but are not considered for further analysis. For example “*The focal length is the distance from the center of the lens to the focal point*”. In this sentence, *focal length* is the defined concept, but the parser is not able to identify it as a defined concept. Presently, the algorithm is incapable of handling sentences following the pattern of the example sentence mentioned above. To improve the performance of the algorithm, the inference rules for types of sentences like these have to be discovered and all those sentences have to be taken into account.

Topic Identification

Many researchers have carried out the work on automatic generation of topics from web documents, and they have used different approaches. The work of automatic generation of the subject of a document by Li et al. (2004) is based on a neural network. Haruechaiyasak et al. (2002) proposed a method of automatically classifying web documents into a set of categories using fuzzy association.

Jovanovic et al. (2006) uses domain ontology for annotating documents with a subject attribute. In their work they have worked on documents which contain a number of slides. The whole document is the learning object and the different slides of the document forms the content object. Initially, the author provides the subject of the learning object. The annotation of the different content objects (or slides) is done by looking at the related concepts of the subject of the learning object in their domain knowledge. The annotation of content objects depends on the subject of the learning object. It fails to annotate the content objects if the subject of the learning object is not available. A limitation of their work on subject identification is that it needs the author's supplied information.

Our approach also uses the knowledge base for automatic identification of the topic of the document. However, we want to identify the topic of a document in a fully automated way.

The knowledge base contains topic taxonomy. Each topic of this taxonomy is associated with a list of concepts present in it. We identify the topic of a document on the basis of the concepts that occur in the document. We first identify the list of concepts present in a document. A concept can belong to more than one topic. For each concept, we find the set of topics that include that concept from the domain knowledge. From these we find out the possible topics for the document. When we consider a concept, the counters corresponding to each topic that includes the concept is incremented by 1. The topic with the maximum score is returned as the topic of the document.

But when we applied the above-discussed algorithm, we observed that in some cases, it identifies a document as belonging to a sibling topic. This is because in our domain knowledge, the sibling topics often share many common concepts. So we augmented the above algorithm by using the frequency values of the concepts. Further we note that each concept is not of the same significance for a topic. For example, if we take documents of topics mirror and lens, we find that they have many common concepts. But the concept *mirror* is very specific to the topic mirror and *lens* for the topic lens. If we give more weights to these significant concepts while scoring, then the classification

accuracy is improved. For this reason, we augment the domain knowledge structure, and a specificity index (SI) of value 1 is associated with the significant concepts of a topic. The other concepts are associated with a specificity value of 0.5. We note that a similar principle is used by Gelbukh et al. (1999) for detecting the main topics of the document.

Thus, the modified algorithm is as follows. For each concept we find the topic name and the *SI value* with respect to the topic. The score for a topic is incremented by the *concept frequency* × (*SI value*). The topic or list of topics with the maximum score gives the topic/topics for the document.

The performance of the topic identification algorithm is tested on documents belonging to different topics. Table 4 shows some of the selected topics from Physics, Biology and Geography on which the algorithm is tested. 770 documents are used for testing, and the output of the algorithm is manually verified. The performance evaluation of the algorithm in terms of recall and precision is shown in Table 4.

Table 4
Accuracy of the topic identification algorithm in percentage

Topic	Recall	Precision
Photosynthesis	91.66	84.61
Human circulatory system	80	83.33
Respiration	64.28	90
Human eye	78.94	88.23
Human ear	70	93.33
Chromosome	61.33	88.88
Lens	97.88	84.21
Mirror	88.66	92.85
Newton's law of motion	93	86.66
Soil	86.95	78.57
Rock	90	64.28

We find that our system is fairly successful in classifying documents. The few cases of error mostly categorize a document that has been manually labeled as belonging to a topic *topic_x* to a topic *topic_y*, where *topic_y* is the sibling of *topic_x*.

Learning Resource Type Identification

As discussed earlier, the pedagogic type of a document like *Narrative text type*, *questionnaire type* etc. are important metadata for e-learning. To identify the document type, we have identified some of the surface level features of the text (Kessler et al., 1997; Rauber & Muller-Kogler, 2001) and used these features to classify the documents into different types using neural network.

Feature set

The feature set consists of a set of specific verbs, trigger words, phrases and special characters.

The verbs in a document can be viewed as part of a conceptual map of the events and actions in a document. A verb is an important factor in providing an event profile, which in turn is useful for categorizing documents. *Narrative text type* documents contain discussion about a concept or concepts. Such documents generally contain definitions, statements of laws or facts about concepts (<http://www.cancore.ca/en/help/44.html>). Therefore verbs like “*define*”, “*known*”, “*state*”, “*described*”, “*explained*”, “*discussed*”, “*illustrated*” etc. are frequently found in *Narrative text type* documents. *Experiment type* documents often contain verbs like “*study*”, “*observe*”, “*design*”, “*measure*” etc. *Questionnaire type* documents usually contain verbs like “*evaluate*”, “*find*” etc.

Apart from verbs, the occurrences of some words and phrases (Grover et al., 2003) play an important role in describing documents. Documents belonging to the category *experiment* usually contain words like “*introduction*”, “*objective*”, “*results*”, “*goal*” etc. *Questionnaire type* documents usually contains phrases such as “*describe how*”, “*show that*”, “*why does*”, “*how can*” etc.

Some special characters like punctuation marks and special symbols also play an important role. *Questionnaire type* documents may contain interrogative sentences, which can be identified as they end with a question mark. These character level cues are important and used for classification.

The distribution of features in documents

The distribution of the features in some of the documents randomly selected from all the three classes *Narrative text*, *Experiment* and *Questionnaire* are shown in Figure 2. The X-axis represents the feature set (consisting of 60 features like *defined*, *discussed*, *evaluate*, etc.) and the Y-axis represents the frequency of occurrence of a feature in a given type of document. In Figure 2, documents 1, 2, and 3 belong to the *Narrative text type*, documents 4, 5, and 6 belong to *Experiment type* and documents 7, 8, and 9 belong to *Questionnaire type*. From visual inspection, it is apparent that the document types cannot be classified using a linear classifier. We have chosen Artificial Neural Networks (ANN) for the task of classification of the given data set. We have experimented with two types of neural networks for the classification task, a feed forward back propagation network with non-linear activation function (BPNN) and a generalized regression neural network (GRNN) with a Gaussian basis function. It was found experimentally that the raw count of trigger words was not enough for the classifiers to work well specially with words like *what*, *which* etc. Some preprocessing of the documents was done to modify the weights of these features according to their immediate context. Due to the limitation of space, only the basic structure of the classifiers is presented in this paper. The detailed discussion and the design aspects of the classifiers are available at <http://www.facweb.iitkgp.ernet.in/~sudeshna/devshrithesis.pdf>.

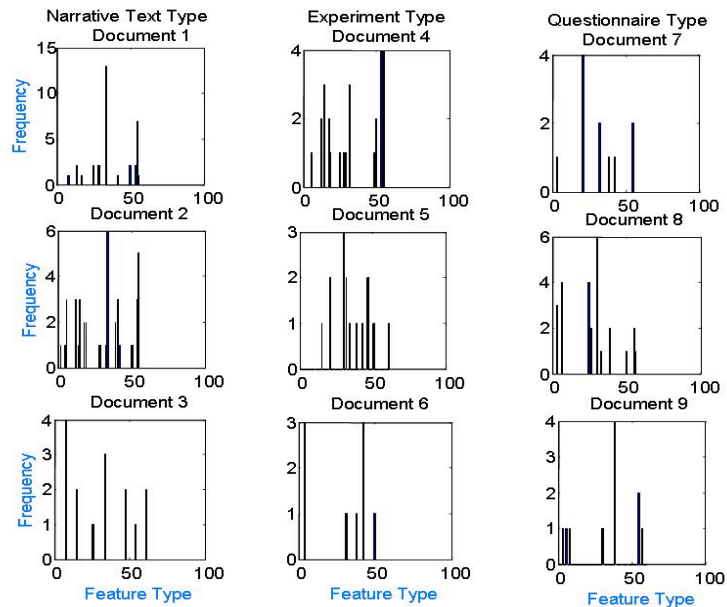


Fig.2. Distribution of features.

Feed Forward Back Propagation Neural Network

We have chosen a 2-layer feed forward network. The number of hidden neurons has been fixed by trial and error over a number of training and test sets (Wasserman, 1993). The structure of the network is as follows:

Number of input nodes [same as number of features] = 60

Number of hidden Neurons = 6 with tan-hyperbolic activation function

Number of output [same as number of classes] = 3 with tan-hyperbolic activation function

There are cases where a document can belong to more than one category. For example, a document may contain both an explanation of some topic and also questions on the topic. To handle these cases, the classifier output classes are rendered into vectors as:

Class 1 + Ve, - Ve, - Ve

Class 2 - Ve, + Ve, - Ve

Class 3 - Ve, - Ve, + Ve

and the positive values at the output are taken as the output classes.

Generalized Regression Neural Network

Generalized Regression Neural Networks are based on non-linear regression. The network consists of three layers i.e. (1) the input layer, (2) the hidden layer where a nonlinear transformation is applied on the data from the input space to the hidden space and (3) the linear output layer. There are weights connecting the input to the hidden nodes. We have chosen a multivariate Gaussian function with an appropriate mean and autocovariance matrix. The principal advantage of GRNN is fast learning and

convergence to an optimal regression surface. Therefore it is particularly advantageous to use this method.

We consider three classes of learning resource types namely *Narrative Text type*, *Experiment type* and *Questionnaire type*. We have collected web documents and manually categorized them into the above classes. Out of a total of 150 documents, 120 documents are randomly chosen and used to train the Feed forward back-propagation classifier and generalized regression neural network classifier. The remaining 30 documents are tested with the trained classifier. The above process is repeated 5 times with different sets of documents randomly chosen from the data set.

Table 5
Classification Performance of Feed Forward Back Propagation Neural Network

Type of learning resource type	Precision	Recall
Narrative Text	89.13	92.56
Experiment	78.63	92.244
Questionnaire	79.284	76.52

Table 6
Classification Performance of Generalized Regression Neural Network

Type of learning resource type	Precision	Recall
Narrative Text	86.81	81.36
Experiment	98.75	94.42
Questionnaire	72.14	77.88

The classification performances of Feed Forward Back Propagation Neural Network and Generalized Regression Neural Network are given in Table 5 and Table 6. It may be noted that the above performance is obtained using the preprocessed feature vector values as mentioned earlier. The misclassifications are mainly in *Questionnaire* documents, which are of multiple-choice types or fill in the blanks type. The classifier classified them as *Narrative Text* type. We again found that in some documents of the experimental type, the document does not contain the trigger words like *Objective*, *Equipment*, *Apparatus*, etc., and thus our classifier fails on those documents. Some improvement may be obtained by increasing the number of features. We will be working on improving the feature set.

AUTOMATIC ANNOTATION TOOL

We incorporated the various algorithms discussed above and built an automatic annotation tool. The input interface of the automatic annotation tool is very simple. It has two input buttons *Document Submission* and *Web Based Submission*. The *Document Submission* input accepts documents from contributors. The *Web Based Submission* provides the facility of accepting documents directly from the web. It accepts the global address of a document (url) on World Wide Web.

The submitted document is sent to the automatic metadata extractor module of the system that extracts the available metadata automatically. It extracts all types of metadata information such as *general*, *technical*, *educational*, *classification* etc. from the document. General and technical category metadata such as *size* of the document, *format*, *date* etc. are extracted automatically from the system

properties (not discussed in this paper). The different pedagogic attributes of the document are extracted by different methods as discussed in the section on Automatic Metadata Extraction.

Metadata annotation is done in a machine comprehensible format. The metadata annotation is expressed in a semantic web language. The metadata elements discussed in the section on Metadata Schema are compliant with IEEE LOM RDF binding specification (Nilsson, 2002) except for the metadata attributes, which are extended by us. For classification category, we use the *dc:subject* to point to a topic from the domain ontology. This does not fully conform to the IEEE LOM RDF Binding since topic hierarchy in our domain knowledge is not according to LOM taxonomy. The metadata annotation of a document with LOM RDF binding is shown below.

```
<?xml version="1.0" ?>
=> <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:lom="http://ltsc.ieee.org/2002/09/lom-base#" xmlns:lom-
gen="http://ltsc.ieee.org/2002/09/lom-general#" xmlns:lom-
life="http://ltsc.ieee.org/2002/09/lom-lifecycle#" xmlns:lom-
meta="http://ltsc.ieee.org/2002/09/lom-metametadata#" xmlns:lom-
tech="http://ltsc.ieee.org/2002/09/lom-technical#" xmlns:lom-
edu="http://ltsc.ieee.org/2002/09/lom-educational#" xmlns:lom-
cls="http://ltsc.ieee.org/2002/09/lom-classification#"
  xmlns:myVoc="http://www.myVocabulary.com/someVocab#">
=> <rdf:Description>
  rdf:about="http://www.physicsclassroom.com/Class/refrn/U14L5a.html"/>
=> <dc:date>
  => <dcterms:W3CDTF>
    <rdf:value>2005-10-23</rdf:value>
  </dcterms:W3CDTF>
</dc:date>
=> <dc:format>
  => <dcterms:IMT>
    <rdf:value>html</rdf:value>
  </dcterms:IMT>
</dc:format>
=> <dcterms:extent>
  => <lom-tech:ByteSize>
    <rdf:value>124032</rdf:value>
  </lom-tech:ByteSize>
</dcterms:extent>
<lom-tech:location
  rdf:resource="http://www.physicsclassroom.com/Class/refrn/U14L5a.html"
/>
<rdf:type rdf:resource="http://ltsc.ieee.org/2002/09/lom-
educational/NarrativeText" />
<lom-edu:context rdf:resource="myVoc;grade 7" />
<dc:subject>concave mirror</dc:subject>
=> <myVoc:keyword-list>
  => <keyword>
    <name>distance</name>
    <frequency>2</frequency>
  </keyword>
  => <keyword>
    <name>concave mirror</name>
```

```

        <frequency>31</frequency>
    </keyword>
    .
    .
    = <keyword>
        <name>angle of incidence</name>
        <frequency>4</frequency>
    </keyword>
</myVoc:keyword-list>
= <myVoc:concept-list>
    = <concept>
        <name>concave mirror</name>
        <significance>36</significance>
        <type>used concept</type>
    </concept>
    .
    .
    = <concept>
        <name>incident angle</name>
        <significance>9</significance>
        <type>used concept</type>
    </concept>
</myVoc:concept-list>
<myVoc:location-metadata rdf:resource="D:/project-
repository/repository/http://www.physicsclassroom.com/Class/refrn/U14L
5a.rdf" />
</rdf:Description>
</rdf:RDF>

```

CONCLUSION

Our work addresses the need for making learning materials reusable and interoperable between different learning systems. The web is a large source of learning materials. The ability to automatically annotate learning materials will enable learners and learning systems to harness the resources available in the web and build an open repository. To increase the potential reuse of learning contents of the open repository, we have annotated documents with a subset of effective and descriptive metadata from the IEEE LOM specification.

We have developed an automatic annotation tool for annotating documents with metadata such as *concepts*, *type of concepts*, *topic* and the *learning resource type*. The domain knowledge has been used to extract the concepts and its type from documents. The *type of concept* is identified by analyzing the documents with a shallow parsing approach and by using some inference rules. As can be seen for concept type identification, we have introduced the idea of common concepts, which appear in the ontology and the documents but are of a general nature and not perceived to be a direct prerequisite for the document. Clearly, the number of such concepts would depend on the number of subjects included in the ontology and the distinction between common concepts and used concepts in manual checking would have a personal bias. This bias could be corrected if manual checking could be standardized by using a number of persons and some formal methodology. However it was not possible to adopt this due to practical limitations. The results given in the paper for concept type identification should thus be taken as indicative. For the automatic identification of concept type there seems to be a trade off

between the generality of the ontology and the depth of analysis required for *concept type* identification.

In order to automatically identify the pedagogic attributes of relevance, we have at present classified documents from the instructional perspective into the following categories *narrative text*, *questionnaire* and *experiment* from the IEEE LOM resource type. However, it is also important to identify the other resource types from the IEEE LOM, RDN/LTSN LOM application profile, UKLOM core etc. by a learning system and this may be an important direction of future work.

It may be noted that for *document type* classification, we have used only three types of documents from the large set provided in the IEEE LOM resource type. Even with these few types, some semantic preprocessing was necessary in order to get acceptable results. In order to include more document types, a more thorough and deeper semantic preprocessing of the learning materials seems to be necessary. Further research is required in this area.

The results of the algorithms mentioned in the paper are compared with manual observations. A single person did the manual observation for the attribute *concept type*, whereas several people did the manual tagging of documents with the attributes *topic* and *document type*. The difference in the opinions were mutually discussed and resolved. A formal approach for human bias elimination could be necessary when the results from improved algorithms approach more acceptable levels of performance for widespread deployment in future.

The usefulness of the kind of repository created through our system needs to be validated through field trials, especially among students. This could not be undertaken as yet but it would be an interesting study and feedback from such a study may enable us to further refine the set of metadata.

ACKNOWLEDGEMENT

We are grateful to the reviewers for their valuable suggestions that have enabled us to improve the quality of the paper.

REFERENCES

- Aitken, S., & Reid, S. (2000). Evaluation of an Ontology-Based Information Retrieval Tool. In A. Gomez-Perez, V. R. Benjamins, N. Guarino & M. Uschold (Eds.) *Proceedings of Workshop on the Applications of Ontologies and Problem-Solving Methods*. <http://www.aii.ed.ac.uk/~stuart/Papers/ontologyeval.pdf>, European Conference on Artificial Intelligence 2000, Berlin, Germany.
- Anderson, J. D., & Perez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: research and the nature of human indexing. *Information Processing Management*, 37, 231-254.
- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. New York: Addison Wesley Longman Publishing Co. Inc.
- Biemann, C. (2005). Ontology Learning from Text: A Survey of Methods. *LDV-Forum*, 20(2), 75-93.
- Brooks, C., McCalla, G., & Winter M. (2005). Flexible Learning Object Metadata. In L. Aroyo & D. Dicheva (Eds.) *SW-EL'05: 3rd International Workshop on Applications of Semantic Web Technologies for e-Learning*. Held in conjunction with the 12th International Conference on Artificial Intelligence in Education (AIED 2005). <http://www.win.tue.nl/SW-EL/2005/swel05-aied05/proceedings/2-Brooks-final-full.pdf>. Amsterdam, The Netherlands.

- Cardinaels, K., Meire, M., & Duval, E. (2005). Automatic Metadata Generation: the Simple Indexing Interface. In A. Ellis & T. Hagina (Eds.) *Proceedings of the 14th International Conference on World Wide Web Committee* (pp. 548–556). Chiba, Japan. ACM Press.
- Cimiano, P., Ladwig, G., & Staab, S. (2005). Gimme' the context: context-driven automatic semantic annotation with C-PANKOW. In A. Ellis & T. Hagina (Eds.) *Proceedings of the 14th International WWW conference* (pp 332-341). Chiba, Japan. ACM Press.
- Dicheva, D., & Dichev, C. (2004). A Framework for Concept-Based Digital Course Libraries. *Journal of Interactive Learning Research*, 15(4), 347-364.
- Dicheva, D., Dichev, C., Sun, Y., & Nao S. (2004). Authoring Topic Maps-based Digital Course Libraries. In L. Aroyo & D. Dicheva (Eds.) *Proceedings of Workshop on Applications of Semantic Web Technologies for Adaptive Educational Hypermedia* (pp. 331-337). Held in conjunction with AH 2004, Eindhoven, The Netherlands.
- Dicheva, D., Sosnovsky, S., Gavrilova, T., & Brusilovsky, P. (2005). Ontological Web Portal for Educational Ontologies. In L. Aroyo & D. Dicheva (Eds.) *SW-EL'05: 3rd International Workshop on Applications of Semantic Web Technologies for e-Learning*. Held in conjunction with the 12th International Conference on Artificial Intelligence in Education (AIED 2005). <http://www.win.tue.nl/SW-EL/2005/swel05-aied05/proceedings/4-Dicheva-final-full.pdf>. Amsterdam, The Netherlands.
- Downes, S. (2004). Resource Profiles. *Journal of Interactive Media in Education*, 5. *Special Issue on the Educational Semantic Web*. Special Issue on the Educational Semantic Web. <http://www-jime.open.ac.uk/2004/5>
- Duval, E., & Hodgins, W. (2004). Making metadata go away - hiding everything but the benefits. In W. Jianzhong (Ed.) *Proceedings of DC 2004: the International Conference on Dublin Core and Metadata Applications* (pp 29-35). Shanghai, China. Emerald Group Publishing Ltd.
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., Weld, D. S., & Yates, A. (2004). Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. In J. A. Hendler, G. Ferguson & D. L. McGuinness (Eds.) *Proceedings of the 19th AAAI National Conference on Artificial Intelligence* (pp. 391-398). AAAI Press.
- Friesen, N. (2004), International LOM survey: Report. ISO/IEC JTC1/SC36 sub-committee.
- Gasevic, D., Jovanovic, J., Devedzic, V., & Boskovic, M. (2005). Ontologies for Reusing Learning Object Content. In P. Goodyear, D. G. Sampson, Kinshuk, D. J. Yang, T. Okamoto, R. Hartley & N. Chen (Eds.) *Proceedings of 3rd International Workshop on Applications of Semantic Web Technologies for E-Learning* (pp. 944-945). The 5th IEEE International Conference on Advanced Learning Technologies. Kaohsiung, Taiwan. IEEE Computer Society Press.
- Gelbukh, A. F., Sidorov, G., & Guzman-Arenas, A. (1999). Document Comparison with a Weighted Topic Hierarchy. In H. Kosch, L. Brunnie & A. Hameurlain (Eds.) *Proceedings of 10th International Workshop on Database & Expert Systems Applications* (pp. 566-570). University of Florence, Italy. IEEE Computer Society.
- Grover, C., Hachey, B., Hughson, I., & Korycinski, C. (2003). Automatic Summarisation of Legal Documents. In G. Sartor, J. Zeleznikow & L. Edwards (Eds.) *Proceedings of ICAIL 03: International Conference on Artificial Intelligence and Law* (pp. 243-251). New York :ACM.
- Han, H. C., Giles, L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E.A. (2003). Automatic Document Metadata Extraction using Support Vector Machines. In C. Marshall, G. Henry & L. Delcambre (Eds.) *Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Library* (pp. 37- 48). New York: ACM Press.
- Haruechaiyasak, C., Shyu, M., Chen, S., & Li, X. (2002). Web Document Classification Based on Fuzzy Association. In I. Sommerville & H. Yang (Eds.) *Proceedings of COMPSAC'02: the 26th Annual International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopment* (pp. 487-492). Oxford, England. IEEE Computer Society.
- Henstock, P. V., Pack, D. J., Lee, Y., & Weinstein, C. J. (2001). Toward An Improved Concept-Based Information Retrieval System. In P. V. Henstock, D. J. Pack, Y-S Lee & C. J. Weinstein (Eds.)

- Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR'01* (pp. 384-385). New York: ACM Press.
- Hoermann, S., Seeberg, C., Divac-Krnic, L., Merkel, O., Faatz, A., & Steinmetz, R. (2003). Building Structures of Reusable Educational Content Based on LOM. In J. Eder, R. Mittermeir, B. Pernici, M. Bessagnet & M. Sala (Eds.) *Proceedings of SW-WL'03: Workshop on Semantic Web for Web-based Learning* (pp. 234-243). The 15th Conference on Advanced Information Systems Engineering Klagenfurt/Velden, Austria, Europe. CEUR-WS.org.
- Jenkins, C., & Inman, D. (2000). Server-side Automatic Metadata Generation using Qualified Dublin Core and RDF. In Y. Kambayashi, G. Wiederhold, J. Klavans, W. Winiwater & H. Tarumi (Eds.) *Proceedings of International Conference on Digital Libraries: research and practice* (pp. 262-269). Kyoto, California. IEEE Computer Society.
- Jovanovic, J., Gasevic, D., & Devedzic, V. (2006). Ontology-Based Automatic Annotation of Learning Content. *International Journal on Semantic Web and Information Systems*, 2(2), 91-119.
- Kessler, B., Nunberg G., & Schutze H. (1997). Automatic Detection of Text Genre. In P. R. Cohen & W. Wahlster (Eds.) *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European chapter of the Association for Computational Linguistics* (pp. 32-38). Somerset, NJ: Association for Computational Linguistics.
- Lauser, B., Wildemann, T., Poulos, A., Fisseha, F., Keizer, J., & Katz, S. (2002). A Comprehensive Framework for Building Multilingual Domain Ontologies: Creating a Prototype Biosecurity Ontology. In L. Bertini, P. Cotoneschi & A. Farsetti (Eds.) *Proceedings of International Conference on Dublin Core and Metadata for e-Communities* (pp. 113-123). Florence, Italy: Firenze University Press.
- Li, Y., Zhu, Q., & Cao, Y. (2004). Automatic metadata generation based on Neural Network. In J. R. White, H. Sheng (Eds.) *Proceedings of the 3rd International Conference on Information Security* (pp. 192-197). New York: ACM.
- Liu, J., & Greer, J. (2004). Individualized Selection of Learning Object. In L. Aroyo & D. Dicheva (Eds.) *Proceedings of SW-EL'04: Workshop on Applications of Semantic Web Technologies for Web-based ITS* (pp 29-34). Maceió, Brazil.
- McCalla, G. (2004). The Ecological Approach to the Design of E-learning Environments: Purpose-based Capture and Use of Information About Learners. *Journal of Interactive Media in Education*, 7. Special Issue on the Educational Semantic Web. www-jime.open.ac.uk/2004/7.
- Mohan, P., & Greer, J. (2003). E-learning Specification in the context of Instructional Planning. In U. Hoppe, F. Verdejo & J. Kay (Eds.) *Proceedings of AIED 2003: International Conference on Artificial Intelligence in Education* (pp. 307-314). Amsterdam: IOS Press.
- Nilsson, M.(Ed.). (2002). IEEE Learning Object Metadata RDF binding. <http://kmr.nada.kth.se/el/ims/md-lomrdf.html>
- Ochoa, X., Cardinaels, K., Meire, M., & Duval, E. (2005). Frameworks for the Automatic Indexation of Learning Management Systems Content into Learning Object Repositories. In P. Kommers & G. Richards (Eds.) *Proceedings of World Conference on Educational Multimedia, Hypermedia & Telecommunications EDMEDIA 2005* (pp. 1407-1414). Charlottesville, VA: AACE.
- Papanikolaou, K. A., Grigoriadou, M, Magoulas, G. D., & Kornilakis, H. (2002). Towards New Forms of Knowledge Communication: The Adaptive Dimensions of a Web-based Learning Environment. *Computers and Education*, 39, 333-360.
- Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., & Goranov, M. (2003). KIM - Semantic Annotation Platform. In G. Goos, J. Hartmanis & J. van Leeuwen (Eds.) *Proceedings of the 2nd International Semantic Web Conference* (pp. 834-849). Berlin Heidelberg: Springer.
- Rauber, A., & Muller-Kogler, A. (2001). Integrating Automatic Genre Analysis into Digital Libraries. In E. A. Fox & C. L. Borgman (Eds.) *Proceedings of JCDL 2001: First ACM/IEEE Joint Conference on Digital Libraries* (pp. 1-10). New York: ACM Press.
- Salton, G., & McGill M. (1984). *Introduction to Modern Information Retrieval*. London: McGraw-Hill Book Company.

- Simic, G. (2004). The multi-courses Tutoring Systems Design. *Computer Science and Information Systems, ComSIS*, 1, 1, 141-155.
- Simon, B., Dolog, P., Miklos, Z., Olmedilla, D., & Michael, S. (2004). Conceptualizing Smart Spaces for Learning. *Journal of Interactive Media in Education*, 9. Special Issue on the Educational Semantic Web www-jime.open.ac.uk/2004/9.
- Song, H., Zhong, L., Wang, H., Li R., & Xia, H. (2005). Constructing an Ontology for Web-based Education Resource Library. In L. Aroyo & D. Dicheva (Eds.) Proceedings of the ICALT (International Conference on Advanced Learning Technologies) 2005 Workshop on Applications of Semantic Web Technologies for e-Learning. Banff, Canada. <http://www.win.tue.nl/SW-EL/2005/swel05-kcap05/proceedings/Poster-3-Huazhu.pdf>.
- Tan, M., & Goh, A. (2004). The Use of Ontologies in Web-based Learning. In L. Aroyo & D. Dicheva, R. Mizoguchi & Y. Itoh (Eds.) *Proceedings of Workshop on Applications of Semantic Web Technologies for e-Learning* (pp 75-80). International Semantic Web Conference (ISWC 2004), Hiroshima, Japan.
- Ullrich C. (2004). Description of an Instructional Ontology and its Application in Web Services for Education. In S.A. McIlraith, D. Plexousakis & F.V. Harmelen (Eds.) *Poster proceeding of ISWC2004: 3rd International Semantic Web Conference* (pp 93-94). Berlin Heidelberg: Springer.
- Wasserman, P.D. (1993). *Advanced Methods in Neural Computing*. New York: Van Nostrand Reinhold.
- Weber, G., Kuhl, H., & Weibelzahl, S. (2002). Developing Adaptive Internet Based Courses with the Authoring System NetCoach. In S. Reich, M. Tzagarakis & P. De Bra (Eds.) *Proceedings of International Workshop OHS-7, SC-3 and AH-3 on Hypermedia: Openness, Structural Awareness and Adaptivity* (pp. 226-238). Berlin Heidelberg: Springer.